Amsterdam
University
Press

COMPUTATIONAL COMMUNICATION RESEARCH 4.1 (2022) 173-207
HTTPS://DOI.ORG/10.5117/CCR2022.1.005.JURG

ARTICLE

Age and Gender Representation on German TV

A Longitudinal Computational Analysis

Pascal Jürgens
University of Mainz
pascal.juergens@uni-mainz.de

Christine E. Meltzer
University of Mainz
meltzer@uni-mainz.de

Michael Scharkow
University of Mainz
scharkow@uni-mainz.de

Abstract

Television offers an enticing glimpse into the world, but its perspective is often skewed. When societal groups are systematically excluded from appearing on the screen, they lose the chance to represent their characteristics and interests. Recipients may then form distorted perceptions and attitudes towards those groups. Empirical research on the prevalence of such biases - especially across stations, time, and genre - has been limited by the effort of manual content analyses. We develop and validate a deep-learning based method for measuring age and gender of faces in video material. An analysis of approximately 16 million faces from six years of German mainstream TV across six stations is fused with existing program metadata indicating timing and genre of broadcasts, including advertisements. Multilevel regression models show a consistent and temporally stable discrimination against women and elderly people, along with a double discrimination of elderly women. A significant amount of variation across genres and systematic differences between public and private broadcasters furthermore indicate previously undocumented heterogeneity in the representation of societal groups on TV. We discuss potential implications of a genre-specific differentiation against the backdrop of societal trends.

Keywords: deep learning, television, gender representation, age representation, image analysis

Introduction

Mainstream media, especially television, provides a constant flow of images of women and men, from which the audience can learn about meanings and values ascribed to gender in society. In doing so, they provide opportunities to form attitudes towards others and to see one's own social group represented. Decades of research have provided evidence that some of TV's content is not balanced, offering a misleading picture to learn from. Yet stations broadcast 24 hours a day, and it remains unclear how biased their program is, both overall and internally. A scarcity of large-scale analyses leaves crucial open questions: Are male-dominated shows counteracted by all-female shows? Do stations differ systematically in the gender they prefer? Which role does the genre play in determining representation? And, given the intense societal debate surrounding equality — has there been any meaningful change? We address these questions using a combination of deep learning methods and a large video dataset.

Television is of course neither a uniform product, nor is it consumed in its entirety. Because stations, shows and genres are produced and seen individually, we must consider second-order biases — where the representation of gender and age varies across structural dimensions of the TV programs, such as station, genre and time. Whereas aggregate biases are well-documented in the literature, comprehensive studies of visual representation across secondary dimensions remain elusive. Measurements of such structural differences are, however, becoming a pressing research gap: A heterogenous TV landscape may on average appear balanced — while containing highly skewed individual genres. At the same time, the growing supply of online video erodes the role of mainstream stations and necessitate larger and more expansive analyses. Research on the representation of age and gender is therefore hampered by a methodological bottleneck: The labor-intensive nature of manual video content analysis meant that studies so far could either focus on detailed coverage of few programs, or a superficial coding of many, but not both. Our contribution provides a crucial methodological advance that unlocks comprehensive research regardless of the scale of material under investigation. To this end, we (1) develop and validate an automated method for measuring age and gender representation and (2) perform a large-scale analysis of six years of German mainstream TV programs.

In our analysis, we draw on a comprehensive archive of TV broadcasts along with structural program metadata to offer both the substantial insights of a broadly representative analysis of age and gender on as well as a rigorous treatment of the computational methodological aspects. Our dataset comprises a total of 16 million faces seen on six German mainstream stations across six years of German television. This allows for a threefold quantitative comparative investigation on how (1) younger and older (2) women and men are represented in television (3) over time. For the first time we can present structural predictors of deficient gender representation of TV, using a comprehensive, yet fine-grained assessment. In line with previous research, multilevel regression models show that women on TV remain universally underrepresented and are consistently younger than men. We additionally find a significant amount of variation across genres and systematic differences between public and private broadcasters.

Gender and ageing in the media

Why study gendered ageism?

Past studies consistently show that women are underrepresented across a wide variety of media, including television (e.g., Collins, 2011; Edström, 2018; Eisend, 2010). This is even the case in developed countries that are highly ranked on gender equality indices (Matthes et al., 2016). When it intersects with gender, age – like race (see Lind & Meltzer, 2020), sexuality, or class – can result in double discrimination. In line with this, prior research has found that the media typically underrepresented older age groups, and this underrepresentation is more severe for women than for men (Baumann & de Laat, 2012; Kessler et al., 2010; Lee et al., 2007; Prieler et al., 2015).

The consequences of such (systematic) underrepresentation are not entirely clear and, so far, have not been rigorously tested (Collins, 2011). Part of the challenge lies in the large variety of very different content that exists across TV programs, each of which may exhibit distinct characteristics and elicit distinct effects. As a result, there is an empirically and theoretically heterogenous literature (which we hope will benefit from the broad empirical basis of our method). Some robust core findings warrant renewed scholarly interest in the role of gender and age representation as antecedents for harmful media effects. First, there are clear indications that underrepresentation in the media affects the "unseen" themselves as well as how they are viewed by society as a whole. The media's selection of actors plays an active role in who is entitled – and equally important, who is not – to participate in the public debate (McCombs, 2007). Hence, underrepresentation in the media prevents debates about the needs of the marginalized groups.

Second, cultivation theory suggests that television affects the idea of social reality in the audience by creating a symbolic environment over the life span of its viewers (Gerbner & Gross, 1976). In doing so, it provides the audiences with norms, with stories about winners and losers in society, hence telling who is meaningful, newsworthy, and desirable in society – and who is not (Fryberg & Townsend, 2008). This explains why those who are not visible are less inclined to be politically active (Campbell & Wolbrecht, 2006) and more constrained from taking powerful positions in society (Haraldsson & Wängnerud, 2019).

Third, going beyond the idea of mere representation and focusing on context, characters on screen can serve as role models (Bandura, 2001) and thus can have severe consequences for social reality. For instance, surveys have shown that female role models on television can increase young women's confidence to excel in a male-dominated profession (Fox, 2018) or even give them the courage to leave an abusive relationship (Geena Davis institute on Gender in the Media, 2015).

Testing these societal consequences goes beyond the scope of this paper. Yet, it is the aforementioned socializing and society-shaping role of TV that makes it necessary to understand the fundamental composition of the television world. We argue that TV should provide its viewers with a program that, as far as possible, does not depict a distorted, unrealistic, or discriminatory realty concerning who is and who is not seen. Television viewers with a wide variety of interests and viewing motives should see themselves reflected in the program they watch. Hence, overall television as well as specific genres in themselves should reflect a share of female and old persons that is comparable to their real share in society.

The analysis of gendered age on television

Acknowledging the evidently desolate state of gender representation and its plausible effects on individuals and societies, past studies have offered relatively focused yet isolated insights into empirical reality. When the team around George Gerbner started analyzing "television reality" in the 1970s, one major finding was that men were almost three times more visible than women (Gerbner & Signorielli, 1997). Tuchman (1978) summarized the state of women as being "symbolically annihilated" by the media.

In Germany, research in this field was pioneered by Küchenhoff and Bossmann (1975). The authors showed that women are quantitatively underrepresented in German television and, even if they speak, are not part of serious dialogues. It was concluded that “men act, women appear” (Küchenhoff & Bossmann , 1975, p. 142). Subsequent studies of the German television landscape corroborated these findings. Weiderer (1993) showed that women were still underrepresented to nearly the same extent as in 1975 in all genres under investigation. Yet, there seemed to be some differences between genres, and especially fictional vs. nonfictional television content: In 1975, there were no female presenters, anchors, or experts in television news, while in 1990, there were 32% female presenters and 4% female experts in the news (Weiderer, 1993). The most recent content analysis of German television found that across all television programs, there were two men for every woman (Prommer & Linke, 2019). This held true across all genres, for information as well as entertainment programs, and across all television stations under investigation, no matter if private or public. More than half (55%) of the programs under investigation had no female protagonist at all, whether as a presenter, guest, expert or similar. At the same time, only 16% of the programs lacked a male main protagonist. One striking exception to this pattern were Daily Soaps, where women and men characters were shown equally.

The pattern of imbalanced gender representation on German TV is mirrored in international studies. Women are underrepresented in popular television fiction (Hether & Murphy, 2010; Sink & Mastro, 2017; Smith et al., 2010; Van Bauwel, 2018), in advertising (Eisend, 2010; Matthes et al., 2016), and in music videos (Turner, 2011). In children’s programs, the gender disparity seems to be even more severe than that in television for the adult population (Prommer & Linke, 2019). The ratio seems to be slightly more balanced in television news (Cann & Mohr, 2001; Desmond & Danilewicz, 2010). As far as existing studies allow longitudinal comparisons, there seems to be a trend toward more gender equality (or adequate gender representation) in the media. Nevertheless, even across multiple decades, progress is of a rather slow nature, if significant at all (Sink & Mastro, 2017; Weiderer, 1993).

As outlined above, many studies have investigated gender representation on television. Research on the representation of different age groups, however, is scant. As for gender, a large share of the relevant studies was conducted in the context of advertisement (e.g., Baumann & de Laat, 2012; Kessler et al., 2010; Prieler et al., 2015). These studies find that, even though the representation of older people has changed in the direction of a more adequate representation of older age groups, it still remains far from realistic. For instance, Kessler and colleagues (2010) found only about 4.5% of the characters on German prime-time advertisements to be older than 60 years. Comparing commercials of five television stations in Japan, Prieler and colleagues (2015), found that the number of older people increased between 1997 and 2007 but that older people were still significantly underrepresented in 2007. The few studies that have looked at the representation of older people on television aside from advertisements come to a similar conclusion. Older people become almost invisible on screen, and this has not changed much over time (Edström, 2018; Signorielli, 2004). This is interesting in light of the fact that more and more countries in the world are ageing increasingly (United Nations, 2017).

When it intersects with gender, age can result in double discrimination. In line with that, research finds the underrepresentation of older women to be more extreme than the underrepresentation of older men (Baumann & de Laat, 2012; Edström, 2018). In the German context, Prommer and Linke (2019, p.54) note that there are hardly any older women to be seen on television, with a progressive underrepresentation starting around the age of 30 and getting worse towards higher ages. These findings nearly mirror the findings of Weiderer (1993); thus, the gendered ageism of women over 30 has not changed significantly in the last 30 years.

Summary and Research Gaps

Despite its evident importance, the literature currently offers an incomplete picture of the prevalence of age and gender on TV. Women are certainly less visible than men, and doubly discriminated when older. There seem to be programs that have nearly achieve a gender-equal representation (e.g., Daily Soaps, Prommer & Linke, 2019) while in others, women are severely underrepresented. Further, some genres have been investigated intensively (e.g., advertising), while others (e.g., non-fictional, editorial content) remain a white spot on the map. In the few studies that considered a longitudinal approach, some indicate a change towards more gender equality over time, but this seems to not be true concerning a realistic depiction of age. Hence, even if there might be some progress in gender equality on screen, it may come at the expense of older women — creating a specifically discriminated group in society. Looking at the intersectional dynamics of gender and age, it seems of great relevance to consider the two characteristics in the television landscape together, comparatively for different genres, and over a longer period, in order to get a holistic picture of the television worlds depiction of age and gender.

As we argued above, a modern assessment of the state of gender and age representation on TV needs to offer a broad survey — of the bulk of mainstream stations, throughout the day, and incorporate structural factors along which audiences might differentiate. Such an undertaking, while prohibitively costly with conventional content analysis, has become feasible in automated approaches. The extensive methodological section below outlines the logistics of this data-intensive work before discussing the deep learning-based automated face analysis and its verification. Our results draw on multilevel regression models to show that there is, in fact, a significant heterogeneity to be found. Gender representation remains universally deficient but varies strongly across genres.

Method and data

Our work aims to reduce the bottleneck of manual content analysis, whose effort and cost limit the scope of empirical research in this field. To do so, we draw on advances in artificial neural networks (or "deep learning") to create and validate a classifier for age and gender of images containing faces. This classifier is then applied to a large dataset of TV coverage from six German mainstream TV stations to provide the most comprehensive empirical assessment yet. A repository containing all code and a combination of aggregate and raw data used in the study is available on OSF [1].

Sample

A major challenge in the analysis of television material is the effort and cost associated with obtaining suitable samples. Stations typically charge around 100€ for recordings of a single show, and commercial databases of program data are scarce and expensive (WDR, 2021). We draw on a television archive system maintained by one of the authors which comprises ten years of daytime programming (9.10 am to 10.50 pm) from six German TV stations with the largest reach: Public service broadcasters ARD and ZDF as well as private stations Pro7, Sat.1, RTL, and VOX.

As outlined above, the representation of age and gender may vary across structural variables including station, time and genre. Genre is a particularly important confounder, since it includes advertisements which are less frequent in highly regulated public service broadcasters (Die Medienanstalten, 2019). We were able to obtain detailed data describing genres on a per-second basis – a dataset used by German media regulation authorities (ALM) in their continuous yearly assessment of key content parameters (Becker et al., 2018). These records are designed to produce a representative sample of TV broadcasts used in the scrutiny of, among others, compliance to advertising regulations and content diversity mandates (Weiß et al., 2020). Advertisement slots are listed as distinct entries. To enhance cross-station comparisons, we recoded genres into broader categories: News, editorial content (which includes information programs such as talk shows, documentaries or magazines), non-fictional entertainment, fictional entertainment, sports, and advertisements.

The ALM program metadata spans the years from 2012 to 2017, containing two weeks from each except for 2017, when a new sample scheme was introduced, for a total of 77 days. A shift in the responsibility of data collection meant that we lack structural data for 2018 and beyond. Due to technical maintenance periods, the final four days of the 2017 sample and one day from 2012 are missing in the TV archive (as are some hours from the rest of the sample - see figure 3), leaving us with a total of 72 broadcast days with metadata, comprising 13 hours of material. The total amount of analyzed video material is therefore 6 stations x 13 hours x 72 days = approximately 5,600 hours. Since human perception barely notices sub-second visual cues, we sampled one frame (out of 24) per second, yielding a total of 20,217,600 images. Using the neural network face detection model described below, we extracted 16,109,738 faces from 3,364,207 of these images - the rest contained no detectable person or failed to meet a minimum threshold for face size (see below). Fusion with the program metadata yielded a total of n=57,919 distinct shows, which will form the base of our further analysis.

Methods - Face Analysis

Actors on TV leave viewers with numerous impressions through their appearance, utterances, and actions. The single most important expression, however, are faces: These serve as communicative anchor points for speaking and acting, for personality, expression of emotion, identification with characters, and of course as cues for gender and age attributes. Just like human viewers do, we draw on the appearance of manifest faces as an indicator for the prevalence of genders and ages on TV programs. Recent advances in computer science, notably the rapid evolution of neural networks, have made feasible the automatic extraction and analysis of faces from digital images (Parkhi et al., 2015). Where human coders employ their innate capacity to assess other people's faces, the corresponding computational analysis requires an explicit integration of multiple steps. In the following sections, we briefly outline the background and current state of the required steps of face detection, extraction, and classification, leading to our choices of RetinaFace (a state-of-the-art face detection network) and ResNet (a well-established convolutional neural network architecture which we use to simultaneously predict age and gender).

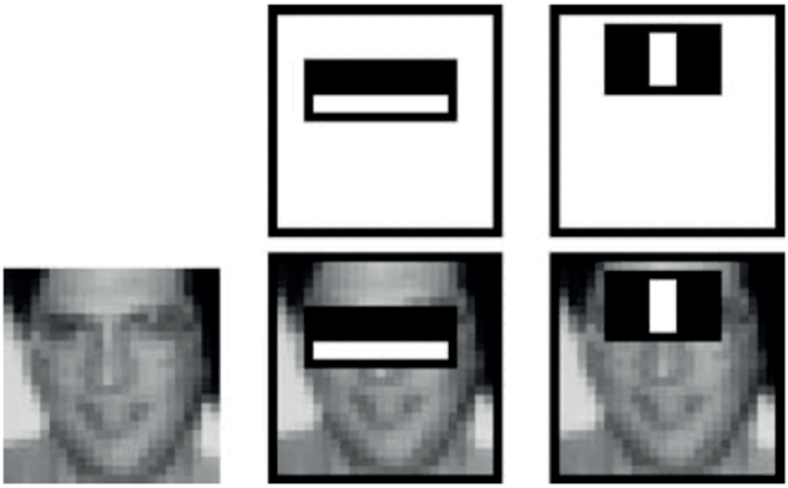


Figure 1: Example of fixed face detection features (Viola & Jones, 2004, p. 144)

Face Detection

The initial step of any strategy for analyzing faces must consist of a method to locate them: A clear distinction between face and no-face areas helps classification by increasing the information to noise ratio, since classifiers need to spend less resources on discarding irrelevant regions. Early approaches to face detection mostly utilized purpose-built patterns (features), such as edges, colors (Hjelmås & Low, 2001), or contrasting rectangles ("Haar Features", Viola & Jones, 2004). These were typically combined with simple machine learning algorithms to select the most useful out of a larger number of features (ibid.). For example, a horizontal bright rectangle framed by dark rectangles may be indicative of a face, as the dark area corresponds to the eyes with brighter skin surrounding them (see figure 1). Since such patterns are geometrically static, they perform worse when faces deviate from a standard frontal angle - e.g., in low-light conditions, with partially covered regions, or for rotated heads (ibid.).

Contemporary methods no longer rely on manual feature design. Deep neural network architectures sidestep the limitations of rigid manual patterns by "learning", i.e., developing their own visual indicators on subsequent layers of abstraction (Taigman et al, 2014; Schroff et al., 2015, compare figure 2). Modern neural networks treat face detection as a regression problem for predicting "bounding boxes" — the coordinates of a rectangle enclosing faces (Girshick, 2015). Each image produces one or multiple candidate regions, in conjunction with a confidence score indicating how well the region

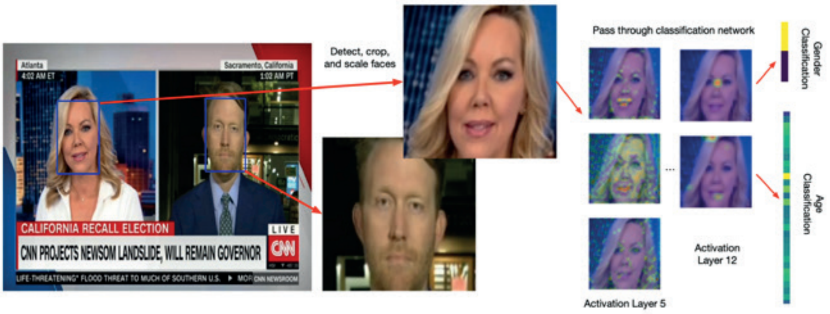

Atlanta
4:02 AM ET
Sacramento, California
1:02 AM PT
CALIFORNIA RECALL ELECTION
LIVE
CNN PROJECTS NEWSOM LANDSLIDE, WILL REMAIN GOVERNOR
CNN
LIFE-THREATENING FLOOD THREAT TO MUCH OF SOUTHERN U.S.
Detect, crop, and scale faces
Pass through classification network
Gender Classification
Age Classification
Activation Layer 12
Activation Layer 5

Figure 2: Detection, scaling and classification of faces from video data. Own data. Activation layers show facial regions which are activated in classification.

matches the criteria for a face. Subsequent improvements in performance have been made by increasing the complexity of the network architectures. In addition to deeper and broader networks (with more and wider layers), some approaches introduce custom cross-links between layers that combine information from different levels of abstraction (Najibi et al., 2017). Consecutive layers typically reduce information and thus decrease in size; combining multiple of them, therefore, allows the detection of faces with varying sizes [2].

In addition to the adaptiveness of neural networks, our method employs a series of additional strategies to enhance detection and classification of faces. So-called "multi-task" architectures (Neverova et al., 2018) integrate multiple distinct branches, each of which is designed to predict a distinct criterion. When combining a branch for gender prediction and a branch for age prediction, as we do, a network can profit from cross-information: Features that describe gender might be relevant to the age-classification task, and vice versa [3]. As mentioned above, earlier computer vision approaches were limited by rigid features that match frontal faces but do less well when people are seen from the side. Instead of learning all possible angles separately, researchers have therefore introduced explicit models of distortion which capture rotation, skew, or even more complex deformations (deformable convolutional networks, Dai et al., 2017). When combined, such strategies allow automated detection of people's faces at a much higher rate than human coders can achieve (Parkhi et al., 2015).

Our analysis employs a pre-trained and validated reference model from the RetinaFace project (Deng et al., 2019), a state-of-the-art face detection method that achieves > 99% precision and best-of-class AOC on standard test datasets (LFW: Huang et al., 2008; WIDER: Yang et al., 2016). Retina Face uses a multi-task architecture to simultaneously learn face classification (face in region: yes/no), regression for bounding boxes, and facial landmarks - i.e., anchors that locate the eyes, nose, and corner of the mouth.

Face alignment

RetinaFace is able to identify faces even when they are not perfectly frontal due to its use of deformable convolutions (Deng et al., 2019). The extracted bounding boxes (rectangles containing the face) still relate to raw regions in the original image and thus remain problematic. Subsequent analysis steps benefit from a normalized (i.e., frontal, centered, and aligned) representation; it therefore makes sense to transform the raw face bounding box. This is done by using the five reference points ("landmarks") identified by the RetinaFace model. Each facial crop is transformed so that the position of its landmarks matches a predefined location [4]. The individual normalized faces are then stored along with their metadata.

Age and gender prediction

Our target variables — age and gender — are estimated in a second step using a separate neural network. In contrast to the task of face detection, there are comparatively few unique network models for gender and especially age. Contributing to this scarcity might be the fact that freely accessible training datasets remain limited. While a manual labeling of apparent gender is certainly feasible, determining people's age presents several challenges:

1. Although apparent age is correlated with biological age, the two are not identical and vary between subjects as well as coders (Agustsson et al., 2017).
2. Images may not be up to date — they could depict a person when she was younger. It is therefore not strictly sufficient to know the age of a person *now*; instead, one needs to know the age of a person *when the picture was taken*. Such potential age bias is especially common with TV broadcasts that are re-run.
3. Facial indicators of biological age vary across ethnicities; training datasets from asian subjects might not lend themselves to robust inferences on caucasian whites.

The influence of training material on neural classifiers is well-known. We therefore evaluated four common datasets used for age and gender classifications. These are: (1) IMDB-Wiki (Rothe et al., 2015; 2018) with approximately 550'000 faces, (2) UTKFace (Zhang et al., 2017), a racially diverse set of 20'000 faces labeled with DEX, (3) Appa-Real (Agustsson et al., 2017), a densely annotated datasets containing real and apparent age of 7591 faces, and (4) fourth “FairFace”, a novel 108,501-entry dataset on age, gender and ethnicity, with special focus on the granularity of ethnicities and balanced class prevalence (Karkkainen & Joo, 2021). We evaluated these datasets with respect to the target material, which (originating from Germany) contains mostly adult white caucasian people. Our primary research interest lies in the minimization of prediction errors; therefore IMDB-Wiki was chosen due to its large size and content proximity to TV material (in fact, many faces from our broadcasts should be contained in the training data), along with UTKFace as a secondary reference (see below).

The classification step of our face analysis predicts age and gender using a joint (multi-task) network (Uchida, 2020). It uses pre-trained convolutional architectures from Keras (Chollet & others, 2015) which are extended by two prediction heads: One two-neuron softmax classifier for gender (yielding a continuous score from 0–1 that we thresholded at .5 to split gender into binary poles), and one 101-neuron softmax classifier for individual ages (spanning 1 to 101 years). Although regression might intuitively make more sense for the estimation of a continuous measure such as age, a classification approach has been shown to significantly outperform its counterpart (Rothe et al., 2015). The models that ultimately provided our data are pre-trained by the author using Keras’ Resnet50; one with the IMDB-Wiki dataset, and one trained on UTKFace. Both were further fine tuned on the training section of the Appa-Real images. This resulted in an MAE (Mean Average Error) of 4.08 years for the apparent age, and 5.3 years for the real age (Uchida, 2020). The benefit of this simple approach is that it offers robust, well-documented performance that can be trivially replicated.

Validation of these classifiers was performed in three stages. First, we (successfully) replicated the Resnet50 (26 million parameters) results reported in the original repository. The findings were then corroborated by training classifiers with both simpler and more complex architectures (see supplementary material).

Second, we compared face classification across both training datasets - IMDB-Wiki and UTKFace. Given that both comprise a disjunct set of faces, we expect systematic errors in one of them to manifest in asymmetric differences between the classifications. This was not the case: Overall agreement was moderately high for both gender [5] (r = .837) and age predictions (r = .765), and furthermore symmetrical across the range of ages and gender values. To obtain the best predictions, we merged both classifications and discarded faces smaller than 64 pixels in size [6].

Third, we employed a set of four student coders and one author to manually verify gender classifications produced by the neural network. The same was not possible for ages due to the effort involved in researching identities of actors along with their real age. A reliability test with 588 images revealed an adequate though imperfect intercoder consistency of Krippendorff's alpha = .803 for classifying female/male/undecidable and alpha = 0.867 for classifying female/non-female. Subsequent coding of 1500 unfiltered images showed an agreement between gender classification from the neural classifier and human coders in 80.7% of the cases, rising to 83, 86.1 and 90.2 percent at thresholds of 64, 100 and 200 pixel width with a confidence above .95. Krippendorff's alpha among all six coders (five humans and the network) was .872 after excluding cases which humans deemed undecidable (no face or not sufficient resolution to decide) [7]. As indicated by the slight increase, the network fit well into the group of human coders. We conclude that the accuracy of our neural network is by no means perfect but seems comparable to human levels of detection. Small, low-information and androgynous faces impose penalties for both, while faces with meaningful sizes offer very high reliability.

Results

Gender Representation

The results of our extensive computational analysis are broadly compatible with existing findings. Figure 3 presents a detailed overview of both measured age and gender across stations — displayed as hourly aggregates (cells) within sample days (x axis). Women on average remained underrepresented on TV, with 6.3 million female faces out of 16 million total (estimated proportion .39, 95% CI: .37-.42). This strong overall bias was mirrored across specific subsamples (figure 3 and 4). Out of all stations, there is not a single year in which the share of female faces reached parity (with a maximum of .455 for VOX in 2017). The same goes for genres, where the most balanced yearly average of .493 (for editorial content on Sat.1 in 2012) barely misses the mark. In fact, there are even few individual hourly slots that exhibit a clear majority of female faces (figure 3, left panel: purple color ranges). In other words: Gender representation remains deficient no matter where we look.

To untangle the individual differences and produce robust estimates from the sample, we computed multilevel logistic models estimating the share of female faces on the level of individual programs (i.e. shows as individual measurements). A first model was run across all genres ("Overall") with varying slopes for years nested in stations and including

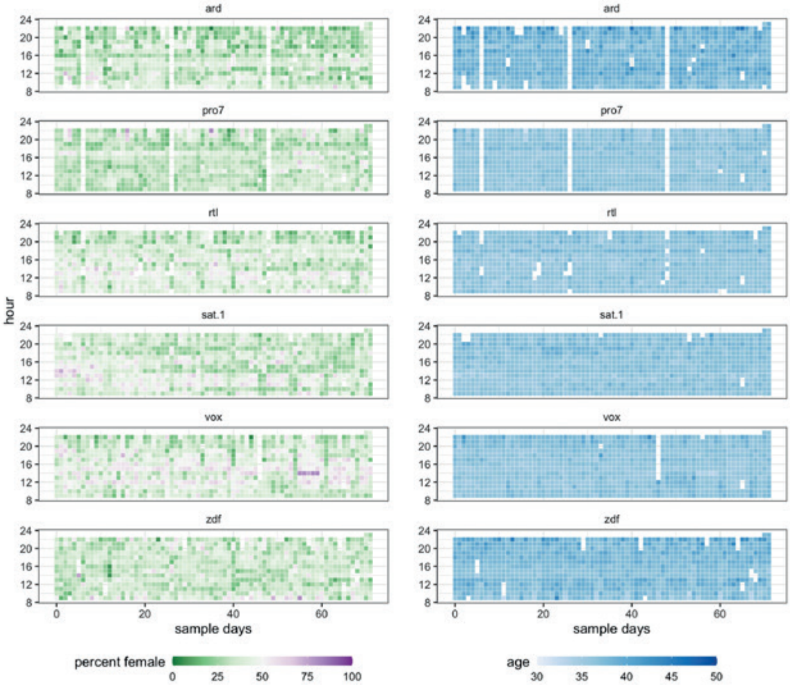


Figure 3: Proportion of female faces as well as age across stations and time. Hourly aggregates over 72 sample days.

an interaction term for public service broadcasters. A second model additionally included program genres as predictors. All models were fitted using lme4 (Bates et al., 2015) and the R statistical software (R Core Team, 2021). Readers may refer to the supplementary material and supplied replication package for the precise model specifications, code and data.

Looking at the model results, there was indeed a significant amount of variation across structural variables – even between stations (see Figure 4). Pro7, a private broadcaster, had the lowest aggregate share of .356, followed by both public service stations (ARD and ZDF) with .367. The top half is represented by VOX (.445), Sat.1 (.414) and RTL (.41). This means that there was a discrepancy of nine percent between the most and least balanced channels. Different genres exhibited even stronger heterogeneity of female representation, with sports ranking lowest (.112 overall, .076 on Pro7 in 2012) followed by news (.3 overall, .229 on Pro7 in 2012), and advertisements taking the top rank (.425 overall, .487 on Sat.1

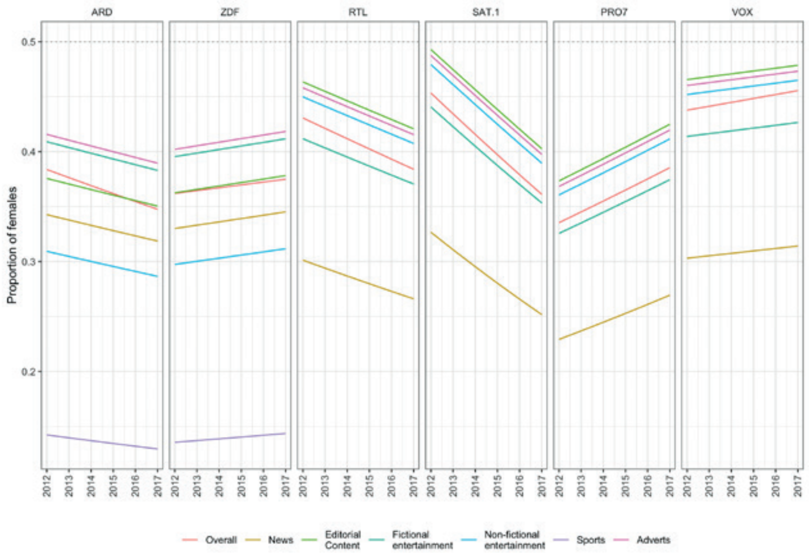


Figure 4: Proportion of female faces across stations, genres and time. Per-station slope estimates (year | station) from multilevel logistic regression model, n = 57,919 shows.

in 2012). Broadly speaking, women were more frequent in commercial and entertainment content (ads, editorial content which includes shows, and fictional movies), and much less visible in non-fictional formats (sports, news, and non-fictional entertainment). In contrast to the station-level and genre-level effects, longitudinal changes were not significant for the share of female faces (B = .00, 95% CI: -.02-.01), despite the sample's span of six years.

A particularly interesting contrast lies in the split between public service broadcasters (PSB, in Germany ARD and ZDF), and private stations. The distinct economic model and regulatory framework of PSBs introduces significant mandates regarding content and diversity, including a quota representation of societal groups in oversight boards (Medienanstalten, 2019). Given these strict measures, we would expect public service broadcasters to offer a significantly more balanced program, in any regard. Results from the multilevel models show, however, that this was not trivially so (Figure 4). Even though there were significantly more women in PSB news programs (.33, 95%-CI .37-.44 vs. .27, 95%-CI .26-.29), these public stations were less diverse in their non-fictional entertainment (.30, 95%-CI .27-.33 vs. .42, 95%-CI .39-.44) and editorial content (.36, 95%-CI .33-.4 vs. .43, 95%-CI .41-.46).

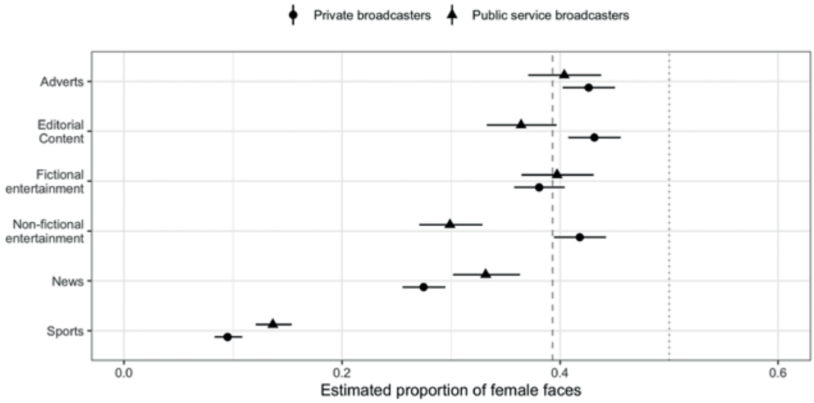


Figure 5: Estimates of gender prevalence by genre and type of broadcaster, results from a multilevel logistic regression model, n = 57,919.

Age Representation

Just like gender, the distribution of age groups on TV is known to be highly distorted with regard to the general population (Prommer & Linke, 2019), with children and elderly people being underrepresented. A phenomenon of particular theoretical interest in this context has been the phenomenon of gendered ageism, whereby a bias in age representation expresses differently for men and women. Figure 6 shows the average age of both genders across stations and genres. In support of the gendered ageism hypothesis, women were indeed consistently younger than their male counterparts across all contexts. Across all programs, the estimated age gap between men and women on television was 4.44 years (95% CI: 3.68-5.19).

Linear multilevel models for the average age of men and women (again nesting yearly slopes in stations and including station type as a predictor, Figure 4) revealed that the gendered ageism effect was furthermore distinct in different program genres, with entertainment programs (both fictional and non-fictional) displaying more age parity than news or sports.

Discussion

Starting from the observation that gender representation is a simultaneously important issue and empirically under-researched, we developed and validated a computational automated analysis of facial age and gender. Drawing on a comprehensive archive of German mainstream TV along with program metadata, we presented a systematic assessment of the diversity

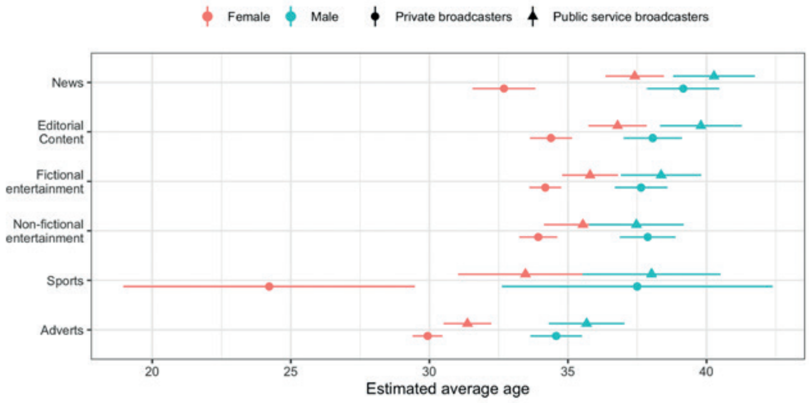


Figure 6: Estimated average age by gender, genre and type of broadcaster, results from a multilevel logistic regression model, n = 57,919.

of approximately 16 million faces shown on six stations across 72 days from 2012 through 2017. Our findings broadly support the overall impression that women remain underrepresented, and furthermore corroborated the assumption that age representation varies by gender, in disfavor of older women ("gendered ageism", Edström, 2018). Hence, by providing a distorted symbolic environment, the media might shape and reinforce both ageism and sexism in its audience. We subsequently explored additional contextual variables: Station, station type (public vs. private broadcasters), and genre. Multilevel models showed that there is a significant degree of variation across those predictors (especially genre); this means that the effective balance of genders which viewers are exposed to is in some cases lower than the grand average. Somewhat surprisingly, we did not identify an increase in gender parity on screen, or a decrease in the gender age gap from 2012 to 2017.

Against the backdrop of research from prior decades, it is not quite surprising to see that the theoretical ideal of equal representation has not been attained so far, and that no change could be measured across the span of six years. But it would be premature to dismiss this as "just the status quo". First, recent years have seen a strong increase in political and societal attention to issues of representation. Online spaces for counter-publics, an increased permeability of mass media, and persistent vocal activism might have made disenfranchisement of societal groups harder to ignore. Second, Germany has passed the sixteenth year of Angela Merkel's chancellorship and thus looks back on an era that brought significant visibility, attention, oxford and respect for women in power. We should therefore assume that the context of our measurements is untypically favorable, and that other countries' TV content lacks these “boosts”. It is disconcerting, alas, to observe that despite these favorable conditions, there is no unequivocal trend towards equal representation. The observation of measurable regressions shows that even if a societal consensus might have been achieved, it has not yet developed to the sort of binding imperative that shapes media and social reality. A final worrying observation lies in the marked bias of news towards men and, where they appear, younger women. News are not merely another genre: they serve as the central point of contact for societal reality, bring together many otherwise disjoint audiences, and represent a particularly prestigious stage for journalists. To see this genre among the least balanced suggests that inequalities are not merely random, but rather follow a societal topology of power.

Note that the normative argument here is not that any individual *show* or even every individual second or frame should be balanced — such a demand would be artistically and pragmatically unrealistic. Instead, the implicit goal is a balance across some sort of aggregate; in most cases a station’s entire program. This grand total standard, as we have shown, is limited in its own right, since the depiction of gender and age vary greatly across genres. Our analysis of structural predictors reveals that viewers may be exposed to strikingly different portrayals – depending on their choice of shows to watch. Any conclusion that wants to offer concrete recommendations towards the regulation of broadcast content needs to make a choice towards the level on which fair representation should be attained. If the goal is to offer societal groups a space in which they can observe their peers, learn from role models, and foster positive self-perception, then a few select shows with sufficient representation would do. The desire for stronger female presence in newscasts, for example, could be partially satisfied by PSBs which exhibit a less biased selection than their private counterparts.

Yet it is not immediately evident that such *conditional fair representation* is a desirable state. Offering groups their “own” space on TV might help *internal* strength but does probably not translate into a general fostering of understanding and tolerance throughout society. For that, an *external* interaction would be needed – they would need to appear more frequently in the **other** groups’ programming. To give a somewhat cliché example: Having an all-female crime show will do little to improve women’s recognition; for that, they would need to be more visible in football broadcasts (and presumably not in the form that they currently are, with an average age that is 10 years younger than mens’). So, there is an argument to be made for some universal benchmark that ensures not only a fair depiction of the demographic makeup of society, but also that this makeup is somewhat hard to elude on mainstream content.

Interestingly, the distinction between public service broadcasters (whose raison d'être is to safeguard a supply of balanced and sensible content) and private stations suggests that strong regulation may only be part of the solution. PSBs do much better in age representation, with smaller age gaps, especially in news and sports. Yet they offer significantly worse gender balance in non-fictional and editorial content. The complex multi-causal factors that influence broadcast programming (including international content, availability of talent, audience demand, economic incentives) will require similarly broad developments to enable a lasting shift in the way that TV portrays people.

Ethical Aspects of Recognition

Precise automated detection and classification of faces is a potentially highly invasive technology with severe ethical implications. Numerous existing real-world applications have drawn scrutiny and justified criticism. Among those are the large face database built by Clearview (Hill, 2020) and liberal use by police forces in the US (Garvie et al., 2016). The technology's risk profile appears severe enough that numerous legislatures are considering prohibitions or at least limits on its use (Schneier, 2020). Opposition against the use of face detection addresses two major issues: (1) Lacking consent in the generation, use and re-distribution of facial data, and (2) biases resulting in the differential treatment of particular groups based on their age, gender, ethnicity, or other attributes.

Our work with television material sidesteps the first aspect, since appearances on TV are regulated and protected by copyright, personality rights and libel law (among others). Furthermore, subjects of coverage have adressees in the form of stations against which they can seek legal recourse. The second aspect should similarly be mitigated by the fact that these biases are in fact our primary research interest and as such explicitly considered. Problematic consequences of deep learning, we feel, arise primarily out of power imbalances between observers and observed, coupled with methodological bias. Far from contributing to these, our paper instead seeks to counteract both through transparent, critical application of existing technologies.

Limitations and potential for follow-up research

With its computational approach, our study goes beyond existing empirical evidence in depth and length. The substantial effort linked to the initial development of computational analyses still incurred some limitations that offer ample opportunities for future enhancements. Linking TV material to structural content metadata limited our sample size to 72 days. Although we could have processed over forty times as much material, it seemed much more important to investigate genres (including advertisements) than to add further data points to an already large sample. Though our detector and classifier networks offer high precision, they rely largely on manifest information that might not represent some invisible truth: Apparent age differs from true age, and perceived gender may differ from self-identified gender. In this regard, the simple dichotomization oversimplifies the reality of gender representation and may in fact contribute to disenfranchisement of people who feel mislabeled by deep learning.

In terms of the substantial questions addressed, our analysis leaves many worthwhile avenues unpursued. *Context* matters, e.g., research has shown that women are often portrayed in stereotyped roles as nonprofessionals or homemakers, wives, or parents (Collins, 2011). This suggests that certain domains of society or specific expertise are reserved for or only appropriate for men. The same goes for *voice*: Who speaks and who remains silent? Addressing such contextual differences in a computational setting would mean analyzing cues for roles as well as voices and active speakers, which is substantially more difficult (and less reliable) than detecting faces (Roth et al., 2020). A related enhancement could enrich the counting of gender with an identification of *identity*, to figure out how the number and popularity distribution of distinct men and women. Doing so is certainly possible through the use of numeric representations of facial features (embeddings). Yet the large number of faces poses a significant risk of misclassification and approaches for labeling identities (such as those employed by Facebook and Google) rely on immense amounts of manual work by their users. Finally, the vulnerability of machine learning to incomplete, biased, and disproportionate training material is well-documented. We took particular care to assess and select suitable existing training material, but were limited by the lack of nuanced resources. Gender as we measured it is particularly superficial, given that it is operationalized as a binary classification of apparent attributes. This issue extends to other attributes not measured, such as ethnicity/race. In addition to the question of categories, we must assume that classification errors are particularly large for minority groups whose faces appear rarely in the training datasets. The creation and curation of large, well-balanced annotation datasets remains a pressing problem for researchers and institutions.

Conclusion

Our computational study has combined significant effort from several endeavors: Its video material stems from a TV archive that, if stored on DVDs, would weigh 12 tons. Processing the sample with deep neural networks consumed between 100 and 200 kilowatts of energy across two weeks, while the manual creation of time-stamped genre annotations must have incurred a costly amount of time as well. Yet the takeaway is that the progress of tools and technology has made it much easier to get precise answers to extremely broad questions. In stark contrast to the television studies of the 20th century, we now have means to eliminate empirical reservations about representativity – even in the face of an immensely increased supply of video material. Two true challenges remain: For scholars, the development and validation of deeper, more meaningful indicators; for society to establish a path that leaves behind the injustices of the past. We would therefore hope that this study is perceived as an encouraging example of the possibilities instead of a showcase of limitations. Future work should continue the development and especially the critical evaluation of deep learning as a tool that is tailored towards theory, to show that we need neither dumb down theory nor tools when pursuing modern computational communication research.

Acknowledgements: We thank Hans-Jürgen Weiß and Torsten Maurer for generously providing the data on the television program structure 2012-2017.

Supplementary material for: Age and Gender Representation on German TV: A Longitudinal Computational Analysis

Availability of Code and Data

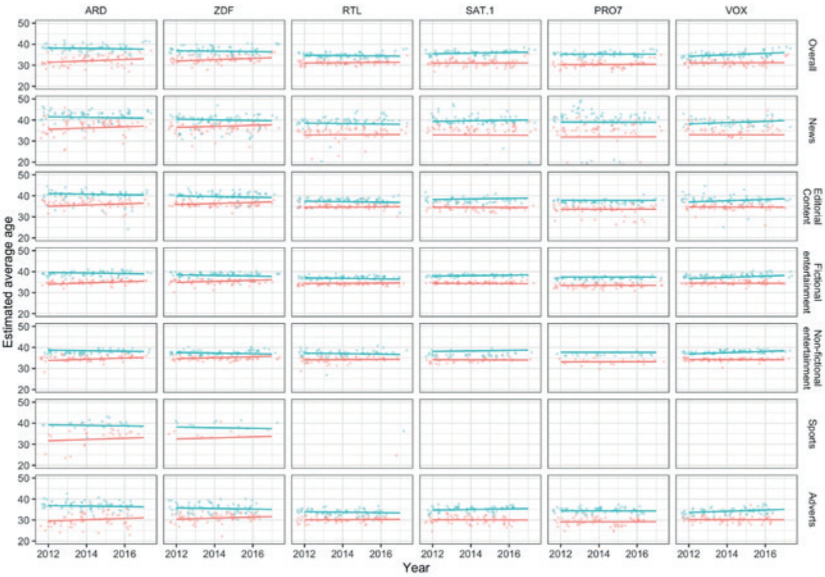
We make all our code and most of our data available in an OSF repository with the URL https://osf.io/fhcgp/. The repository contains two large csv files containing the raw face data (~50 million data points) along with two software packages: One (face.zip) containing the face detection pipeline including models and one (age-gender.zip) containing the R code and aggregate data used for the analysis. Due to the proprietary nature of the ALM program data, we could not include this raw data file, but worked to ensure full replicability otherwise. Both packages contain an extensive readme file, which we include below.

Validation of ResNet Classification Architecture

As indicated in the paper, we verified the overall plausibility of the employed pre-trained ResNet model by training our own networks using the same training dataset (IMDB-Wiki) but different well-known architectures: MobileNetV2 (3.5 million parameters), EfficientNetB1 (7.5 million parameters), EfficientNetB3 (12 million parameters) as well as the very large InceptionResNetV2 (56 million parameters). These extremely different networks achieved a gender classification accuracy of .915, .913, .912 and .914, and an age classification accuracy of .085, .113, .112, and .108 respectively Our replication results show that accuracy is fairly saturated for modern convolutional networks on these tasks; that it is only weakly dependent on the depth and breadth of the network, and that highly complicated architectures (e.g. InceptionResNetV2) do not offer a significant benefit over their simpler peers.

Visualization of Age across Station, Genre and Time

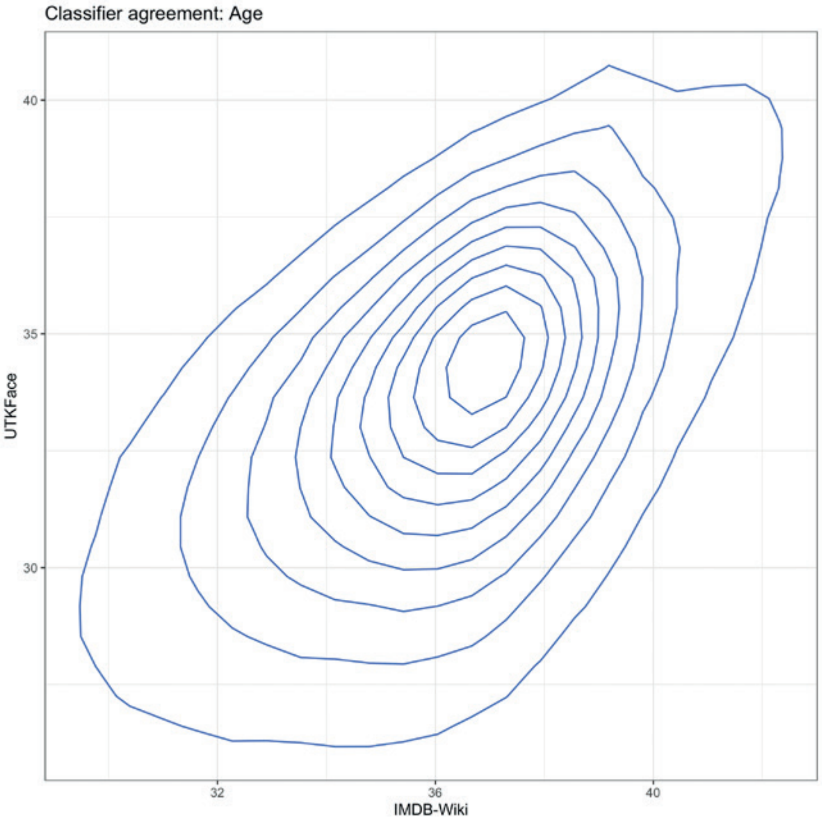
The results of our most complex model estimate slopes for genres within each station and over time. The visualization below displays all of these components in a single, comprehensive graph.

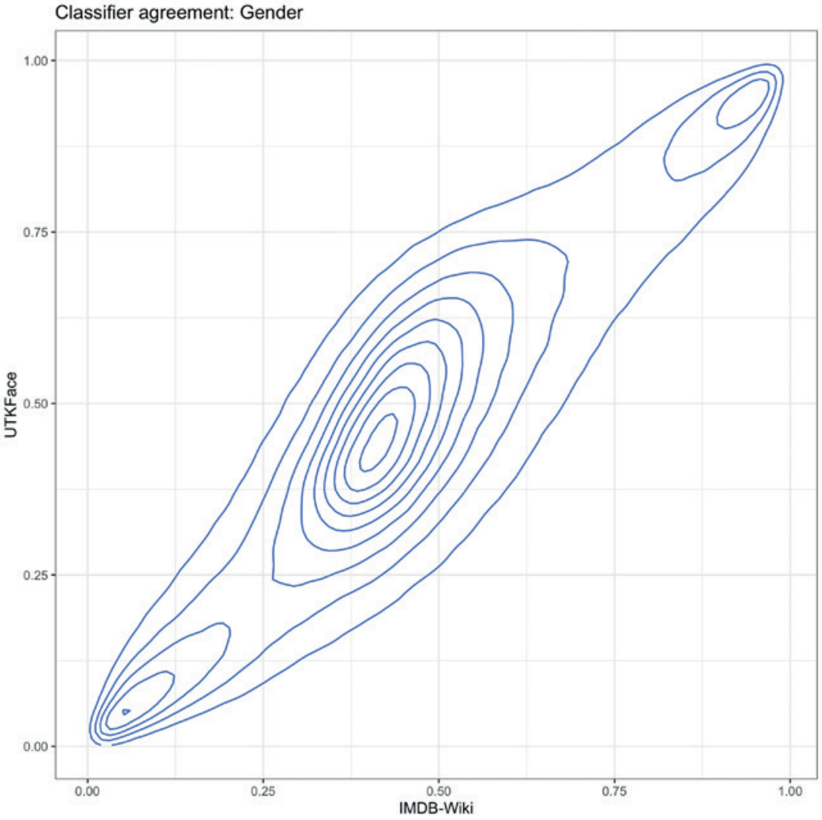


Validation of IMDB and UTK Classifiers

As we argued in the paper, the classifier networks trained on either the IBDM or UTK dataset offer a satisfactory correlation, meaning they do not exhibit a strong, systematic and asymmetric bias. Below we show density plots that indicate correlations for both age and gender. As is to be expected, the correspondence of age is lower than that for gender (due to the higher uncertainty in the underlying measurement and the concept itself).

As expected, the size and detection confidence of faces influence those correlations: Faces narrower than 100 pixels showed a correlation of r=.631 between both classifiers on gender, but those between 100 and 200 agreed with r=.789. A similar trend shows for confidence, where faces detected at .9 show a gender correlation of r=0.597, while those at .99 correlate with r=.837. Picking arbitrary but sensible thresholds, we opted to remove faces with confidence below .95 and less than 64 pixels in size. At this cutoff, both classifiers correlate with r=.869 on gender and r=.797 on age. Since classification errors appear to be symmetrically distributed, we merged values to their mean for subsequent analyses.





Readme File for the Face Extraction and Analysis Pipeline

This repository contains code for the paper Age and Gender Representation on German TV: A Longitudinal Computational Analysis. It implements a performant analysis pipeline to extract faces from video files, scale them into a normalized representation, and estimate their age and gender.

We draw on pre-existing code from a series of projects:

- The retinaface model from insightface implements a high-performance face detection and landmark detection model
- To maintain comparability, we employ two pre-trained ResNet architectures whose performance on IMDB-Wiki and UTK was comparable to various self-trained models
- Fast parsing of video files is performed with decord, a gpu-based video loading library for python

The pipeline is able to efficiently process large volumes of video material and produces estimates of age and gender of found faces along with metadata (detection confidence, bounding boxes):

Technical prerequisites and replicability

Many of the python libraries used in this repository are not trivially usable on Windows and macOS computers. Please use a linux system (preferrably Ubuntu 20.04) to run the code. We also provide a Dockerfile to create a self-contained environment in which everything is set up and ready to use.

Note that this repository contains the original model files used along with a sample video, and therefore uses git lfs (for large file storage). You will need to install git-lfs to clone it.

Graphics cards

Deep learning is rather resource-intensive and therefore benefits from computing resources provided by graphics cards (gpus). The code in this repository should work without special hardware, but has been used exclusively with Nvidia gpus. If you do not have one installed on your system, set GPU_ID = -1 in the file face/retina.py.

Despite their performance benefits, graphics cards complicate replicability due to the complexity of the involved software and tooling. Specific versions of python libraries need specific versions of Nvidia's own deep learning libraries (CUDA), which in turn need specific versions of drivers. The paper's results were obtained from a Ubuntu 20.04 system running CUDA 11.2. Building the docker image requires at least the same base version for the driver. Furthermore (and somewhat unintuitively), some libraries are not part of the container but instead need to be installed in the host system - notably nvcuvid files for gpu-accelerated decoding of videos. To make sure the drivers are matching (or exceeding) our baseline, add the official driver repository:

```
sudo add-apt-repository ppa:graphics-drivers
```

and then install the required apt packages for nvidia-470. A list of packages installed on our system can be found in apt-nvidia-libraries.txt.

Docker image

To use the docker image, install docker to your linux computer (and nvidia-docker to make use of Nvidia gpus), and run docker build -t face-analysis -f face-gpu.Dockerfile. Then launch the image with docker run –rm -it –gpus all -it face-analaysis /bin/bash. Note that building the image requires a gpu, and the default docker executor needs to be set to nvidia. To do so, add the line "default-runtime": "nvidia" to /etc/docker/daemon.json and restart the docker daemon. If the system does not have a gpu, use docker build -t face-analysis -f face-cpu.Dockerfile instead.

Library versions

Keeping with best practices, we have set up a requirements.txt file which lists the versions of python libraries we used. Where possible, we have directly included the relevant source files in our repository - this goes for the architecture of our classifier (in wide_resnet.py), the architecture of the face extraction model (retinaface.py), the face normalization (retinaface_align.py) and the models (R50 for retinaface, weights.28-3.73.hdf5 for the IMDB-wiki-classifier and weights.29-3.76_utk.hdf5 for the UTK-classifier, all found in the models directory).

Note that due to disjoint requirements, installing precisely these versions may fail in the future, especially when using pip's new dependency lookup. Because mxnet (the deep learning framework used by retinaface) and tensorflow/keras (the deep learning framework used by the age/gender classifier) track different versions of libraries (numpy in particular), resulting conflicts can become unresolvable. In that case, it should still be possible to use a separate environment for mxnet-cu112 and tensorflow respectively.

Running the code

Once all prerequisites are installed, video files may be placed in the input directory of this repository. Then run python3 pipeline.py. For each video, a tar file (without compression) is created containing the extracted faces. These tar files are then processed by the classifiers. The result are per-video TSV files in the output directory, which can be further analyzed in R. There are a couple of optional command line options:

- input sets a directory from which to load video files
- output sets the directory to which output data is written
- frames N enables writing of annotated full frames (not just faces) to the output directories. These frames have identified faces annotated with a bounding box.

To run the included example, set up the docker container and run the following steps:

```
docker run –rm –gpus all -it face-analysis /bin/bash python3 pipeline.py –input diagnosis –output output –frames 100
```

You should find the result in the output folder. We have included the sample video segment along with the results in diagnosis.

Some notes on performance

Large-scale analyses of video material is bound to encounter several bottlenecks which have been worked around in this repository. We want to make these observations available in the hope that future research may benefit.

Disk I/O

Video material takes up significant disk space; even our modest sample reached approximately 6TB of raw recordings. More importantly though, the individual images of extracted faces can easily overwhelm a disk system, either by using up all available inodes or by slowing down directory access. Possible solutions are

1. a tiered directory structure that stores images in subfolders named by tiered hash bytes (123/456/789/123456789.jpg),
2. directly analyzing faces without storing them on disk (this requires that all applicable neural networks fit into the gpu memory at once), and
3. our approach: Storing face images in an uncompressed tar per video file.

Decoding videos

Decoding videos from their compressed source formats makes up a surprisingly large amount of the resources consumed by the overall pipeline, especially if sources are in h.264 or h.265 or other modern formats and/or exist in high resolution. Processing a one-hour segment can take up to several minutes, even on contemporary machines.

Decoding video is much faster when done by a modern gpu - that is why we include the decord library for gpu-accelerated video loading. Since gpu-based decoding has quite a few prerequisited, it's disabled by default. You can enable it by setting the decoding context in face/retina.py: Switch vr = VideoReader(str(video_path), ctx=cpu(0)) to vr = VideoReader(str(video_path), ctx=gpu(0)).

Some care is still warranted in deciding for or against this technology.

1. gpu-based video decoders are not identical to software-based ones and do encounter more errors.
2. To decode a video, the file needs to be transferred from main memory to the gpu - a bottleneck that may slow down especially serial decoding of many small files.
3. Decoding on the gpu requires free memory; video files may be too large to be processed, especially when neural models are loaded at the same time.
4. In the past, there were several bugs in decord that would slow down random access to a video's frames; as a result, CPU parsing was faster.
5. When a fast CPU is available, it may be more efficient to perform video decoding on the CPU and leave the GPU free to simultatenously analyze results.

Parallelization and handling gpu memory

Complex pipelines often include multiple neural networks (as ours does - retinaface and two wide resnets). But modern architectures also tend to be very large and require a lot of gpu memory. In between running these networks, the memory needs to be freed, or scripts will crash with an out of memory-error. It is sometimes still possible to run the entire pipeline in one go - by spawning child processes with multiprocessing, which in turn import the required libraries that set up the network. Note that this strategy may fail for mxnet when the library is not imported before creating additional processes. Finally, memoization (caching finished results) is crucial; we included a simple file-based memoization utility in utils.py.

Optimal performance furthermore may require parallelization - i.e. running multiple neural networks at once on the same gpu, or spreading computation across multiple gpus. We did run our bulk analysis with parallel processes, but the details are too finnicky to offer them pre-made. The best strategy for efficient handling of very large volumes should be running each step of the pipeline as a separate process: The built-in memoziation will prevent repeated ingestion of already processed files, and each stage can independently watch for and work on new results. Do not hesitate to contact us for further details.

Readme File for the Analysis Project

This repository contains analysis code for the paper Age and Gender Representation on German TV: A Longitudinal Computational Analysis. It is the second part of the replication package: The first repository (face) contains the python-based code for extracting faces from video files and classifying age and gender.

Prerequisites

The datasets used here are fairly large (with approximately 50 million lines). Our data preparation code is mostly optimized for legibility instead of efficiency. As a result, more than 32GB of memory (ram) are needed to run the entire pipeline.

Non-public data

Our analysis contains propietary content analysis coding from ALM (ALM-2012-2018.sav), which we cannot publish in full. This omission unfortunately renders a complete replication impossible. We have, however, retained the entire code used to prepare, transform, merge and analyze all datasets, so that these steps may be scrutinized.

The two final aggregated datasets used in the analysis are included in this repository: data/agm_program-aggregates.tsv.gz contains per-show aggregates of age, gender share, gender counts and counts per age bracket, along with the station, date, ALM genre classification and a cryptographic hash of the title. The second file, data/hourly_age_gender.tsv.gz, contains per-hour aggregates of age and gender plus date and station; these are used to present a broad overview of variation in the data.

Reproducibility

Changing versions of R packages can lead to differences between published and repeated results. We therefore include a snapshot of all package versions from the R library packrat.

Data preparation

As the R code might not be self-explanatory to all readers, we offer a brief high-level description.

There are two main input files originating from the previous neural network analysis: data/2021_age_gender_combined.tsv.gz contains classification results for all identified faces. For robustness, we employ two distinct classifiers (one trained on IBDM-Wiki data, the other on the UTK dataset); each face therefore has one line with age and gender estimates for each of them.

The second file, data/2021_meta_combined.tsv.gz, contains extraction metadata, specifically the size of the face in pixels and the detection confidence.

The first script, 1_prepare_data.R loads both files, joins them and pivots the data so that each line represents a single face, with classifier-specific age and gender estimates in columns.

It then parses the face filenames - which contain station, date, and precise time information - into their own columns. We then perform a date range join, attaching program information from the content analysis dataset to each face. This creates additional columns for title, genre, start and end date of the program.

Next, the individual datapoints for each face are annotated with age and gender groups and aggregated, to produce per-show entries with mean age, mean gender share etc. We then recode genre information into english, descriptive and coherent categories.

Analysis

Our analysis primarily comprises lme4 multilevel models and ggplot visualizations, with some aggregated raw data added to visualize variance. Figure 6, a heatmap, is derived directly from raw data.

Notes

[1] https://osf.io/fhcgp/

[2] A simpler and popular way of ensuring size invariant detection is to perform analysis across a stack (called “pyramid”) containing multiple resized versions of the same input (Viola & Jones, 2004)

[3] The challenge in multi-task learning lies primarily in simultaneous optimization of mul-tiple distinct loss functions. For an overview, see Ruder(2017).

4. See figure 2 for an example of the scaling procedure that aligns faces to appear per-fectly symmetrical, frontal pictures.
5. Since the dataset comprises around 25 million cases, we omit reporting of p values which are approximating zero.
6. See supplementary material for details.
7. There are multiple options for comparing human and neural codings, given the neural network's floating point value for female and human coders' female, male or undecidable. Split-ting the machine classification at .5 into a dichotomous variable and comparing this to the hu-man three-value annotations yields a lower alpha of 0.798 — still rather good given the penalty that there was no "undecidable" option for the machine. Although it would be feasible to introduce a middle option for the automatic classification (at, say .4 to .6), this decision would be arbitrary. The alpha reliability reported in the text (which excludes cases where humans were unsure) thus seems the most pragmatic measurement.

REFERENCES

Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., & Rothe, R. (2017). Apparent and real age estimation in still images with deep residual regressors on Appa-Real Database. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG 2017*), 87–94. https://doi.org/10.1109/FG.2017.20

Bandura, A. (2001). Social cognitive theory of mass communication. *Media Psychology*, 3, 265–299.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Becker, J., Maurer, T., Spittka, E., Benert, V., Weiß, H.-J., Beier, A., Hölig, S., Hasebrink, U., Volpers, H., Bernhard, U., Dammler, A., Blase, R., Langer, C., Walter, M., Feldhaus, D., Frerichmann, N., Holsten, C., Hein, D., #x98;die#x9C; medienanstalten – ALM GbR (Hrsg.), & VISTAS Verlag. (2018). *Content-Bericht 2017 Forschung, Fakten, Trends.*

Baumann, S., & de Laat, K. (2012). Socially defunct: A comparative analysis of the underrepresentation of older women in advertising. *Poetics*, 40(6), 514–541. https://doi.org/10.1016/j.poetic.2012.08.002

Campbell, D. E., & Wolbrecht, C. (2006). See Jane run: Women politicians as role models for adolescents. *The Journal of Politics*, 68(2), 233–247. https://doi.org/10.1111/j.1468-2508.2006.00402.x

Cann, D. J., & Mohr, P. B. (2001). Journalist and source gender in Australian television news. *Journal of Broadcasting & Electronic Media*, 45(1), 162–174. https://doi.org/10.1207/s15506878jobem4501_10

Chollet, F. & others. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

Collins, R. L. (2011). Content analysis of gender roles in media: Where are we now and where should we go? *Sex Roles*, 64(3–4), 290–298. https://doi.org/10.1007/s11199-010-9929-5

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. *2017 IEEE International Conference on Computer Vision* (*ICCV*), 764–773. https://doi.org/10.1109/ICCV.2017.89

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. (2019). RetinaFace: Single-stage dense face localisation in the wild. *ArXiv:1905.00641* [*Cs*]. http://arxiv.org/abs/1905.00641

Desmond, R., & Danilewicz, A. (2010). Women are on, but not in the news: Gender Roles in local television news. *Sex Roles*, 62(11–12), 822–829. https://doi.org/10.1007/s11199-009-9686-5

Edström, M. (2018). Visibility patterns of gendered ageism in the media buzz: A study of the representation of gender and age over three decades. *Feminist Media Studies*, 18(1), 77–93. https://doi.org/10.1080/14680777.2018.1409989

Eisend, M. (2010). A meta-analysis of gender roles in advertising. *Journal of the Academy of Marketing Science*, 38(4), 418–440. https://doi.org/10.1007/s11747-009-0181-x

Fox, C. (2018). The scully effect: I want to believe in stem. https://seejane.org/wp-content/uploads/x-files-scully-effect-report-geena-davis-institute.pdf. Accessed 15 Sep 2019.

Fryberg, S. A., & Townsend, S. M. (2008). The psychology of invisibility. In G. Adams, M. Biernat, N. R. Branscombe, C. S. Crandall, & L. S. Wrightsman (Eds.), *Commemorating brown: The social psychology of racism and discrimination* (pp. 173–193). American Psychological Association.

Garvie, C., Bedoya, A., & Frankle, J. (2016). *The perpetual line-up. Unregulated police face recognition in America. Georgetown Law Center on Privacy & Technology, October 18, 2016*.

Geena Davis Institute on Gender in the Media (2015). Cinema and Society: Shaping our worldview beyond the lens. Investigation on the impact of gender representation in Brazilian films. https://seejane.org/wp-content/uploads/cinema-and-society-investigation-of-the-impact-on-gender-representation-in-brazilian-films.pdf. Accessed 15 Sep 2019.

Gerbner, G., & Gross, L. (1976). Living with television: The violence profile. *Journal of Communication*, 26(2), 173–199.

Gerbner, G., & Signorielli, N. (1979). *Women and minorities in television drama 1969–1978*. The Annenberg School of Communication.

Girshick, R. B. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision* (*ICCV*). https://doi.org/10.1109/ICCV.2015.169

Haraldsson, A., & Wängnerud, L. (2019). The effect of media sexism on women's political ambition: Evidence from a worldwide study. *Feminist Media Studies*, 19(4), 525–541. https://doi.org/10.1080/14680777.2018.1468797

Hether, H. J., & Murphy, S. T. (2010). Sex roles in health storylines on prime time television: A content analysis. *Sex Roles*, 62(11–12), 810–821. https://doi.org/10.1007/s11199-009-9654-0

Hill, K. (2020, January 18). The secretive company that might end privacy as we know it. *The New York Times*. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

Hjelmås, E., & Low, B. K. (2001). Face detection: A survey. *Computer Vision and Image Understanding*, 83(3), 236–274. https://doi.org/10.1006/cviu.2001.0921

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008, October). *Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments*. Workshop on faces in "real-life" images: Detection, alignment, and recognition. https://hal.inria.fr/inria-00321923

Karkkainen, K., & Joo, J. (2021). FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–1558.

Kessler, E.-M., Schwender, C., & Bowen, C. E. (2010). The portrayal of older people's social participation on German prime-time TV advertisements. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 65B(1), 97–106. https://doi.org/10.1093/geronb/gbp084

Küchenhoff, E. & Bossmann, W. (1975). Die Darstellung der Frau und die Behandlung von Frauenfragen im Fernsehen: Eine empirische Untersuchung einer Forschungsgruppe der Universität Münster. Kohlkammer.

Liebler, C. M., & Smith, S. J. (1997). Tracking gender differences: A comparative analysis of network correspondents and their sources. *Journal of Broadcasting & Electronic Media*, 41(1), 58–68. https://doi.org/10.1080/08838159709364390

Lind, F., & Meltzer, C. E. (2020). Now you see me, now you don't: Applying automated content analysis to track migrant women's salience in German news. *Feminist Media Studies*, 1–18. https://doi.org/10.1080/14680777.2020.1713840

Lukesch, H. & Schneider, I. (2004). *Das Weltbild des Fernsehens: Eine Untersuchung der Sendungsangebote öffentlich-rechtlicher und privater Sender in Deutschland*. Band 2: Theorie - Methode - Ergebnisse. Roderer.

Matthes, J., Prieler, M., & Adam, K. (2016). Gender-Role portrayals in television advertising across the globe. *Sex Roles*, 75(7–8), 314–327. https://doi.org/10.1007/s11199-016-0617-y

McCombs, M. E. (2007). *Setting the agenda: The mass media and public opinion* (Reprinted.). Polity Press.

Medienanstalten. (2019). *Interstate treaty on broadcasting and telemedia* (*Interstate Broadcasting Treaty*).

Najibi, M., Samangouei, P., Chellappa, R., & Davis, L. S. (2017). SSH: Single stage headless face detector. *2017 IEEE International Conference on Computer Vision* (*ICCV*), 4885–4894. https://doi.org/10.1109/ICCV.2017.522

Neverova, N., Alp Güler, R., & Kokkinos, I. (2018). Dense pose transfer. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11207, pp. 128–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-01219-9_8

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *Proceedings of the British Machine Vision Conference 2015*, 41.1-41.12. https://doi.org/10.5244/C.29.41

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge Univ. Press.

Prieler, M., Kohlbacher, F., Hagiwara, S., & Arima, A. (2015). The representation of older people in television advertisements and social change: The case of Japan. *Ageing and Society*, 35(4), 865–887. https://doi.org/10.1017/S0144686X1400004X

Prommer, E., & Linke, C. (2019). *Ausgeblendet: Frauen im deutschen Film und Fernsehen*. Herbert von Halem Verlag.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., & Pantofaru, C. (2020). Ava active speaker: An audio-Visual dataset for active Sspeaker detection. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 4492–4496. https://doi.org/10.1109/ICASSP40776.2020.9053900

Rothe, R., Timofte, R., & Gool, L. V. (2015). DEX: Deep EXpectation of apparent age from a single image. *2015 IEEE International Conference on Computer Vision Workshop* (*ICCVW*), 252–257. https://doi.org/10.1109/ICCVW.2015.41

Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2), 144–157. https://doi.org/10.1007/s11263-016-0940-3

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *ArXiv:1706.05098* [*Cs, Stat*]. http://arxiv.org/abs/1706.05098

United Nations, Department of Economic and Social Affairs, Population Division (2017).*World Population Ageing 2017 - Highlights* (ST/ESA/SER.A/397).

Schneier, B. (2020, January 20). Opinion | We're Banning facial Recognition. We're missing the point. *The New York Times*. https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). *FaceNet: A unified embedding for face recognition and clustering*. 815–823. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html

Signorielli, N. (2004). Aging on television: Messages relating to gender, race, and occupation in prime time. *Journal of Broadcasting & Electronic Media*, 48(2), 279–301. https://doi.org/10.1207/s15506878jobem4802_7

Sink, A., & Mastro, D. (2017). Depictions of gender on primetime television: A quantitative content analysis. *Mass Communication and Society*, 20(1), 3–22. https://doi.org/10.1080/15205436.2016.1212243

Smith, S. L., Pieper, K. M., Granados, A., & Choueiti, M. (2010). Assessing gender-related portrayals in top-grossing G-rated films. *Sex Roles*, 62(11–12), 774–786. https://doi.org/10.1007/s11199-009-9736-z

SWR. (2018). *SWR Mitschnittdienst*. https://swrservice.de/mitschnitt/

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). *DeepFace: Closing the gap to human-level performance in face verification*. 1701–1708.

Tuchman, G. (1978). The symbolic annihilation of women by the media. In G. Tuchman, A. K. Daniels, & J. Benét (Eds.), *Hearth and home: Images of women in the mass media* (pp. 3–38). Oxford University Press.

Turner, J. S. (2011). Sex and the spectacle of music videos: An examination of the portrayal of race and sexuality in music videos. *Sex Roles*, 64(3–4), 173–191. https://doi.org/10.1007/s11199-010-9766-6

Uchida, Y. (2021). *Yu4u/age-gender-estimation* [Jupyter Notebook]. https://github.com/yu4u/age-gender-estimation (Original work published 2017)

Van Bauwel, S. (2018). Invisible golden girls? Post-feminist discourses and female ageing bodies in contemporary television fiction. *Feminist Media Studies*, 18(1), 21–33. https://doi.org/10.1080/14680777.2018.1409969

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2), 137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

Weiß, H.J., Maurer, T. & Beier, A. (2020). ARD/ZDF-Programmanalyse 2019: Kontinuität und Wandel. Media Perspektiven 6/2020, 226-225.

WDR. (2021). *Geschäftsfelder—I-O - Mitschnittservice*. WDR Mediagroup. https://wdr-mediagroup.com/geschaeftsfelder/i-o/mitschnittservice/

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). WIDER FACE: A face detection benchmark. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5525–5533. https://doi.org/10.1109/CVPR.2016.596x