How can computer vision widen the evidence base around on-screen representation

Raphael Leung
Creative Industries Policy and Evidence Centre, Nesta, UK
raphael.leung@nesta.org.uk

Abstract

There is strong demand for more complete and better data around diversity in the screen industries. Focusing on on-screen diversity and representation in the UK, the evidence base around representation on-screen has been narrow so far. Diversity evaluation needs to consider more than on-screen presence – it should also consider prominence and portrayal. In this position paper, the ethics of applying computer vision to study on-screen characters is discussed via a conceptual framework of on-screen diversity metrics. Computer vision should be applied to identify character occurrences, rather than demographic classification. An illustrative example of measuring character prominence using a short video clip is shown. It concludes with four areas of applications where adopting computational methods can create a measurably more inclusive and representative broadcast landscape.

1 Context

Representation is defined as "how the media *represents* aspects of reality and/or staged realities. [It] tends to portray people, groups, organisations and topics in structured, [...] often ideologically predisposed ways. This constitutes the media's symbolic power." Representation also has a second meaning of "how the media represents their public(s)" which includes "how race and gender are represented (quantitatively) in the workforce" as well as covering "social and economic class and interest groups" [Fourie, 2010]. Increasingly, screen industry bodies are formally acknowledging and addressing diversity and representation. There is reflected by recent developments in the UK: Bafta forming a diversity steering group; the British Film Institute (BFI) strengthening its Diversity Standards as a contractual requirement for funding and eligibility requirement, as well as renewed inclusion and diversity commitments from the UK's biggest broadcasters (the BBC, Channel 4, ITV, and Sky) in 2020.

1.1 State of the evidence base

There are two big initiatives in the UK regularly collecting evidence across broadcasting channels and different demographics about representation on screen: Project Diamond by the Creative Diversity Network, beginning in 2016, and the Office of Communications (Ofcom)'s annual diversity in television broadcasting reports, beginning in 2017. While informative, existing continual data collection misses out on important parts of the picture. There are three areas where more evidence could be particularly impactful.

First, current evidence misses out on some key aspects of diversity entirely, even though it is possible to collate this data. Much of the current evidence base focuses mostly on presence (whether a character appears on screen) and very little on prominence (how much screen time a character has, and how centred or foregrounding they are).

Second, data coverage is low and uneven across demographic groups. Project Diamond had an average response rate of 28% across five contributing broadcasters [Creative Diversity Network, 2019]. There is no direct knowledge of how representative the remaining 72% of the production landscape is. A review of Diamond data found the "inevitable possibility of reporting bias due to non-response as a consequence of the low response rate" [NatCen, 2018]. Ofcom similarly highlight "data gaps" and "insufficient collection" of some demographic data [Ofcom, 2019]. There is comparatively much less data on some underrepresented and minoritised groups, too. BFI's evidence review found "sexual orientation and religion and belief were seldom explored in detail" in the 80 studies examined [BFI, 2016]. Only 1% of productions meeting the new BFI's Diversity Standard on on-screen representation (Standard A) do so via gender identity compared to 63% for gender and 50% for race and ethnicity [BFI, 2020]. For a production that ended many years ago, it is also difficult to conduct retrospective self-reporting.

Third, existing methods to annotate on-screen demographic traits (manually) is limited. The laborious manual coding approach is adopted by state-of-diversity reports on BBC channels [Cumberbatch *et al.*, 2018] as well as in-depth case studies in media analysis, e.g. [Markov and Yoon, 2020; Mastro and Stern, 2003]. The coverage is limited, usually representing a snapshot of available data, constrained by what the researcher can feasibly annotate.

1.2 Reasons for lack of evidence

The application of computer vision to on-screen diversity measurement has been slow. There are computer vision and audio processing applications: from the Geena Davis Insti-

tute [GD-IQ, 2015], the French National Audiovisual Institute [Doukhan *et al.*, 2018] and tools like Ceretai [Ceretai, 2020] - which focus on gender representation. Different reasons may account for the lack of evidence more widely and across different underrepresented demographics.

First, some researchers consider facial attribute identification (relating to emotion, race, gender and age) a solved problem [Wang *et al.*, 2019]. Second, algorithmic audits have found disparate impacts in commercial face detection models, with lower accuracy rates for darker-skinned females [Buolamwini and Gebru, 2018; Grother *et al.*, 2019]. In February 2020, Google removed the gendered labels such as 'man' and 'woman' from its Cloud Vision API, "given that a person's gender cannot be inferred by appearance" [Lyons, 2020] The perpetuation of bias is an especially amplified concern: there rightfully should be careful consideration of fairness and transparency criteria before models are productionised. Third, conventional diversity form categories do not map well to categories in labelled data. For example, existing public datasets of faces only tend to have three to four types of race labels [Kärkkäinen and Joo, 2019] but the 2011 Census had 18 ethnic groups (grouped into 5 including "other") and the Office for National Statistics is consulting on how to expand to more inclusive categories [GOV.UK, 2020]. While there are approaches to re-balance the data to better mimic real-world distribution of age, gender and skin colour, e.g. [Yang *et al.*, 2020], a perfectly balanced dataset does not absolve all moral responsibility. Demographic inference using classification assumes clean categorisations when in reality demographic identities are heavily laden with social context, e.g. see social construction of gender [Risman, 2004], ethnicity [Ford and Harawa, 2010] and disability [Goering, 2015]. Skin reflectance classification cannot directly inform ethnic representation. Similarly, detection of walking aids e.g. [Weinrich *et al.*, 2014] would miss out entirely on non-visible disabilities. Fourth, there are issues with copyright and access to data. Licensing of the relevant components of broadcast data is complex. While the UK has a copyright exception to text and data mining, there are specific conditions under which it can be applied. Difficulty of data access and fear of infringement may deter interested researchers.

2 A framework for measuring on-screen representation

A conceptual framework of on-screen representation measurements is presented here. The framework is most useful for measurements of on-screen groups who fall under Equality Act protected characteristics [Hepple, 2010]. This is because most reporting on workforce diversity (on- and off-screen) examine protected characteristics: e.g. the Diamond report covers gender, gender identity, age, ethnicity, sexual orientation and disability. However, the framework can be used for non-protected characteristics too like socioeconomic diversity. It asks three key questions about the method being used to measure representation on-screen:

- What aspect of diversity does the measurement capture?
- Who is being tracked on-screen, and what potential biases exist in the method being used?
- How are character occurrences identified?

This first question helps make explicit the aspect of diversity that is being measured. They generally fall under one of '3P's – presence, prominence and portrayal. Measures of presence focus on determining if someone is shown on-screen or not. Much current evidence concentrates on this aspect. However, it is optimal to expand beyond presence, for example, studying characters' prominence, relative to each other in a programme. Computer vision has potential to contribute here. The last 'P' is portrayal. This analyses the authenticity of portrayals, and what narratives and stereotypes the story may be subverting or perpetuating. The 3Ps be captured in a wide range of metrics, for example 'presence' by the cast make-up by gender, ethnicity, etc; 'prominence' by the duration of screen time, likelihood to appear as a solo face, most central nodes in a character network, etc; and 'portrayal' by the emotion of faces or the words uttered, likelihood of appearing next to particular objects like weapons or drinks.

The second part of the framework helps make explicit who is being tracked, and raises key considerations around whether this is possible with a purely visual approach, and any ethical concerns. The framework groups the considerations under feasibility (can we) and ethics (should we), considered in parallel. Under feasibility, the considerations include if a human face is shown (it could be computer-generated), the level of occlusion (a mask can cover a face or a character can speak off-screen), whether a demographic group exhibits identifiable traits using just visual of face, and any available proxies (like a regional accent). Under ethics, there are considerations specific to character monitoring. A character's demographic can be left ambiguous (no ground truth) or not fit into conventional diversity categories. It is important to acknowledge if the method captures intersectionality, especially as there is an identified need for more insights into the intersectional dynamics of underrepresented groups on-screen [Nwonka, 2020]. Interdisciplinary efforts – from computer science, digital humanities, anthropology, critical theory, data justice – are required for thoughtful answers.

The third part of the framework covers how measurements are generated. To minimise the chance of reinforcing unfair bias, computationally inferring characters' or people's demographics is not advised, but computer vision can still be usefully applied to identify character occurrences. There is often a trade-off between accuracy and speed for face detection. [Huang *et al.*, 2017] Still, slower face detection models will be quicker than manual annotation. The face tracks can be clustered to identify occurrences of the same character, but the optimal model parameters will vary by the type of programme being analysed. A show with recurring frontal faces filmed under similar conditions allows for easier clustering and identification of most to all appearances of the same character, e.g. [Tapaswi *et al.*, 2019]. However, programmes with higher variation in viewpoint, more crowds (smaller faces) and darker lighting are more difficult to yield auto-generated representation metrics. More research is needed to study causes for dropped detections. Investigations into optimal downsampling of frame rates, missed detection thresholds for different demographic groups and implementation of human-in-the-loop systems would be particularly relevant.

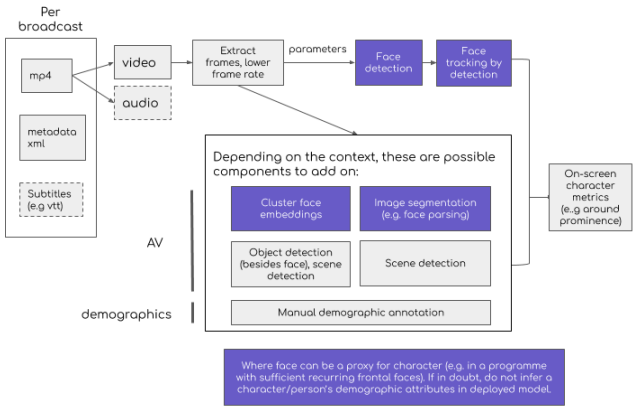

Figure 1: Pipeline for generating on-screen representation metrics

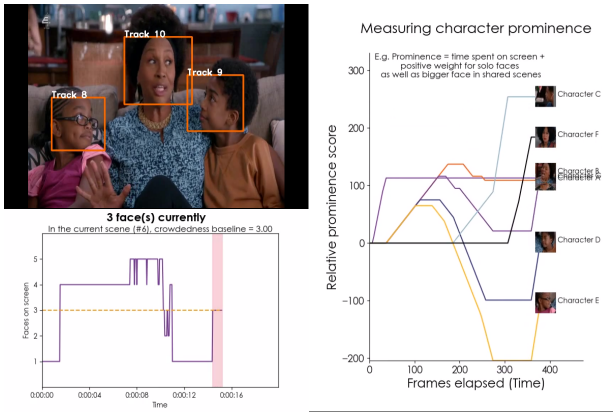

Figure 2: An illustrative example of prominence metrics

3 Illustrative example

An episode from Black-ish Season 1 (originally aired on ABC) was used to test the feasibility of generating character prominence metrics to study representation on-screen. The broadcast was made available via the Box of Broadcasts (BoB) archive and the licensed use of data via the Educational Recording Agency. BoB is an on-demand TV and radio service for education [Learning on Screen, 2020].

After downsampling the video to 1 FPS, S3FD was used for face detection [Zhang *et al.*, 2017], pyscenedetect for scene detection [Castellano, 2018], and IOU tracking by detection [Bochinski *et al.*, 2017]. The pipeline is shown in Figure 1. In the illustrative example of a 20 second clip shown in Figure 2, prominence was defined as time spent on screen (as a clear and big enough face). A metric was generated that combines different aspects of prominence: positive weights were added for a larger face in a sea of faces, as well as for longer periods of screen time when the face is the only face on screen. Characters were labelled manually. In the demo clip, the grandmother (played by Jenifer Lewis) was the most prominent followed by other characters in the scene. This ranking of prominence is based entirely on the visual information and takes quicker than real-time to process.

There are limitations around subtle aspects of prominence: a character can be a scene stealer with limited screen time, or say or do something which is highly impactful. The metric can be extended by applying active speaker detection [Roth *et al.*, 2020]. Despite the limitations, the illustrative example shows that it is feasible to measure relative character prominence in TV programmes without computational inference of demographic attributes. If scaled up, the resulting ranking of relatively more (and less) prominent characters could generate novel insights to understand prominence, a less studied aspect of representation on-screen. To identify all occurrences of the same character, recent techniques such as face clustering with unknown number of characters [Tapaswi *et al.*, 2019] or one-shot retrieval [Nagrani *et al.*, 2018] can help. But current benchmark datasets, which research methods are evaluated against, often include just one to two TV programmes. In reality, faces on screen have greater variation in viewpoint, head pose, face size, skin reflectance and lighting. A better understanding of how the methods scale up, specifically addressing the ethical and logistical barriers to wider deployment of such methods across different types of filmed content, will be beneficial. More annotated datasets can be shared and data standards for on-screen representation can be developed.

4 Applications to widen the evidence base

Recognising that inclusion is more than capturing numerical measurements, and through the lens of the four proposed roles for computing in social change [Abebe *et al.*, 2020], here are some applications to widen the evidence base around on-screen representation in its broader social context.

First, computer vision can be used, supplemented with manual review, to generate more frequent and richer data about representation. Through plugging evidence gaps, computer vision can generate closer to real-time insights of on-screen representation, in other words, acting *"as diagnostic"*. Measurements can prompt rethinking about the stories which are told and funded. Second, the content producer can potentially use richer data on character prominence to create new product features for viewers to look for major and minor characters. Computer vision here can generate additional value for viewers and fans, perhaps uncovering new dimensions to understand representation, acting *"as synecdoche"*. Third, faster processing allows content to be analysed before a show airs. Currently, diversity evidence is gathered long after the broadcast date. Processing an episode post-production for character prominence can allow concerns to be addressed or discussed upstream, during commissioning, screen-writing or editing between series. Through explicit specification of inputs and goals, computing can promote as useful, a particular lens – i.e. that relative prominence is an informative aspect of representation – hence acting *"as formalizer"*. Finally, research partnerships can be formed to better understand the models under which content holders can open up broadcast data for research, clarifying the limits of computer vision and its most appropriate interpretations – acting *"as rebuttal"*, so that as cultural heritage collections are more commonly treated as data [Ziegler, 2020], they can be responsibly opened up to answer important social research questions.

References

[Abebe *et al.*, 2020] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260, 2020.

[BFI, 2016] BFI. Workforce diversity in the UK screen sector: evidence review. Technical report, British Film Institute, 2016.

[BFI, 2020] BFI. BFI Diversity Standards initial findings. Technical report, British Film Institute, 2020.

[Bochinski *et al.*, 2017] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

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

[Castellano, 2018] Brandon Castellano. Pyscenedetect, 2018.

[Ceretai, 2020] Ceretai. Ceretai. https://ceretai.com/, 2020. Accessed: 2020-11-12.

[Creative Diversity Network, 2019] Creative Diversity Network. Diamond - The Third Cut Report. Technical report, Creative Diversity Network, 2019.

[Cumberbatch *et al.*, 2018] Guy Cumberbatch, Andrea Bailey, Victoria Lyne, and Sally Gauntlett. On-screen diversity monitoring: BBC One and BBC Two 2018. Technical report, Cumberbatch Research Group, 2018.

[Doukhan *et al.*, 2018] David Doukhan, Géraldine Poels, Zohra Rezgui, and Jean Carrive. Describing gender equality in french audiovisual streams with a deep learning approach. *VIEW Journal of European Television History and Culture*, 7(14), 2018.

[Ford and Harawa, 2010] Chandra L Ford and Nina T Harawa. A new conceptualization of ethnicity for social epidemiologic and health equity research. *Social science & medicine*, 71(2):251–258, 2010.

[Fourie, 2010] Pieter J Fourie. *Media studies: Media history, media and society*, volume 2. Juta and Company Ltd, 2010.

[GD-IQ, 2015] GD-IQ. The Reel Truth: Women Aren't Seen or Heard. An Automated Analysis of Gender Representation in Popular Films. Technical report, Geena Davis Institute on Gender in Media, 2015.

[Goering, 2015] Sara Goering. Rethinking disability: the social model of disability and chronic disease. *Current reviews in musculoskeletal medicine*, 8(2):134–138, 2015.

[GOV.UK, 2020] GOV.UK. List of ethnic groups. https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups, 2020. Accessed: 2020-11-12.

[Grother *et al.*, 2019] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. National Institute of Standards and Technology, 2019.

[Hepple, 2010] Bob Hepple. The new single equality act in Britain. *The Equal Rights Review*, 5:11–24, 2010.

[Huang *et al.*, 2017] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.

[Kärkkäinen and Joo, 2019] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[Learning on Screen, 2020] Learning on Screen. Teaching and research: BoB for AI. https://learningonscreen.ac.uk/teaching-and-research/research-resources/bob-for-ai/, 2020. Accessed: 2020-11-12.

[Lyons, 2020] Kim Lyons. Google AI tool will no longer use gendered labels like 'woman' or 'man' in photos of people. The Verge, February 2020.

[Markov and Yoon, 2020] Čedomir Markov and Youngmin Yoon. Diversity and age stereotypes in portrayals of older adults in popular American primetime television series. *Ageing & Society*, pages 1–21, 2020.

[Mastro and Stern, 2003] Dana E Mastro and Susannah R Stern. Representations of race in television commercials: A content analysis of prime-time advertising. *Journal of Broadcasting & Electronic Media*, 47(4):638–647, 2003.

[Nagrani *et al.*, 2018] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–88, 2018.

[NatCen, 2018] NatCen. Independent review to determine the quality of the Diamond dataset. Technical report, The National Centre for Social Research, 2018.

[Nwonka, 2020] Clive James Nwonka. Race and ethnicity in the UK film industry: an analysis of the BFI diversity standards. 2020.

[Ofcom, 2019] Ofcom. Diversity and equal opportunities in television: Monitoring report on the UK-based broadcasting industry. Technical report, Ofcom, 2019.

[Risman, 2004] Barbara J Risman. Gender as a social structure: Theory wrestling with activism. *Gender & society*, 18(4):429–450, 2004.

[Roth *et al.*, 2020] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP*

2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4492–4496. IEEE, 2020.

[Tapaswi *et al.*, 2019] Makarand Tapaswi, Marc T Law, and Sanja Fidler. Video face clustering with unknown number of clusters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5027–5036, 2019.

[Wang *et al.*, 2019] Cunrui Wang, Qingling Zhang, Wanquan Liu, Yu Liu, and Lixin Miao. Facial feature discovery for ethnicity recognition. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(1):e1278, 2019.

[Weinrich *et al.*, 2014] Christoph Weinrich, Tim Wengefeld, Christof Schroeter, and Horst-Michael Gross. People detection and distinction of their walking aids in 2d laser range data based on generic distance-invariant features. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 767–773. IEEE, 2014.

[Yang *et al.*, 2020] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.

[Zhang *et al.*, 2017] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017.

[Ziegler, 2020] SL Ziegler. Open data in cultural heritage institutions: Can we be better than data brokers? *Digital Humanities Quarterl*, 14(2), 2020.