AI in Production: Video Analysis and Machine Learning for Expanded Live Events Coverage

By Craig Wright, Jack Allnutt, Rosie Campbell, Michael Evans, Ronan Forman, James Gibson, Stephen Jolly, Lianne Kerlin, Susan Lechelt, Graeme Phillipson, and Matthew Shotton

Abstract

As with many industries, TV and video production is likely to be transformed by artificial intelligence (AI) and machine learning (ML), with software and algorithms assisting production tasks that, conventionally, could only be carried out by people. Expanded coverage of a diverse range of live events is particularly constrained by the relative scarcity of skilled people, and it is a strong use case for AI-based automation. This article describes the recent research conducted by the British Broadcasting Corporation (BBC) on the potential production benefits of AI algorithms, using visual analysis and other techniques. Rigging small, static ultrahigh-definition (UHD) cameras, we have enabled a one-person crew to crop UHD footage in multiple ways and cut between the resulting shots, effectively creating multicamera HD coverage of events that cannot accommodate a camera crew. By working with programmakers to develop simple deterministic rules and, increasingly, training systems using advanced video analysis, we are developing a system of algorithms to automatically frame, sequence, and select shots, and construct acceptable multicamera coverage of previously untelevised types of events.

Artificial intelligence (AI) and machine learning (ML) have the potential to substantially increase the range and scale of events that broadcasters and other content producers can cover. It is not clear as to what the timescale and impact of these technologies will be or the extent to which they will assist existing human craft roles rather than automate parts of them.

Keywords

Broadcast technology, intelligent cinematography, TV broadcasting, user evaluation

Introduction

Artificial intelligence (AI) and machine learning (ML) have the potential to substantially increase the range and scale of events that broadcasters and other content producers can cover. It is not clear as to what the timescale and impact of these technologies will be or the extent to which they will assist existing human craft roles rather than automate parts of them. In this article, we present our first efforts to investigate these opportunities.

We will describe our recent work to simplify the process of covering staged events such as stand-up comedy or panel shows using new software tools and novel craft workflow. British Broadcasting Corporation (BBC) prototypes *Primer* and single operator vision mixer (*SOMA*)[1,2] use web technologies and our IP Studio implementation of the Advanced Media Workflow Association (AMWA) Networked Media Open Specifications (NMOS) standards[3] to allow a single operator to produce *nearly live* coverage of such performances. We then describe our experiences in developing *Ed*, a system that attempts to automate the work of a craftsperson using a rules-based AI approach. The challenges associated with evaluating the performance of such a system, as well as the prospects for improving it using ML, are discussed.

Our objective in developing automation for a specific production workflow is to learn where the limitations of AI lie. Our expectation is that our industry will benefit most from AI and ML in the short term by using these technologies to make people more effective—automating their most time-consuming or repetitive tasks—rather than by supplanting them.

Video Coverage of Hard-to-Reach Events

Providing video coverage of cultural and sports events, using conventional outside broadcast (OB) technologies, is challenging. Even if coverage is not required to be live (which mitigates the immediate need to get content from the event site to viewers' devices, probably via a broadcast center), OBs still need a significant amount of

Digital Object Identifier 10.5594/JMI.2020.2967204
Date of publication: 6 March 2020

1545-0279/20©2020SMPTE

equipment and people. From a video perspective, a typical OB requires several cameras with operators, and a gallery/video production area with a vision mixer, director, and other staff, as well as cabling from cameras to a gallery that conveys video and other signals.

The complexity and lack of scalability of this approach are limiting, which means that a large proportion of events that viewers might enjoy experiencing via video coverage are not covered. At the 2017 Edinburgh Fringe Festival—the largest cultural event in the world—there were more than 50,000 performances across 300 venues. Only a tiny fraction of these could be captured using conventional OB workflow. The BBC provides coverage from only around six of the nearly 100 places where music is performed at the Glastonbury festival.

Recently, the industry has begun to develop the workflow required for the kind of increase in video capture capacity that would support much more comprehensive coverage of this type of event. At the Edinburgh Fringe Festival in 2015 and 2016, BBC Research and Development experimented with using static UHD cameras in a variety of difficult-to-cover venues. UHD resolution means that each of these static wide shots can be cropped in multiple ways, in realtime, to create a much higher number of HD "virtual" camera shots. These were composed and sequenced by a single craftsperson, using a simple web application called *Primer*, allowing operators to create reasonable quality multicamera video footage, from performances that, previously, would have been too impractical to cover.[1]

Subsequently, this work helped enable a current BBC Research and Development project, *SOMA*, which is in use on an experimental basis.[2] We have also developed a highly compact, low-cost capture device suitable for these use cases, on the basis of IP Studio and the Raspberry Pi platform.

Outside the BBC, similar approaches are being used in a number of products and domains. Mevo[4] is a web-connected camera designed to be mounted statically, while an associated mobile phone application is used to create multiple crops of its imaging. Products like this could facilitate simple quasi-multicamera workflow for Vloggers or similar producers working on platforms such as YouTube and Facebook Live. Beyond web video, aimed at the potentially higher end requirements of broadcast, Datavideo's KMU-100 product is just one example of a camera processing unit for studios and OBs that allows the setting up of multiple crops of a 4K camera input, forming HD virtual cameras.[5] Enabling logistically straightforward location shoots is a key purpose of compact and heavily integrated *flypack* video production systems.[6]

The combined effect of these innovations is to increase the scope, in terms of infrastructure and crew size, for lightweight video production workflow at live events. There is evident potential for even more lightweight video capture, and broadcasting of many more events to audiences, by harnessing the power of AI-based automation.

ED—A Rule-based AI System for Automated Coverage

A proof-of-concept system, called *Ed*, has been built to capture and edit live events. Like SOMA, *Ed* takes one or more video streams as input, with each event captured using static UHD cameras, positioned for contrasting wide shots of the stage. SOMA requires a human operator to frame shots, and then switches between these to form output sequences. *Ed*, on the other hand, performs shot framing, sequencing, and selection autonomously. *Ed* has been developed to enable expanded coverage of a specific performance type—the live panel show common at Edinburgh and other festivals. However, the processes applied are largely invariant of the genre. *Ed* is a rule-based system, and its rules are based on recommendations made by real editorial staff during formative user experience (UX) research interviews. Implementation uses low-level feature extraction for framing and methods for sequencing and selecting shots. Examples of shot-framing guidelines include:

Position focal points of a shot in the center or on the third lines (rule-of-thirds)	Looking room should be given in the direction a person is facing

Examples of the shot sequencing and selection guidelines captured include:

Speakers are generally kept in shot	Switch between one-shots and two-shots for variety
Occasional cutaway to reaction shot	Occasional cutaway to establishing shot
Fast-paced shows should have fast-paced cuts	Shot durations should be similar but not linear

Feature Extraction

Ed software extracts several features from the video streams, using face detection and tracking, facial landmarking and pose estimation, and visual speaker detection. This indicates where people are in each frame, the directions they are facing, and when they are speaking. Our face- and speaker-detection methods are tuned to minimize false positives at the expense of more false negatives. Thus, faces or periods of speech are more likely to be *undetected* than *misdetected*. The left side of **Fig. 1** shows the detected face region, facial landmarks and pose from an example frame.

Framing

During our UX research, craftspeople described the need to center a shot around a focal point or place focal points

FIGURE 1. The face detection bounding box (green), facial landmarks (blue), and head pose projection (red) (left), and a camera view labelled with three candidate crops: two mid-close shots (green and blue) and a midshot (red) (right).

around invisible horizontal and vertical lines dividing the frame into thirds (the rule of thirds). In a panel show setting, the focal points are the panelists. When framing a shot on a single person, the facing direction of the person indicates whether he/she should be framed in the center of the shot or on one of the third lines.

The face detections and corresponding pose estimations are used to frame candidate *wide* (WS), *mid* (MS), and *close-up* (CU) crops, for each combination of faces: per individual, for each pair of people, each three, etc. Crops are framed to allow adequate head- and look-room and obey the rule of thirds. The right side of **Fig. 1** shows three candidate crops.

Shot Sequencing

Sequencing is the process of defining when shot changes will occur. The sequence cadence is a function of the minimum and maximum shot duration. No shots should be outside these. Given the requirement to generally keep the speaker in shot, the method of sequencing in *Ed* is to schedule shot changes to be near speech events (i.e., when people start or stop talking). The detected periods of speech are used to inform shot sequencing.

A heuristic method of estimating sequences of shot changes temporally close to the detected speech events is used: the algorithm generates a linearly spaced shot timeline, before each shot change is adjusted in the direction of the nearest speech event, as much as is permitted. Where the minimum and maximum shot lengths are l_{min} and l_{max}, respectively, the linear spacing is given by $(l_{max} + l_{min})/2$, and the maximum permitted adjustment is given by $(l_{max} - l_{min})/4$. This heuristic method is illustrated in **Fig. 2**.

Shot Selection

Shot selection is the process of assigning one of the framed crops to the period between each pair of shot boundaries in the sequence. In our UX interviews, craftspeople stated that they: 1) generally keep speakers in shot; 2) occasionally cutaway to a reaction shot; and 3) occasionally cutaway to an establishing shot. In the live panel show setting, the hosts and panelists do not generally move around once they have taken their seats. (As the cameras are all positioned in an arc around the front of the panel, it should be impossible to break continuity editing rules such as the 180° rule or continuity of movement.) The suitability of a framed crop for a given shot region is given by:

- the amount of speech originating from within the crop;
- the number of people in the crop;
- the crop type (close, mid, wide);
- how recently the crop was used.

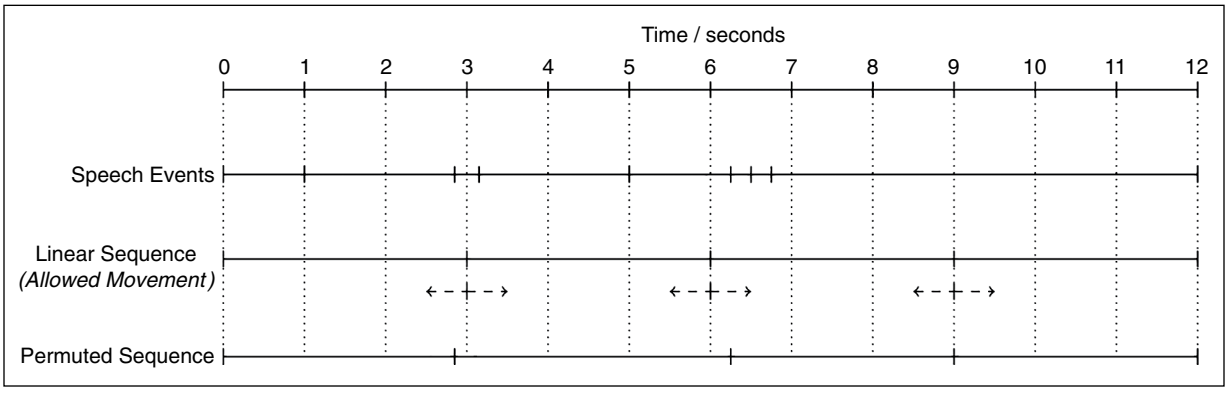

FIGURE 2. Speech events, linear sequence with allowed movements, and favorable permuted sequence using the heuristic approach over a 12-sec period with minimum and maximum shot lengths of 2 and 4 sec, respectively.

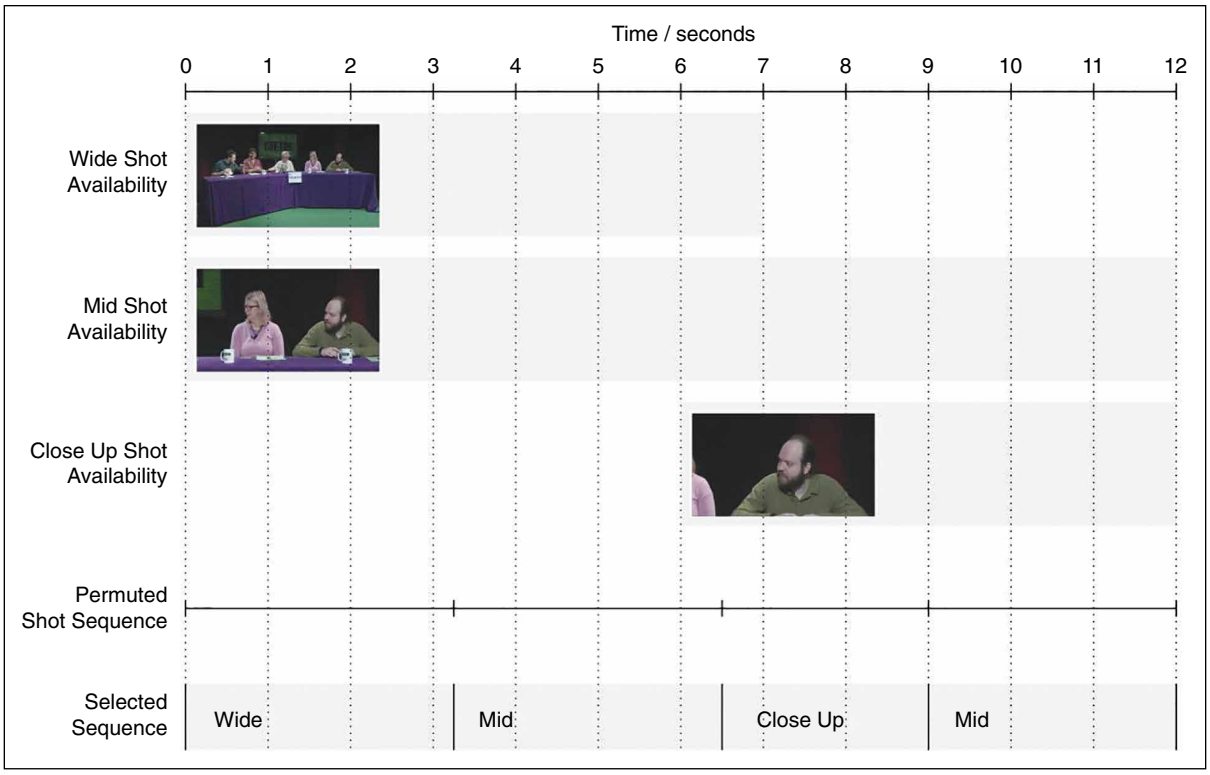

FIGURE 3. Availability of candidate crops and an example shot selection.

When speech is detected during a shot, a closer crop containing fewer people and more speech is favorable. Conversely, when no speech is detected, a more distant crop containing more people is preferable. A crop that was not recently used is always favored. Each shot in the generated shot sequence is selected in time order. All the framed crops that are available in the video content for the corresponding time period are considered, and the crop that scores most favorably is selected. This method is illustrated in **Fig. 3**.

Evaluation and Improvement

Motivation

The performance of *Ed*, and the perceived quality of the system's output, can be described by answering a pair of related research questions.

Shot framing: How do the shot framing, sequencing, and selection decisions made by *Ed* compare to those a human programmaker would have made with the same material and brief?

Viewing experience: Secondly, what is the quality of the viewing experience for the audience?

Answering these questions requires empirical work with people: specifically, with viewers and production professionals. Also, to inform, evaluate, and iterate engineering decisions, it is important to conduct this human-centered work in parallel with algorithmic development. As discussed earlier, the shot-framing decisions made by the *Ed* prototype are based on a relatively simple set of guidelines, distilled from research interviews with professionals. Therefore, a practical investigation of how effective and satisfactory these rules are for viewers has been an early priority for the project—to support progressive refinement. We have conducted a subjective study to compare human and algorithmic shot framing by having reference footage cropped both by experienced professionals and *Ed*, allowing us to investigate the impact of the differences on viewer experience.

Shot-Framing Study Methodology

We developed and conducted a shot-framing study consisting of two empirical phases. First, to investigate (a), we asked four experienced professional filmmakers (a combination of directors and camera operators) to each frame a large set of shots. *Ed* was also used to produce an equivalent set of shots. Second, we asked a number of viewers each to compare *Ed's* shots to those framed by the humans, to understand (b).

Stage 1—Professionals: Reference video material for the shot-framing study was captured in a dedicated studio shoot, consisting of a specially staged panel show (**Fig. 4**). The performance was comprised of five people, in two different seating configurations, captured in very wide, 4K shots from the center, left, and right. Cameras were static and positioned in such a way as to be able to support their output being cropped to cover every individual, pair, or larger group within the panel. Researchers used the shot footage to select 2 sec clips from multiple angles, collectively featuring a broad variety of face direction, interactions, and combinations of speaker across the five people in the shot.

FIGURE 4. Capturing reference footage in studio for the shot framing study.

Using this corpus of reference video, four professional programmakers were each asked to frame various one (person) shots, two shots, and three shots of the panel, using four specified shot types; CU, medium close-up (MCU), MS, and medium long shot (MLS). Exactly the same framing instructions were given to *Ed*, yielding comparable but distinct individual crops. In total, several hundred framed clips were obtained, making an extensive pairwise comparison—between human and human, and human and machine—possible. The professionals were asked to speak aloud while performing framing to understand their reasoning.

Stage 2—Viewers: Twenty four nonexpert viewers were individually presented with a uniquely ordered sequence of clip pairs, including a combination of human-to-human and human-to-algorithmic comparisons. For every pair, each viewer was asked whether the clip on the left or on the right was more appealing, or if they had no preference. Viewers were encouraged to think aloud during a number of their selections and undertook a semistructured interview afterward, providing qualitative data to enable us to understand the factors behind their preferences.

Outcomes and Impact

Viewer participants selected their preferred shot framings, spoke their considerations aloud, and had the factors affecting their clip preferences probed in the interview. Based on this qualitative data around preferences, it has been possible to derive a list of high-priority improvements to the framing guidelines used by *Ed*, expressed as engineering tasks for the next iteration of the system. We expect implementation of these findings to represent *quick wins* for improving the subjective performance of *Ed* with more appealing shot framing.

These five guidelines are illustrated in the example shot framings in **Figs. 5–9**. In each case, the human-framed shot on the right was preferred to the shot that was algorithmically framed by *Ed*, shown on the left: note that, across the study, the left–right arrangement of the shots was balanced between *Ed* and human-framed material, and viewers were never told whether or not any given clip had been framed by a professional programmaker.

Guideline #1—Edges Should Be Clear of Objects

Viewers expressed a clear preference for any objects in clips (e.g., a plant, sign, or mug) to be framed fully in or fully out of shot. Views of objects truncated by the edge of the frame were regarded as distracting and unprofessional. Participant V8 pointed out that it was "annoying to see a quarter of the sign" as shown in the left-hand clip in **Fig. 5**.

Guideline #2—Edges Should Be Clear of Partially Seen People

Very similar to Guideline #1, viewers disliked shots in which the edge of the frame cut through people's faces,

FIGURE 5. MS framed by *Ed* (left) and by a human professional (right, preferred).

FIGURE 6. MLS framed by *Ed* (left) and by a human professional (right, preferred).

figures, or limbs because it distracted their attention away from the focus of the shot (such as the conversation among panel members in **Fig. 6**). As described by Participant V4, with "somebody else on the side…" she feels that she "can't focus." Participants consistently demonstrated a preference for clips that contained panel members, and especially their faces, either fully in or fully out of frame.

Guideline #3—Avoid Excessive Zoom on One Shots

The preference for one shots was to avoid excessively zoomed-in views of the face. We found that participants preferred one shots to contain the full head and a little bit of body, as the right-hand view in **Fig. 7**. In describing the clips shown in **Fig. 7**, Participant V1 suggested that it was "better to see more of head," as on the right. On the whole, viewers suggested that too much face on screen was intrusive, as pointed out by Participant V12 who stated that "There's just something really weird about having [faces] really close up."

Guideline #4—Avoid Cutting off Tops of Heads

Similarly, viewers preferred one shots that kept the full face in view with a little background space surrounding the head, as on the right of **Fig. 8**. Participants described clips in which the top of the head had been cut off as being uncomfortable. Participant V7 asked "Why cut off his head? and much preferred to have ... the whole head in, better to get the whole person in," as suggested by Participant V9.

Guideline #5—Avoid/Minimize Empty Space

Participants disliked clips that contained too much empty space, as in the left-hand clip in **Fig. 9**. As Participant V23 pointed out, "there is a lot of dead space and areas of block color so it feels a bit empty. It feels like there is too much of nothing. It's more the black than the purple but feels like there should be more there." In practice, adding a rule to *Ed* to minimize such spaces means selecting a framing that minimizes the amount of block color, such as the purple of the table cloth or the black of the background.

These five suggestions for enhanced *Ed's* ruleset represent an initial stage of analysis of the framing study and have been selected based on their likely scope for quality improvement and technical feasibility.

Future Evaluative Work

We are preparing further use of a similar human-centered research approach in evaluating and improving the sequencing and selection of shots in our system. The general format will be broadly similar to the framing study. We will ask a cohort of professional programmakers to select shots and their transitions and timing, producing a cut sequence. Viewers will then describe, subjectively, how equivalent sequences produced by the current iteration *Ed* compare to these.

A key question in quality evaluation of this kind (recognizing that an automated system may never fully achieve the subjective quality of skilled human craft) will be: when is an algorithm *good enough* for an audience,

FIGURE 7. CU framed by *Ed* (left) and by a human professional (right, preferred).

FIGURE 8. CU framed by *Ed* (left) and by a human professional (right, preferred).

FIGURE 9. MLS framed by *Ed* (left) and by a human professional (right, preferred).

for a given content type? How will we know when to stop trying to enhance our algorithms? Previous work has shown that subjective viewer evaluation, based on overall quality of experience (QoE) approach, can characterize the relative impact of video, even when there is a wide variation in technical quality.[7]

Application of ML

A limitation of designed approaches—enumerating, as we have done, a finite set of deterministic rules—is that production is at least as much art as science. In addition, ML has demonstrated huge advances in recent years in relevant areas such as image classification, face detection, and pose estimation. Google has demonstrated a system that has learned how to frame and post-process images to produce photographs, a portion of which are comparable in quality to human performance.[8] Similarly, Twitter has been able to use deep learning to rapidly crop image thumbnails and show the most relevant part of an image.[9] Additionally, there are systems available that can automatically or semi-automatically capture certain sports.[10–12] Advances in graphics processing unit (GPU) capability and algorithmic effectiveness[13] make it much easier to process large amounts of data such as that required for training networks using the analysis of broadcast-quality video. Lower cost devices could perform the inference stage of algorithms on-site.

TV archives, full of human-produced programs, could be a rich source of training data for ML, by describing what constitutes (for example) *good* framing. However, when learning from archive data, we only have the single, finished version, even though there would have been many potentially good alternative options reflecting different personal and genre styles.[14] Additionally, it is hard to evaluate the quality of editing directly as, when the quality is high, as many as one-third of the edits will be missed.[15] Large datasets, such as TV archives, still represent significant computational analysis challenges. So far, we have only considered vision mixing of live events. It would be much harder for ML algorithms to carry out nonlinear editing tasks, like the selection of general views and cutaways when editing a news package, or analyzing multiple takes of a scene in a drama for subjective qualities such as comic timing, or chemistry between actors.

Conclusion

This article has described work that applies AI techniques to a specific production challenge, making it possible to provide engaging multicamera coverage from a significantly wider range of live events, performances, and venues. The relative scarcity of conventional OB capacity restricts producers to a narrow range of events. We have shown that automating shot-framing and sequencing decisions that would otherwise require an impractical number of skilled people could permit coverage of events on a potentially larger scale.

The *Ed* prototype is being developed using insights from empirical UX research and from emerging

technologies, most notably, ML. In evaluating the performance of the system, important questions will include understanding when quality is sufficiently good to satisfy viewers' expectations and how broadly deployable a system developed for a specific use case, such as a comedy panel show, will be.

If *Ed* can be developed sufficiently to provide coverage of a panel show that is comparable to a human director with moderate skills, how badly would the system perform when used for a similar but distinct use case, such as an on-stage music performance? More broadly, the broadcast industry's archive of human-produced material is a resource of potentially huge value for training AI technology, but can it be analyzed on a large scale? And what are the professional and creative implications if AI/ML can automate currently unforeseen tasks?

Trying to answer these questions and understand the challenges of bringing the potential benefits of AI to media production will continue to be a fascinating and important activity and a valuable catalyst in developing data-driven, algorithmic innovations in production processes well beyond the basic coverage of live events.

Acknowledgement

The authors would like to thank their professional production colleagues for agreeing to be interviewed as part of the development of the techniques and use cases described in the article, and for participating in the empirical studies.

References

1. R. Campbell, D. Hett, C. Hoare, M. Lomas, T. Pearce, M. Evans, S. Jolly, J. Merrick, and J. Cox, "Nearly Live Production," 2015. [Online]. Available: https://www.bbc.co.uk/rd/projects/nearly-live-production
2. D. Winter, "Building a Live Television Mixing Application for the Browser," 2017. [Online]. Available: https://www.bbc.co.uk/rd/blog/2017-05-video-mixing-application-browser
3. P. Brightwell, J. Rosser, R. Wadge, and P. Tudor, "The IP Studio," *Proc. of IBC 2013*. [Online]. Available: https://doi.org/10.1049/ibc.2013.0016
4. Available: https://getmevo.com/
5. Available: http://www.datavideo.com/product/KMU-100
6. J. Owens, *Television Sports Production*, 5th ed., Routledge, 2015.
7. M. Evans, L. Kerlin, O. Larner, and R. Campbell, "Feels Like Being There: Viewers Describe the Quality of Experience of Festival Video Using Their Own Words," *Proc. ACM CHI Extended Abstracts (CHI'18 EA)*, 2018. [Online]. Available: https://doi.org/10.1145/3170427.3188507
8. H. Fang and M. Zhang, "Creatism: A Deep-Learning Photographer Capable of Creating Professional Work," 2017. [Online]. Available: https://arxiv.org/abs/1707.03491
9. L. Theis, I. Korshunova, A. Tejani, and F. Huszár, "Faster Gaze Prediction with Dense Networks and Fisher Pruning," 2018. [Online]. Available: https://arxiv.org/abs/1801.05787
10. "Pixellot Automatic Production." [Online]. Available: http://www.pixellot.tv/
11. "Automatic TV: The Automatic sports production system." [Online]. Available: http://automatic.tv/
12. "Hawkeye Innovations." [Online]. Available: https://www.hawkeyeinnovations.com/
13. G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, 18:1527–1554, 2006.
14. C. Lino, R. Ronfard, Q. Galvane, and M. Gleicher, "How Do We Evaluate the Quality of Computational Editing Systems?" AAAI Workshop on Intelligent Cinematography and Editing, Québec, Canada, AAAI, pp. 35–39, 2014.
15. T. J. Smith and J. M. Henderson, "Edit Blindness: The Relationship Between Attention and Global Change Blindness in Dynamic Scenes," *J. Eye Movement Res.*, 2(2):1–7, 2008. [Online]. Available: http://dx.doi.org/10.16910/jemr.2.2.6

About the Authors

Craig Wright is a systems engineer turned research engineer with a background in systems programming, artificial intelligence, and computer vision. Since joining BBC Research and Development in 2016, he has contributed to projects including, the delivery of binaural audio and object-based audiovisual experiences in the browser; and the automatic detection of semantic programme to improve content search. Most recently, he has worked with a team of multidisciplinary engineers to prototype *Ed*: a system that automatically produces an edited video package from several input video streams captured using unmanned cameras.

Jack Allnutt is a software engineer with an interest in responsible technology, particularly on privacy and data protection. Since joining BBC Research and Development in 2014, he has contributed to projects including the development of object-based media experiences, the use of artificial intelligence (AI) in media production, the use of personal data stores in a public service environment and providing audiences access to innovative and experimental technologies.

Rosie Campbell is a program lead at the Partnership on Artificial Intelligence (PAI), San Francisco, CA, where she leads a research program on responsible publication practices for AI and machine learning (ML). Before joining PAI, she was assistant director of the Center for Human-Compatible AI, an AI safety research group at UC Berkeley, Berkeley, CA, working toward provably beneficial AI. Previously, she was a research engineer at the BBC Research and Development in the Future Experience Technologies Team, where she

worked on a variety of experimental technical projects spanning AI, computer vision and graphics, and web-based production interfaces. She holds a master's degree in computer science and a bachelor's degree in physics, and has additional academic experience in Philosophy and ML. She is passionate about emerging technology and cofounded Manchester Futurists, a thriving intellectual community group aiming to explore the social impact of technology and shape a positive future.

Michael Evans is a user experience research lead in Future Experience Technologies at BBC Research and Development, working with the other human–computer interaction (HCI) researchers and engineers to invent the public service media of the future. He has a lot of experience in developing tools and research methods for professional creative tasks, including directing 360° video, intelligible machine learning, and quality of experience evaluation. Before joining the BBC in 1999, he was a lecturer at the University of Reading, Reading, U.K., co-founding the Signal Processing Lab and leading HCI research. He is a chartered engineer and has completed a DPhil in spatial audio and psychoacoustics with BT Labs, Martlesham, U.K., in 1997.

Ronan Forman is a software engineer working at the BBC. After studying computer science and artificial intelligence at the University of York, York, U.K., he joined the BBC Software Engineering Graduate Scheme in 2017, where he worked across the BBC, including in Research and Development. While in Research and Development, he worked on prototyping the *Ed* system for automatically editing videos based on artificial intelligence (AI) learned rules, and running a framing study to evaluate the quality of automated and professional output.

James Gibson is a research and development engineer at the BBC Research and Development, U.K., who has worked on projects looking at how to transition broadcasters from serial digital interface to Internet Protocol (IP) the corresponding security requirements and new production opportunities, as well as new immersive applications that are enabled by 5G such as augmented reality/virtual reality (AR/VR) remote rendering.

Stephen Jolly leads the artificial intelligence (AI) in Media Production project at BBC Research and Development, which is exploring how AI and machine learning will transform the media industry. He is a strong advocate of the potential for Intelligent Cinematography to help automate and democratise the production of film and television, and is always interested to hear from potential partners with an interest in the field. In his career at the BBC, he has also worked on a wide range of other technologies—from 3D and high frame-rate television to multidevice media and the Internet of Things. He joined the BBC in 2004, following the successful completion of a PhD in high energy physics at Imperial College, London and CERN, Geneva. He also holds a BSc in physics from Imperial College, London, U.K.

Lianne Kerlin is a research scientist at BBC Research and Development who is interested in the impact of technology on people and society. She leads research and development work around human values; a project that translates psychological insight into actionable tools to shape future media services with people's values at the heart of the innovation process. She is currently focused on turning human values into psychometrics, and by doing so provide new ways to measure the impact and value of digital services beyond the standard consumption metrics.

Susan Lechelt is a post-doctoral researcher in creative informatics at The University of Edinburgh, Edinburgh, U.K. Her research involves engaging with creative practitioners in art, film, design, and beyond, to help them envision how new forms of technology and data-driven innovation might fit into their future practice. Before joining the University of Edinburgh, she completed a PhD in human–computer interaction at University College London (UCL), in collaboration with BBC Research and Development.

Graeme Phillipson currently works in BBC Research and Development on artificial intelligence (AI) in production project. This project is investigating how BBC can use techniques from AI, machine learning, and computer vision to tackle some of the problems found in television production. Previously in the

BBC, Phillipson has worked on several projects involved in the running of BBC iPlayer, a system which automatically keeps track of what music has been played on TV and radio so that audiences can find music they have heard. Prior to working at the BBC, he studied neuroscience at Edinburgh University, and worked on a 3D motion capture system used in clinical gait analysis.

Matthew Shotton is the founder and chief executive officer of Algoloop, a human in the loop machine learning (ML) consultancy based in Berkeley, CA. He previously worked as a research engineer at BBC Research and Development, where he worked on ML systems for automatically editing media content, interactive video rendering engines, and a range of other projects. He is passionate about the intersection of art and technology and is an avid tinkerer and hobbyist with a degree in electronic engineering from Manchester University, Manchester, U.K.

Presented at IBC2018, Amsterdam, The Netherlands, 13–17 September 2018. This article is published here by kind permission of the IBC and BBC.