AI IN PRODUCTION: VIDEO ANALYSIS AND MACHINE LEARNING FOR EXPANDED LIVE EVENTS COVERAGE

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

BBC Research and Development, UK

ABSTRACT

In common 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 is a strong use case for AI-based automation.

This paper describes recent BBC research into potential production benefits of AI algorithms, using visual analysis and other techniques. Rigging small, static UHD cameras, we have enabled a one-person crew to crop UHD footage in multiple ways and cut between the resulting shots, effectively creating multi-camera HD coverage of events that cannot accommodate a camera crew. By working with programme makers 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 event.

INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have the potential to increase substantially the range and scale of events that broadcasters and other content producers can cover. It is not clear 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 paper, we present our first efforts to investigate these opportunities.

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 is described: the BBC prototypes *Primer* and *SOMA* [1, 2] use web technologies and our *IP Studio* implementation of the AMWA 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 this craftsperson using a rules-based AI approach. The challenges associated with evaluating the performance of such a system are discussed, as well as the prospects for improving it using ML.

Our objective in developing automation for a specific production workflow is to learn where the limitations of AI lie, in the expectation 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

Capacity for providing video coverage of cultural and sport events, using conventional outside broadcast (OB) technologies, is fundamentally constrained: Even if coverage is not required to be live (which mitigates the immediate need to get content from the event site to the viewers' devices, probably via a broadcast centre) OBs still need a significant amount of 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. Cabling from cameras to gallery conveys video and other signals. The complexity and lack of scalability of this approach is limiting, and means that a large proportion of events that viewers might enjoy experiencing via video coverage, are not covered. At the Edinburgh Fringe Festival—the largest cultural event in the world—there were over 50000 performances across 300 venues in 2017. 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 that 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 in 2015 and 2016, BBC R&D 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 real time, 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 impractical [1]. Subsequently, this work helped enable a current BBC R&D project, *SOMA* (single operator vision mixer), 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, based on IP Studio and the Raspberry Pi platform.

Outside the BBC, similar approaches are seen in a number of products and companies addressing particular domains: Mevo [4] is a web-connected camera intended to be mounted statically whilst an associated mobile phone application is used to create multiple crops of its imaging. Products like this could facilitate simple quasi-multi-camera workflow for Vloggers or similar producers working on platforms like YouTube and Facebook Live. Beyond web video, and 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, as exemplified by the IPhrame Flyaway product from the company SuitcaseTV [6].

The combined effect of these innovations is to increase scope for lightweight video production workflow at live events, in terms of infrastructure and crew requirements. There is evident potential for even more lightweight video capture, and potentially to bring many more events to broadcast 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 for capturing and editing live events. Like SOMA, *Ed* takes one or more video streams as input, each captured using static UHD cameras, positioned for contrasting wide shots of the stage. Whereas SOMA requires a human operator to frame shots, and then switch between these to form output sequences, the *Ed* prototype 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 genre. *Ed* is a rules-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 centre 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

The *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 in which they are facing, and when they are speaking. Our face detection and speaker detection methods are tuned to minimise false-positives at the expense of more false-negatives. Therefore, faces or periods of speech are more likely to be *undetected* than *mis-detected*. The left half of Figure 1 the detected face region, facial landmarks and pose from an example frame.

Framing

During our UX research, craftspeople described the need to centre a shot around a focal point or place focal points 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 panellists. When framing a shot on a single person the looking direction of the person indicates whether they should be framed in the centre of the shot or on one of the third lines.

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

The face detections and corresponding pose estimations are used to frame candidate w *ide (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 half of Figure 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 the nearest speech event, as much as is permitted. Where the minimum and maximum shot length 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 Figure 2.

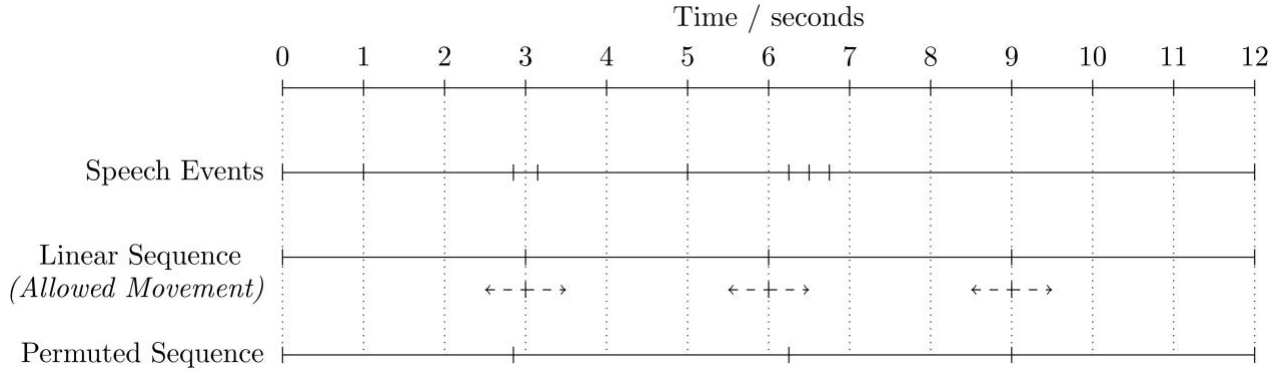

Figure 2 – Speech events, linear sequence with allowed movements, and favourable permuted sequence using the heuristic approach over a 12 second period with minimum and maximum shot length of 2 and 4 seconds respectively

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 advised 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 panellists generally do not 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-degree 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.*

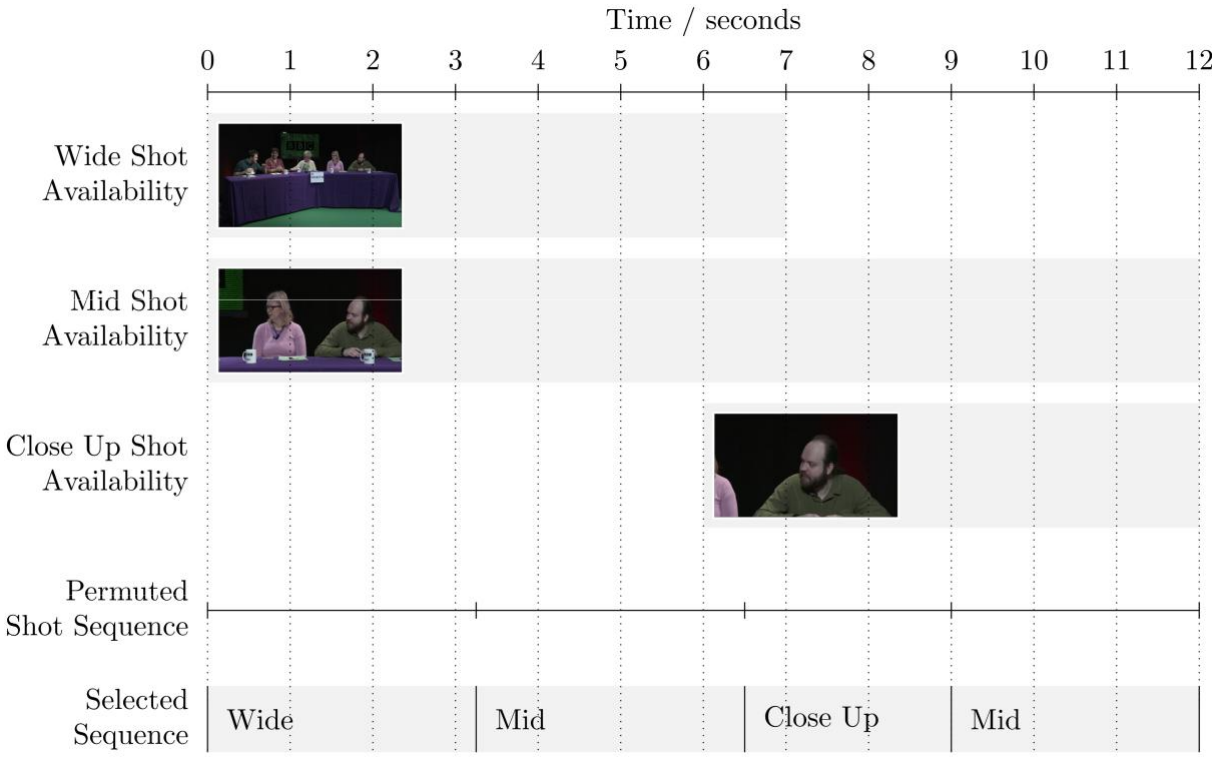

Figure 3 – Availability of candidate crops and an example shot selection

When speech is detected during a shot, a closer crop containing fewer people and more speech is favoured. Conversely, when no speech is detected, a more distant crop containing more people is favoured. A crop that was not recently used is always favoured. 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 favourably selected. The method is illustrated in Figure 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:

(a) *How do the shot framing, sequencing and selection decisions made by* Ed *compare with those that a human programme maker would have made with the same material and brief?*

(b) *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, in order to inform, evaluate and iterate engineering decisions, it is important to conduct this human-centred work in parallel with algorithmic development. As discussed earlier in this paper, 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 - in order 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 by *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: Firstly, 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. Secondly, we asked a number of viewers each to compare *Ed's* shots to those framed by the humans, to understand (b).

Figure 4 – Capturing reference footage in studio for the shot framing study

Stage 1 - Professionals: Reference video material for the shot framing study was captured in a dedicated studio shoot, consisting a specially-staged panel show. The performance comprised five people, in two different seating configurations, captured in very wide, 4K shots from the centre, left and right. Cameras were static and positioned in such a way to be able to support their output being cropped to cover every individual, pair, or larger group within the panel. Researchers used the shoot footage to select two-second clips

from multiple angles, collectively featuring a broad variety of face direction, interactions and combinations of speaker across the five people in shot. Using this corpus of reference video, four professional programme makers were each asked to frame various one (person) shots, two-shots and three-shots of the panel, using four specified shot types; *CU*, *MCU* (medium close-up), *MS* and *MLS* (medium long shot). Exactly the same framing instructions were given to the *Ed* software, yielding comparable but distinct individual crops. In total, several hundred framed clips were obtained, making extensive pairwise comparison—between human and human, and human and machine—possible. The professionals were asked to speak aloud whilst performing framing in order to understand their reasoning.

Stage 2 - Viewers: 24 viewers were each 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 semi-structured interview afterwards; providing qualitative data to enable us to understand 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* by more appealing shot framing. These five guidelines are illustrated in the example shot framings below. 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 programme maker.)

Guideline #1 - Edges should be clear of objects

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

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 Figure 5.

Guideline #2 - Edges should be clear of partially-seen people

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

Very similarly to Guideline #1, viewers disliked shots in which the edge of the frame cut through people's faces, figures or limbs, because it distracted their attention away from the focus of the shot (such as the conversation among panel members in Figure 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

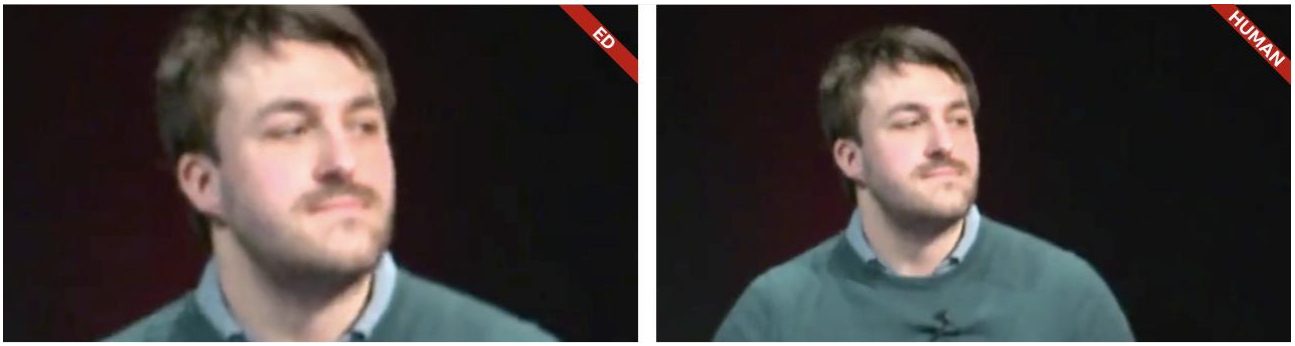

Figure 7 – CU framed by *Ed* (left) and by a human professional (right, preferred)

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 Figure 7. In describing the clips above, Participant V1 suggested 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 *'There's just something really weird about having [faces] really close up'*

Guideline #4 - Avoid cutting off tops of heads

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

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, Figure 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/minimise empty space

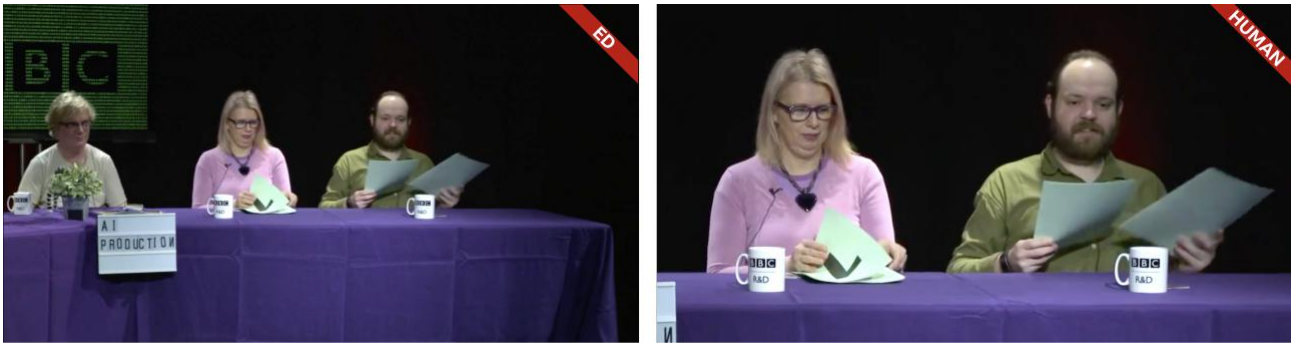

Figure 9 – MLS framed by *Ed* (left) and by a human professional (right, preferred)

Participants disliked clips that contained too much empty space, as in the left-hand clip in Figure 9. As Participant V23 pointed out *'there is a lot of dead space and areas of block colour 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 minimise such space means selecting a framing that minimises the amount of block colour, 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-centred 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 programme makers 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 of the *Ed* prototype compare to these.

A key question in quality evaluation of this kind (recognising 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, 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 an overall quality of experience (QoE) approach, can characterise the relative impact of video, even when there is a wide variation in technical quality [7].

APPLICATION OF MACHINE LEARNING

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, machine learning 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 learnt 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, 11, 12]. Advances in GPU capability and algorithmic effectiveness [13] make it much easier to process large amounts of data such as that required for broadcast-quality video analysis.

TV archives, full of human-produced programmes, could be a rich source of training data for machine learning, 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 non-linear editing tasks, like the selection of general views and cutaways when editing a news package, or analysing multiple takes of a scene in a drama for subjective qualities such as comic timing, or chemistry between actors.

CONCLUSION

This paper 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 constrains producers to a narrow range of events. We have shown that automating shot framing and sequencing decisions that would otherwise require impractical numbers of skilled people, could cover events at potentially huge scale.

The *Ed* prototype is being progressively developed using insights from empirical UX research and from emerging technologies, most notably, machine learning. 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 as comedy panel shows will be. Even 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 analysed at large scale? And what are the professional and creative implications if AI/ ML can automate tasks not currently foreseen?

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 basic coverage of live events.

ACKNOWLEDGEMENT

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

REFERENCES

1. Campbell, R. et al., 2015. Nearly Live Production. https://www.bbc.co.uk/rd/projects/nearly-live-production
2. Winter, D., 2017. Building a Live Television Mixing Application for the Browser. https://www.bbc.co.uk/rd/blog/2017-05-video-mixing-application-browser
3. Brightwell, P. et al., 2012. IP Studio. https://www.bbc.co.uk/rd/projects/ip-studio
4. https://getmevo.com/
5. http://www.datavideo.com/product/KMU-100
6. https://www.suitcasetv.com/live-event-mixing/iphrame-flyaway/
7. Evans, M., Kerlin, L., Larner, O., Campbell, R., 2018. Feels Like Being There: Viewers Describe the Quality of Experience of Festival Video Using Their Own Words. Proceedings of ACM CHI Extended Abstracts (CHI '18 EA), https://doi.org/10.1145/3170427.3188507
8. Fang H., Zhang M., 2017. Creatism: A deep-learning photographer capable of creating professional work, https://arxiv.org/abs/1707.03491
9. Theis, L., Korshunova, I., Tejani, A., Huszár, F., 2018. Faster gaze prediction with dense networks and Fisher pruning
10. http://www.pixellot.tv/
11. http://automatic.tv/
12. https://www.hawkeyeinnovations.com/
13. Hinton, G.E., Osindero, S., Teh, Y-W. 2006. A fast learning algorithm for deep belief nets. Neural Computation 18, pp 1527-1554
14. Lino, C., Ronfard, R., Galvane, Q., Gleicher, M., 2014. How Do We Evaluate the Quality of Computational Editing Systems? AAAI Workshop on Intelligent Cinematography and Editing, Québec, Canada. AAAI, pp.35-39

15. Smith, T.J, Henderson, J.M., 2008. Edit Blindness: The relationship between attention and global change blindness in dynamic scenes. Journal of Eye Movement Research vol. 2, no. 2, http://dx.doi.org/10.16910/jemr.2.2.6