ForecasterFlexOBM: A MULTI-VIEW AUDIO-VISUAL DATASET FOR FLEXIBLE OBJECT-BASED MEDIA PRODUCTION

Davide Berghi[*1], *Craig Cieciura*[*1], *Farshad Einabadi*[*1], *Maxine Glancy*[*2],
Oliver C. Camilleri[1], *Philip Foster*[1], *Asmar Nadeem*[1], *Faegheh Sardari*[1], *Jinzheng Zhao*[1],
Marco Volino[1], *Armin Mustafa*[1], *Philip J. B. Jackson*[1], *Adrian Hilton*[1]
[1] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, U.K.
[2] BBC R&D, London, U.K.

ABSTRACT

Leveraging machine learning techniques, in the context of object-based media production, could enable provision of personalized media experiences to diverse audiences. To fine-tune and evaluate techniques for personalization applications, as well as more broadly, datasets which bridge the gap between research and production are needed. We introduce and publicly release such a dataset, themed around a UK weather forecast and shot against a blue-screen background, of three professional actors/presenters – one male and one female (English) and one female (British Sign Language). Scenes include both production and research-oriented examples, with a range of dialogue, motions, and actions. Capture techniques consisted of a synchronized 4K resolution 16-camera array, production-typical microphones plus professional audio mix, a 16-channel microphone array with collocated Grasshopper3 camera, and a photogrammetry array. We demonstrate applications relevant to virtual production and creation of personalized media including neural radiance fields, shadow casting, action/event detection, speaker source tracking and video captioning.

Index Terms— Object-based Media (OBM) Production, Virtual Production, Professional Quality Real Dataset

1. INTRODUCTION

Systems that use machine learning techniques to provide users with personalized recommendations for *what* to watch or listen to are becoming increasingly prevalent across services that deliver professionally produced content via internet protocol (IP), i.e., downloading or streaming, usage of which continues to increase [1]. The bidirectional communication inherent in IP-based media results in an opportunity to also personalize *how* audio-visual media is presented — to optimize for the particular needs/preferences of individuals.

Machine-learning based techniques can be applied to the task of personalizing *how* media is presented. These include providing alternative viewpoints, relighting, action/event detection, speaker source detection and automatic video captioning. Development of such techniques typically takes place using research-oriented datasets, which offer extensive examples. In comparison to professionally produced media, the eventual target for integration of these techniques, research-oriented datasets typically lack in several attribute areas. In contrast, whilst often of a higher quality, professionally produced material comes with different limitations. Table 1 summarizes the differing attributes between professionally produced media and research datasets and highlights the contribution of ForecasterFlexOBM which we describe and publicly release in this paper.

Table 1: Generalized summary of varying attributes when comparing research-oriented datasets and reusing production material to ForecasterFlexOBM

Attribute	Research Datasets	Production Material	ForecasterFlex-OBM
Acting Quality	Often low	High	High
Capture Quality	Varied	High	High
Realism/Narrative	Low	High	High
Asset Availability	High	Low[1]	High
Range of Sequences	Varied	Low	Moderate
Range of Capture Techniques	Varied	Low	Moderate
Unrestricted by Licensing	High	Low	High[2]

[1] In pre-distribution form. [2] For non-commercial use.

ForecasterFlexOBM is an object-based (OB) audio-visual (AV) media dataset themed around a presentation of a UK weather forecast aiming to bridge the gap between professional production and research-oriented datasets — for fine-tuning and evaluation of techniques for personalized media production. ForecasterFlexOBM consists of sequences derived from production-typical scripts, both linear and modular, and performed by professional actors/presenters — one male and one female (English) and one female (British Sign Language) — and directed by a professional director and assistant director to maximize production quality. Capture of the sequences utilized volumetric video and production-typical audio techniques. Alongside the production-oriented attributes, the dataset also includes specific research-oriented sequences and capture techniques.

Examples of research-oriented datasets, relevant to per-

[*] Equal contributions.

sonalization applications, include 3DVHshadow [2], for shadow casting, and Charades [3] and MultiTHUMOS [4] for action/event detection. 3DVHshadow offers images of 25 synthetic humans. In comparison to computer graphics used in film and television today, they are less realistic and less detailed. The capture approaches in Charades and MultiTHUMOS, crowd-sourcing or publicly available amateur videos, can differ from professionally produced media in attributes of acting and capture quality or realism/narrative. They do, however, offer a wider range of sequences — for the specific applications described — than professionally produced media and with fewer license restrictions. ForecasterFlexOBM offers a range of both production-oriented and research area specific sequences featuring the same professional actors and capture techniques.

Related datasets, although not explicitly for refining techniques for personalization, with comparable attributes bridging research and production but with differing purposes and approaches, are available. These include *Old School* [5], a multi-camera dataset derived from a fictional game show and shot in 8K resolution, produced at the level of professional television production and presented to develop computational cinematography techniques; *Tragic Talkers* [6], a dataset of excerpts from Shakespeare's *Romeo and Juliet* performed by student actors captured with co-located camera and microphone arrays was intended for development of techniques for OB media production; and the *S3A Object-Based Audio Drama Dataset* [7], which consists of three professionally produced audio-only radio dramas distributed in the object-based Audio Definition Model (ADM) format [8]. In comparison, ForecasterFlexOBM provides a greater range of types of sequences, both production and research-oriented, and a more diverse range of both video and audio capture techniques.

Section 2 provides additional background to developments in personalized media. Section 3 presents the design, sequences, and capture processes. The dataset is applicable to the development of a variety of techniques. Section 4 illustrates five such techniques and discusses their applicability to personalized media before concluding in Section 5.

2. BACKGROUND

Prior research in the area of personalization of media presentation has included development of prototype personalized experiences, such as by the British Broadcasting Corporation (BBC) [9, 10] and Netflix [11] [12]. In these examples, personalization affordances [13] are provided to users enabling them to alter the experience to best suit their unique circumstances. This personalization typically takes one of two forms: sequential/narrative — in which chapters, segments or sub-segments are extracted, annotated, and reordered — and layered — in which media layers, such as foreground or background video; audio objects, tracks or channels; and text or graphical overlays; are enabled or disabled, repositioned or altered. One well-known layered personalization affordance is subtitles or captions, available since 1979 in the UK. Subtitling affords comprehension of the narrative to the user who cannot adequately hear the audio component of an experience.

Tools have been developed for authoring personalized experiences [9, 14, 15, 11]. Studies have been performed to assess the feasibility of these tools in mainstream production [16] and what developments of new production systems is required if personalized media is to become a mainstream offering [13]. Other studies considered the user perspective [17].

3. DATASET DESCRIPTION

3.1. Dataset Design

The design evolved from replicating assets from an existing personalized media prototype, *Forecaster* [18], with updated methods, to a multi-faceted design encompassing varying functionalities that might deliver diverse personalized experiences. A range of sequences/scenes were scripted within two main categories: *production-* and *research-focused*. For each scene, multiple takes were captured.

In the production-focused sequences, combinations of linear or modular scripts, production (three-point) or diffuse lighting conditions, and actors (female-English, male-English, female-BSL) were captured. The linear script was generated from a transcription of the Forecaster example [18], which was originally scripted by a professional weather presenter and then edited for brevity. The modular script consists of a set of standard phrases and variables for weather conditions and locations enabling short sequences to be recombined to personalize the duration and specificity of the forecast.

In the research-focused scenes, the following sequences were captured: *Gestures*, presenters gestured to eight background segments with a return to a neutral pose; *Motion and gestures*, gesturing whilst walking and improvising dialogue; *Action sequences*, performing varied sequences of actions.

Media featuring professional actors is often subject to strict licensing agreements. The use of the likeness of actors for machine learning-oriented activities, especially around generative AI, is currently a contentious issue due to the rapid evolution of technologies, limited public understanding, and a lack of explicit consent in the creation of legacy datasets and a range of media outputs. In advance of the shoot, the actors were provided with an information sheet and contribution agreement that detailed the uses to which the captured data might be used (only for non-commercial purposes), including for generative AI, so that we could ensure informed consent. The actors were paid the feature daily rate recommended by the UK's largest performers' union.

The dataset includes foreground video and audio. To illustrate usage, we provide an example combination of linear scene with a background map generated from OpenStreetMap.org data used under an Open Database License.

Table 2: Summary and enumeration of the scenes for action and light conditions, actor, no. of takes, no. of takes captured with the AVA rig, and duration. Action type: linear (Lin), modular (Mod), gesture (Gest), motion&gesture (M-G), actions (Acts), music presenter (Mus). Lighting: 3-point production (Prod), diffuse (Diff). Actor: male (M-), female (F-) and English (ENG), British Sign Language (BSL)

Scene	Action	Light	Actor	# takes	# tk AVA rig	Duration (s)
4	Lin	Prod	F-BSL	3	0	200;203;205
5	Lin	Prod	M-ENG	4	3	205;206;204;207
6	Lin	Prod	F-ENG	3	3	202;206;201
9	Lin	Diff	F-BSL	3	0	208;210;215
10	Lin	Diff	F-ENG	3	3	213;209;212
11	Lin	Diff	M-ENG	3	2	211;110;212
14	Mod	Diff	F-ENG	2	2	209;211
15	Mod	Diff	M-ENG	2	1	211;210
16	Gest	Diff	M-ENG	1	1	100
17	Gest	Diff	F-ENG	1	1	84
18	M-G	Diff	M-ENG	1	1	168
19	M-G	Diff	F-ENG	1	1	171
20	Acts	Diff	F-ENG	1	1	109
21	Acts	Diff	M-ENG	1	1	112

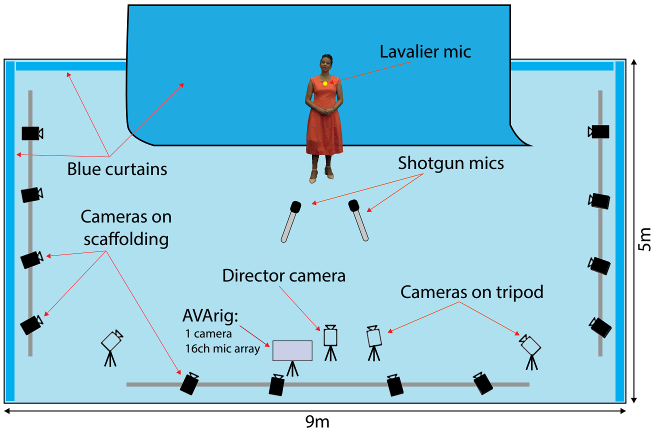

Fig. 1: Schematic top-down view of the studio

The dataset includes 14 scenes, often provided in multiple takes, for a total of 29 sequences, as summarized in Table 2. Combined, ForecasterFlexOBM provides over 1h30min of audio-visual data. Each of the subjects was also captured using a 64-camera photogrammetry system to obtain high-res static 3D textured models. These models can be used as reference when creating digital human characters, e.g., via Epic's MetaHuman platform, or can be used to support research in a wide range of computer vision and graphics tasks.

3.2. Capture Setup

Sequences were captured in a multi-camera studio comprised of 16 synchronized Blackmagic URSA 4k broadcast cameras at 30fps, against a blue chroma background. They were placed in a horseshoe inward-facing fixed configuration to capture the front and sides of the presenter. The cameras were rigged on scaffolding above the capture volume and four were supported by tripods at eye level to provide variations in height perspective. Figure 1 includes a schematic top-down view of the studio showing the camera arrangement. While 15 of these cameras were locked as described, one, frontal, was maneuvered by the director. We refer to it as the *director camera*. All fixed cameras are calibrated in a single coordinate frame, including the intrinsic parameters. Production-focused sequences were shot with both diffuse and 3-point lighting setups to provide versions that could later be re-lit and to also provide versions representative of typical production practices. Lighting data is provided in the form of spherical, high dynamic range (HDR) images captured by two devices: Spheron's SpheroCam HDR, and FARO Focus$^{\text{m}}$ 70.

Production audio, i.e., audio intended for audience reproduction and not for, e.g., sound source localization, was captured via a pair of shotgun microphones and a wireless lavalier microphone attached to each actor's clothing as in Figure 1, a typical production arrangement. The shotgun microphones were a Schoeps CMIT 5 and a Sennheiser MKH 418-S. Post-processing was applied to reduce background noise and to set loudness to –27 LUFS. Raw recordings are also provided.

In addition to the multi-view and production audio data, the dataset was captured from an Audio-Visual Array (AVA) Rig. The AVA rig is a custom device consisting of a quasi-linear 16-element microphone array and 11 Grasshopper3 cameras fixed on a flat perspex sheet of size 0.6mx0.4m mounted on a standard production tripod. It was designed to advance research in the field of audio-visual immersive production [19, 6]. Only the central camera has been employed. Since each sensor is in a fixed relative position to the other sensors, the AVA rig represents a co-located platform with a shared coordinate system across the audio and visual modalities. This makes the recordings ideal for audio-visual or audio-only active speaker localization. The camera's resolution is 2448×2048p, 30 fps. Audio is sampled at 48 kHz, 24 bits. The microphone array presents a horizontal aperture of 450 mm and a vertical aperture of only 40 mm. This provides higher spatial resolution along the azimuth direction than elevation. Horizontally, the microphones of the quasi-linear array are log-spaced for strong coverage of the speech frequency band from 500 Hz to 8 kHz. A clapper-board was employed at the beginning of each take to enable manual alignment of the audio and visual streams. Scripts were played back via an autocue. Playback speed was synchronized plus manual live adjustment to ensure natural performances.

4. APPLICATIONS

In this section, we demonstrate how ForecasterFlexOBM is used for evaluation purposes in five research tasks in personalized media production.

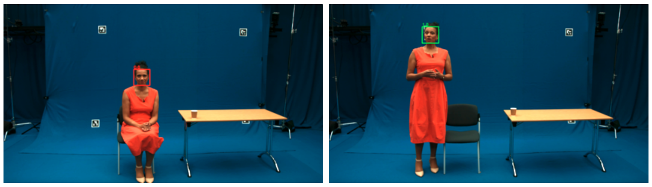

Fig. 2: Example of active speaker detection where red bounding boxes are silent faces, green are actively speaking

4.1. Active Speaker Detection

ForecasterFlexOBM enables the generation of speaker tracking data useful for a variety of OBM applications. This can either be achieved by audio-visual active speaker detection (ASD) methods that leverage monaural audio [20, 21], or by employing the microphone array of the AVA rig [22]. With the microphone array, this can be addressed with audio-only algorithms, e.g., MUSIC [23], or with audio-visual solutions [24]. We provide 2D ASD pseudo-labels generated with the open-source TalkNet model by Tao et al. [21]. They consist of per-frame face bounding boxes associated with the respective voice activity confidence as in Figure 2. ASD pseudo-labels are generated on the director camera and the AVA rig camera.

Tracking data of the speakers can be employed in immersive production. Mohd Izhar et al. [25] predicts 3D tracking data to drive the listening directivity of a spatial beamformer. The filtered speech signals are then replaced by the production audio spatialized at the predicted positions [19].

Another use of the speaker tracking data concerns the idea of an *automated editing system* [5]. AI is leveraged to automate cinematography operations such as shot framing, where pan-tilt-zoom maneuvers are applied in-camera and/or in post-production cropping, and shot sequencing, where a timeline of different shots or camera angles is generated to tell the story in an aesthetically pleasing way. Research from BBC R&D in this direction and proposed an automated editing system called *Ed* [26]. Ed leverages face and speaker detections to drive an automated shot re-framing system.

4.2. Action/Event Detection

Action or event detection aims to determine the boundaries of different actions/events occurring in a video [27]. Event detection can benefit in many aspects, e.g., archiving and editing processes or personalized media distribution. For example, it can be used to find the optimal moment for advertisements, or it can be leveraged for personalized video summarization.

There are several action detection datasets focusing on different types of actions, e.g., Charades [3] include daily activities, such as drinking and eating, while MultiTHUMOS [4] contains sports activities. Large datasets for action detection present significant challenges due to the labor-intensive and complex annotation process. These issues can become more severe with dense multi-label scenarios. The same is valid with datasets for media production. To overcome this, one solution is to train a model on existing datasets and then use machine learning techniques, such as zero-shot learning [28] or few-shot learning [29], to transfer the learned knowledge by the model for media production. ForecasterFlexOBM can be leveraged to evaluate the performance of such approaches as it includes 4 video samples that have been annotated for the action/event detection task. Figure 3 presents and compares the performance of PAT [27] and MS-TCT [30] on one test sample of ForecasterFlexOBM. PAT and MS-TCT have been pre-trained on Charades [3], which includes the same actions as in ForecasterFlexOBM, and then fine-tuned on one video sample of ForecasterFlexOBM.

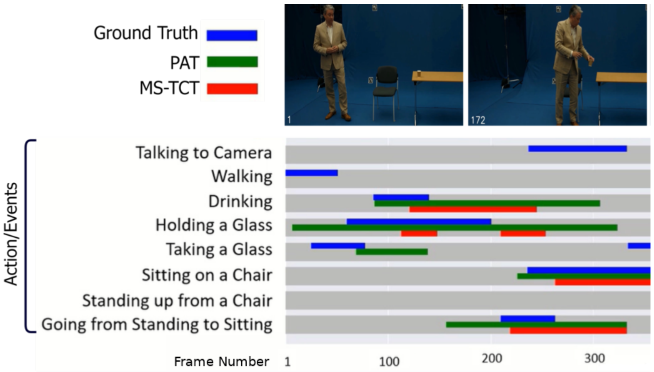

Fig. 3: Action/event predictions by PAT [27] and MS-TCT [30] on a video of ForecasterFlexOBM

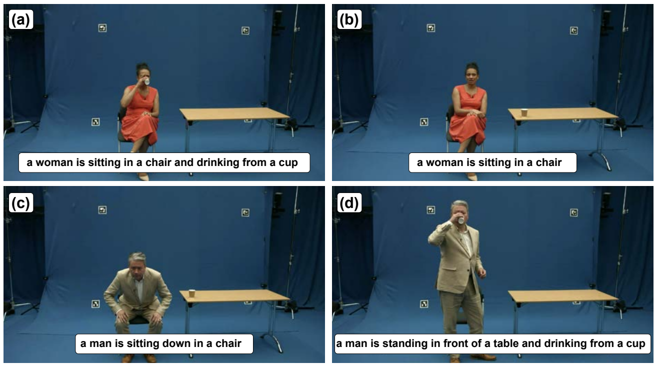

Fig. 4: Captions generated with SEM-POS [31]: (a) a woman is sitting in a chair and drinking from a cup (b) a woman is sitting in a chair (c) a man is sitting down in a chair (d) a man is standing in front of a table and drinking from a cup

The 4 annotated videos of ForecasterFlexOBM contain temporal annotations for 10 action classes, e.g., 'Talking to camera', 'Smiling', etc. There is also high overlap among the instances of different action categories in each video.

4.3. Video Captioning

Video captioning is a crucial task that involves generating natural language descriptions for video content, and enhancing the accessibility and user experience of media. In the context of OBM production, ForecasterFlexOBM serves as a valuable resource for evaluating video captioning algorithms.

In video captioning, the first step involves data prepro-

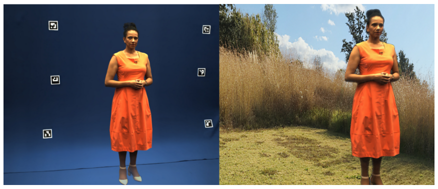

Fig. 5: Scene reconstructed by two separate NeRFs (left); presenter rendered from a novel view and 3D translated into a new environment rendered from a spherical image (right)

cessing. This includes extracting features from the video frames and generating embeddings for the labels. The annotated labels in ForecasterFlexOBM provide pairs of video frames and associated captions, serving as ground truth for evaluation. With the prepared data, the video captioning method SEM-POS [31], which achieves state-of-the-art results on benchmark video captioning datasets MSVD [32] and MSRVTT [33], is evaluated on ForecasterFlexOBM (see Figure 4). SEM-POS uses four parts of speech (POS) blocks and a global-local fusion block (GLFB) to generate grammatically and semantically correct captions. The POS blocks are determinant + subject, auxiliary verb, verb, and determinant + object. The GLFB fuses local and global semantics from the POS blocks and the visual features from the video.

Pre-trained SEM-POS model can be applied in OBM production workflows. It can be integrated to automatically generate captions, contributing to the creation of accessible and personalized media content. Moreover, the model's capability for content indexing and retrieval can be harnessed, enabling users to search for specific actions or dialogues within the archive. In summary, ForecasterFlexOBM not only provides a valuable resource for advancing video captioning algorithms but can also be extended to audio-visual question answering (AVQA) [34] in the realm of OBM production, promoting improved media accessibility and user experience.

4.4. Modular Neural Radiance Fields

ForecasterFlexOBM provides calibrated, multi-view footage and a separately captured background environment. We leverage this modular capture and a clear segmentation to implicitly represent discrete scene components using Neural Radiance Fields (NeRFs) [35]. These can be manipulated and recomposed to create custom scenes for use in immersive and personalized OBM production.

NeRFs typically assume a simple emission-absorption light transport model: $\hat{C} = \sum_{i=1}^{N} W_i(\sigma(x(t))) \cdot c(x(t), \hat{d})$. $\hat{C}$ is the color of a camera ray, x, which is sampled N times, has an origin x_o, direction $\hat{d}$, and is parameterized by t: $x(t) = x_o + t\hat{d}$. $c(x(t), \hat{d})$ is the color at a 3D point as seen from direction $\hat{d}$ while W_i can be interpreted as an analogue to a differential color weight. These weights are functions of volume density, $\sigma(x(t))$.

To manipulate the presenter, an *object* NeRF is trained on images with the background masked out. In addition to the typical mean squared error (MSE) loss that encourages renderings, $\hat{C}$, to match target images, C, two new losses were included. One penalizes high background σ values while the other ensures a fully opaque target object. The loss function for this model is defined as:

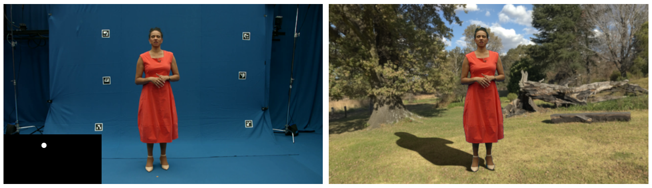

Fig. 6: (Right) projected and composited cast shadows given (left) a frame of ForecasterFlexOBM and lighting

$$\mathcal{L} = ||C - \hat{C}||_2^2 + \frac{1}{N_B}\sum_{\forall b \epsilon B}\sigma_b + \left|1 - \frac{1}{N_{R_o}}\sum_{\forall r \epsilon R_o}\sum_{i=1}^{N} W_{r,i}\right|$$

where B is the set of all background points, and N_B is the number of points in said set. R_o is the set of all object rays, and N_{R_o} is the number of elements in this set.

A *background* NeRF can then be trained on images of the background environment, with just an MSE reconstruction loss. Once trained, each model can be queried with arbitrary sets of sampling rays - the resulting field values are then composed during a joint rendering operation.

Figure 5 illustrates the editing capabilities granted by this compositing pipeline. The full scene is first reconstructed from two separate NeRFs before the presenter is placed within a new environment, rendered from an unseen pose, and subjected to a small 3D translation vector. The combination of modular, photo-realistic objects, novel view synthesis, and translation provides a promising demonstration of flexible media through learned, implicit scene representations.

4.5. Projective Cast Shadows for Actors/Presenters

Shadows can contribute significantly to the realism of a rendering [36] and have been studied in the literature since as early as the seminal work of Williams [37]. In particular, rendering cast shadows of presenters and actors, which are missing from the green screen cut-outs, plays an important role for compositing plausible mixed-reality scenes with real or synthetic background objects in OBM production.

ForecasterFlexOBM data is used to generate cast shadows for foreground weather presenter on existing background objects of corresponding regions. This kind of personalization can enhance the perception of the delivered forecast for younger audiences. The pseudo-labels can be generated by, e.g., multiple-view geometry reconstruction algorithms of [38], given the calibrated views of ForecasterFlexOBM.

Traditionally, cast shadows can be rendered given the geometry of the foreground actor. However, as reconstruction of such geometry requires multiple-camera capturing setups [38], Einabadi et al. [2] propose a neural rendering [39] approach to learn the transformation from the actors' silhouettes captured by the (monocular) director camera to the corresponding cast shadows on arbitrary, given shadow receiving geometry – this is especially beneficial for cost-effective green screen scenarios. More specifically, cast shadows are learned in a proposed intermediate representation *canonical space*, which can then be easily projected onto the rest of the (background) scene. For more details please refer to [2].

Figure 6 illustrates sample rendered cast shadows using [2] (trained on their proposed dataset 3DVHshadow [2]), evaluated on the sequences of ForecasterFlexOBM in a mixed-reality, weather forecaster scenario. The inputs to the algorithm are the silhouette of the weather presenter from the viewpoint of the director production camera, and a target, desired light position. The output is the canonical shadow which is then cast onto a given background geometry.

5. CONCLUSION

We introduced a new dataset themed on a UK weather forecast using professional actors/presenters and demonstrated how it is used for evaluation purposes in different areas of object-based media production such as action/event detection and modular NeRFs. In doing so, we aimed to bridge the gap between the flexibility of research datasets and the quality of professionally produced media. The dataset is multi-view, calibrated and contains multiple audio channels, as well as a rich set of production- and research-oriented sequences.

Acknowledgments
The authors would like to thank Director Gavin Saville, Asst. Director Fiona Rivera; Actors Grace Cookey-Gam, Paul Courtenay Hugh and Lisa Kelly; BSL/English Interpreter Sammi Villabon; BBC's Mark Lomas, Peter Mills, Jack Reynolds, James Shephard and Emma Young; the Institute of Sound Recording. Research supported by EPSRC BBC Prosperity Partnership AI4ME: Future personalized Object-Based Media Experiences Delivered at Scale Anywhere EP/V038087/1. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Availability of data
Data generated as part of this research is freely available under the terms and conditions detailed in the licence agreement enclosed in the data repository. Access details are available from the University of Surrey: doi.org/10.15126/surreydata.900912.

6. REFERENCES

[1] Stastia, "Global: video-on-demand users 2018-2027 by segment," statista.com/forecasts/456771/video-on-demand-users-in-the-world-forecast.

[2] F. Einabadi, J.-Y. Guillemaut, and A. Hilton, "Learning projective shadow textures for neural rendering of human cast shadows from silhouettes," in *EGSR*, 2023, pp. 63–75.

[3] G. A. Sigurdsson et al., "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding," in *ECCV*, 2016, pp. 510–526.

[4] S. Yeung et al., "Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos," *IJCV*, vol. 126, no. 2, pp. 375–389, 2018.

[5] S. Jolly, G. Phillipson, and M. Evans, "Old school: An 8k multicamera shoot to create a dataset for computational cinematography," in *IMXw*, 2023, p. 76.

[6] D. Berghi, M. Volino, and P. J. B. Jackson, "Tragic Talkers: A Shakespearean sound- and light-field dataset for audio-visual machine learning research," in *CVMP*, 2022.

[7] J. Woodcock et al., "Presenting the S3A Object-Based Audio Drama Dataset," 2016, http://www.aes.org/e-lib/browse.cfm?elib=18159.

[8] ITU, "ITU-R BS.2076-2: Audio definition model," Tech. Rep., International Telecommunications Union, Oct. 2019.

[9] J. Cox, M. Brooks, I. Forrester, and M. Armstrong, "Moving Object-Based Media Production from One-Off Examples to Scalable Workflows," *SMPTE Motion Imag. J.*, vol. 127, no. 4, pp. 32–37, 2018.

[10] L. Ward et al., "Casualty Accessible and Enhanced (A&E) Audio: Trialling Object-Based Accessible TV Audio," in *AES*, 2019.

[11] R. Altman, "The editing and tech behind Netflix's Black Mirror: Bandersnatch," May 2019, postperspective.com/netflixs-black-mirror-bandersnatch-lets-viewers-choose.

[12] Ofcom, "Object-based media report," Tech. Rep., Office of Communications, Sept. 2021.

[13] C. Cieciura, M. Glancy, and P. J. B. Jackson, "Producing Personalised Object-Based Audio-Visual Experiences: an Ethnographic Study," in *IMX*, 2023, pp. 71–82.

[14] M. Armstrong et al., "Taking Object-Based Media from the Research Environment Into Mainstream Production," *SMPTE Motion Imag. J.*, vol. 129, no. 5, pp. 30–38, 2020.

[15] K. Hentschel and J. Francombe, "Exploring audio device orchestration in workshops with audio professionals," in *AES*, 2020.

[16] L. Ward, M. Glancy, S. Bowman, and M. Armstrong, "The Impact of New Forms of Media on Production Tools and Practices," in *IBC*, 2020.

[17] M. Glancy et al., "Object-Based Media: An Overview Of The User Experience," in *BBC White Paper*, 2020, vol. 390.

[18] M. Leonard, "Forecaster: our experimental object-based weather forecast," Dec. 2015, bbc.co.uk/rd/blog/2015-11-forecaster-our-experimental-object-based-weather-forecast.

[19] F. Schweiger et al., "Tools for 6-DoF immersive audio-visual content capture and production," in *IBC*, 2022.

[20] J. Roth et al., "AVA active speaker: An audio-visual dataset for active speaker detection," in *ICASSP*, 2020, pp. 4492–4496.

[21] R. Tao et al., "Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection," in *MM*, 2021, p. 3927–3935.

[22] D. Berghi, A. Hilton, and P. J. B. Jackson, "Visually supervised speaker detection and localization via microphone array," in *MMSP*, 2021, pp. 1–6.

[23] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[24] X. Qian et al., "Multi-speaker tracking from an audio–visual sensing device," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2576–2588, 2019.

[25] M. A. Mohd Izhar, M. Volino, A. Hilton, and P. J. B. Jackson, "Tracking sound sources for object-based spatial audio in 3D audio-visual production," in *Forum Acusticum*, 2020, pp. 2051–2058.

[26] C. Wright et al., "AI in production: Video analysis and machine learning for expanded live events coverage," *SMPTE Motion Imag. J.*, vol. 129, no. 2, pp. 36–45, 2020.

[27] F. Sardari, A. Mustafa, P. J. B. Jackson, and A. Hilton, "PAT: Position-aware transformer for dense multi-label action detection," in *ICCV Workshops*, 2023, pp. 2988–2997.

[28] X. Kong et al., "En-Compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning," in *CVPR*, 2022, pp. 9306–9315.

[29] F. Zhou et al., "Revisiting prototypical network for cross domain few-shot learning," in *CVPR*, 2023, pp. 20061–20070.

[30] R. Dai et al., "MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection," in *CVPR*, 2022, pp. 20041–20051.

[31] A. Nadeem et al., "SEM-POS: Grammatically and semantically correct video captioning," in *CVPR Workshops*, 2023, pp. 2605–2615.

[32] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *ACL*, 2011, pp. 190–200.

[33] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *CVPR*, 2016, pp. 5288–5296.

[34] A. Nadeem et al., "CAD-Contextual multi-modal alignment for dynamic AVQA," in *WACV*, 2024, pp. 7251–7263.

[35] B. Mildenhall et al., "NeRF: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[36] F. Einabadi, J.-Y. Guillemaut, and A. Hilton, "Deep neural models for illumination estimation and relighting: A survey," *Comput. Graph. Forum*, vol. 40, no. 6, pp. 315–331, 2021.

[37] L. Williams, "Casting curved shadows on curved surfaces," in *SIGGRAPH*, 1978, pp. 270–274.

[38] J. Starck and A. Hilton, "Surface capture for performance-based animation," *IEEE Comput. Graph. Appl.*, vol. 27, no. 3, pp. 21–31, 2007.

[39] A. Tewari et al., "Advances in neural rendering," *Comput. Graph. Forum*, vol. 41, no. 2, pp. 703–735, 2022.