# A Joint Attention Model for Automated Editing

Hui-Yin Wu
huiyin_wu@ncsu.edu

Arnav Jhala
ahjhala@ncsu.edu

North Carolina State University

**Abstract**

We introduce a model of joint attention for the task of automatically editing video recordings of corporate meetings. In a multi-camera setting, we extract pose data of participants from each frame and audio amplitude on individual headsets. These are used as features to train a system to predict the importance of each camera. Editing decisions and rhythm are learned from a corpus of human-expert edited videos. A Long-Short Term Memory (LSTM) neural network is then used to predict the joint attention by training on video and audio features from expert-edited videos, and editing predictions are made on other test data. The output of the system is an editing plan for the meeting in Edit Description Language (EDL) format.

## 1 Introduction

Joint attention is an element of human communication to draw the attention of others through non-verbal processes such as gaze and gesture [Hea05]. Editors uses continuity and analytical editing to imitate the audience's "natural attention" by using gaze to draw the audience's attention to elements of setting and story [Bor85, Smi12].

The explosive growth of everyday video recordings calls for smart methods that can understand context, and automatically process and present data in a meaningful way. While intelligent camera switching technology is available to a some extent, it is based primarily on audio, movement, and other low level features of video streams. Existing work shows that LSTMs have been effective for video summarization tasks [ZCSG16] due to their ability to model long ranged variable dependencies across shots. However, using LSTMs for a complex task such as video editing is currently insufficiently addressed.

We present a joint attention-based editing model where we extract audio and pose data to train an LSTM to rank each focal point in the room at each second, and produce an automated edit of videos. In order to be unbiased, the videos we choose are meeting recordings from the AMI corpus established by the University of Edinburgh [MCK$^+$05], where 100+ hours of meetings were recorded in smart meeting rooms equipped with multiple cameras, individual headsets, and other annotations. We address four main challenges:

1. detecting head pose and audio amplitude from each camera
2. training an LSTM model to rank joint attention of focal points in the room based on expert edits
3. generating an edit of the meeting using the score output by the LSTM

In: H. Wu, M. Si, A. Jhala (eds.): Proceedings of the Joint Workshop on Intelligent Narrative Technologies and Workshop on Intelligent Cinematography and Editing, Edmonton, Canada, 11-2018, published at http://ceur-ws.org

The extraction of the head pose and audio data, and the training of the model are pre-processed, while the final editing can be done in real time. We envision that the automated editing techniques built in this work could be broadly applicable to intelligent camera systems for both productivity and entertainment.

The rest of the paper first covers the related work in the field. After providing an overview, we introduce the calculation of head pose, and the technical details of the LSTM. Finally, we discuss limitations and future work.

## 2 Related Work

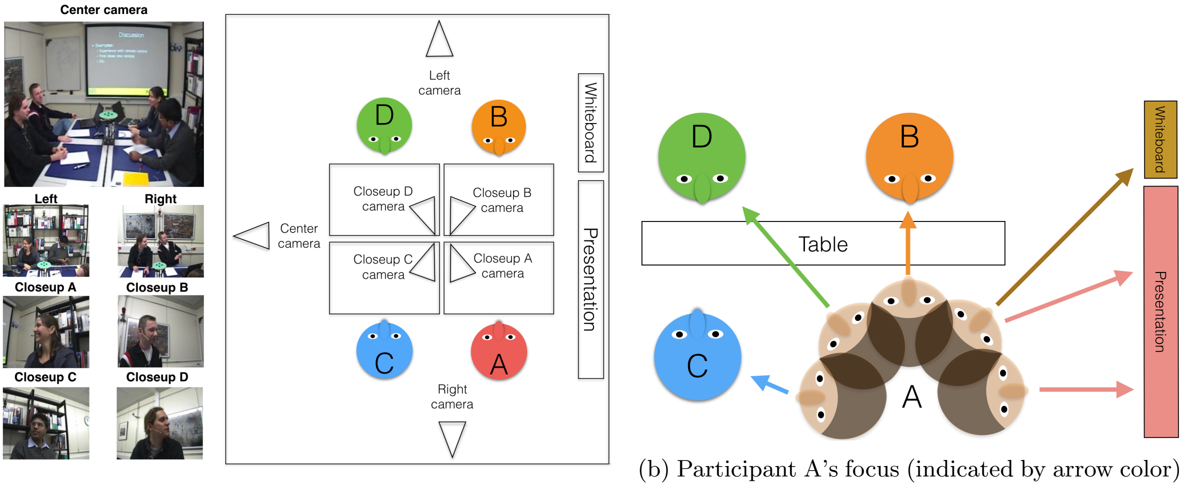
In HCI, sound and motion have been used as metrics for automated camera capture systems for editing meeting or lecture videos [BLA12, RBB08, LRGC01] to create edits. Arev et al. [APS+14] was the first to propose using joint attention in social cameras (i.e. cameras carried around by people during a group activity). Editing tools that take into consideration cinematographic rules are also emerging. Ozeki et al. [ONO04] created a system that generates attention-based edits of cooking shows by detecting speech cues and gestures. Notably Leake et al. [LDTA17] focuses on dialogue scenes and provides a quick approach that evaluates film idioms, and selects a camera for each line of dialogue.

Video abstraction or summarization techniques generally focus on the problem of selecting keyframes or identifying scene sections for a compact representation of a video. Ma et al. [MLZL02] were the first to propose a joint attention model for video summarization comprising the audio, movement, and textual information. Lee et al.[LG15] generated storyboards of daily tasks recorded on a wearable camera based on gaze, hand proximity, and frequency of appearance of an object. LSTMs have been used for video summarization tasks [ZCSG16] due to their ability to model long ranged variables dependencies, outside of a single frame.

Virtual cinematography has been a prominent area of study in graphics, where the challenge is often how to place cameras for story and navigation. Jhala and Young designed the Darshak system which generates a sequence of shots that fulfills goals of showing specific story actions and events [JY11]. Common editing rules [GRLC15] and film editing patterns [WC15] have been applied to virtual camera systems to ensure spatial and temporal continuity, and defining common film idioms.

Our work differs from virtual cinematography in two main ways: (1) in a 3D environment, object locations and camera parameters are precise, whereas in our video data, camera parameters and positions of people and objects can only be estimated, and (2) the virtual camera can be moved anywhere, while the cameras in the meeting recordings are fixed.

## 3 Joint Attention Model



(a) Configuration of the meeting room and cameras. The presentation and whiteboard is situated at the front of the room, visible from the center camera.

(b) Participant A's focus (indicated by arrow color) based on A's head pose and configuration of the meeting room. If two or more targets are close, there can be ambiguity as to what A is looking at.

Figure 1

We present the joint attention model from which we obtain features to train an LSTM. Meetings are selected from the videos in the AMI Corpus [MCK+05]: IDs IS1000a, IS1008d, and IS1009d, representing 3 types of interactions (brainstorming, status update, and project planning). Each meeting is around 25 minutes.

These meetings were held with 4 people in a smart meeting room equipped with 7 cameras and individual microphone headsets. The configuration of the room is shown in Figure 1a. The 7 cameras include 1 overview, 2 side cameras, and 4 closeups on each participant.

### 3.1 Determining Focal Points

To extract the head pose of people in the video, we use the OpenPose library's[CSWS17] MPI 15-keypoint detection, which has five points on the head: 2 ears, 2 eyes, and nose tip. Pose is calculated per frame, 25 fps. For each point, the on-screen ($x$,$y$) coordinates are given, with a confidence score between 0 and 1.

This information is assumed as input for the camera, which then calculates the confidence level of a head orientation based on visibility of five facial points: the two eyes and ears, and the nose. The eight possible head orientations are (F)ront (i.e. looking in the same direction as the camera), (B)ehind (i.e. looking at the camera), (L)eft and (R)ight from the camera's perspective (with only the nose tip and one eye and ear visible), and variations of these: LF (left-front), RF (right-front), LB (left-behind), and RB (right-behind). The confidence level for a head orientation is the sum of the confidence score $c$ of each of the $n$ points $pv$ that should be visible, and the sum of 1-$c$ for the $m$ points $ph$ that should be hidden.

$$\mathit{PoseConfidence} = \sum_{i=1}^{m} c_{pv_i} + \sum_{j=1}^{n} (1 - c_{ph_j}) \tag{1}$$

The pose is then mapped to focal points corresponding to the head orientation, such as the example shown in Figure 1b for participant A. Each pose can refer to one, none, or more than one focal point. The focal point matrix (Table 1) shows the mapping of head orientation-camera to the corresponding focal point(s).

Table 1: The focal point matrix indicates for the L, R and Closeup (CU) cameras, what each camera shows, and what focal point a participants' head orientation would correspond to. The focal points in the room are the presentation (P), the whiteboard (W), and participants A, B, C, and D.

| **Camera** | **Shows** | **L** | **R** | **LF** | **RF** | **F** | **LB** | **RB** | **B** |
|---|---|---|---|---|---|---|---|---|---|
| L | AC | PA | C | - | - | - | PW | D | BD |
| R | BD | D | PWB | - | - | - | C | A | AC |
| CU.A | A | P | C | - | - | - | PW | D | B |
| CU.B | B | D | PW | - | - | - | C | A | C |
| CU.C | C | PA | - | - | - | - | PW | D | B |
| CU.D | D | C | PWB | - | - | - | C | PWB | A |

The confidence level that the focal point is a specific target (a person/object) $t$ for a camera $C$ is the average of the confidence levels for the $x$ head orientations $o$ that have the target as a focal point. This value is summed up for each person $p$ detected by the camera.

$$\mathit{FocalPointConfidence}(C, t) = \sum_{p=1}^{n} \frac{\sum_{o=1}^{x} c_{o,p}(t)}{x} \tag{2}$$

This gives a focal point confidence level to each target in the room from the viewpoint of $C$, which ranks the importance of the target. With the importance of the targets in the room ranked by each camera, we then calculate the focal point of all participants. The focus $F$ for a target $t$ is the accumulated focal point confidence for each camera, since the more cameras that feel that $t$ is the focal point, the higher this score should be for $t$.

$$F(t) = \sum_{q=1}^{7} \mathit{FocalPointConfidence}(C_q, t) \tag{3}$$

This score and the extracted amplitude of individual headsets are used as input features to train our LSTM to identify the joint attention of the room, and generate the edit of the video, which we detail in the next section.

## 4 LSTM Model and Results

Our neural network is composed of an input layer, an output layer, and 3 hidden LSTM layers with 100 neurons each. It uses an MAE loss function and adam optimizer, trained for 1000 epochs. We choose LSTM layers as

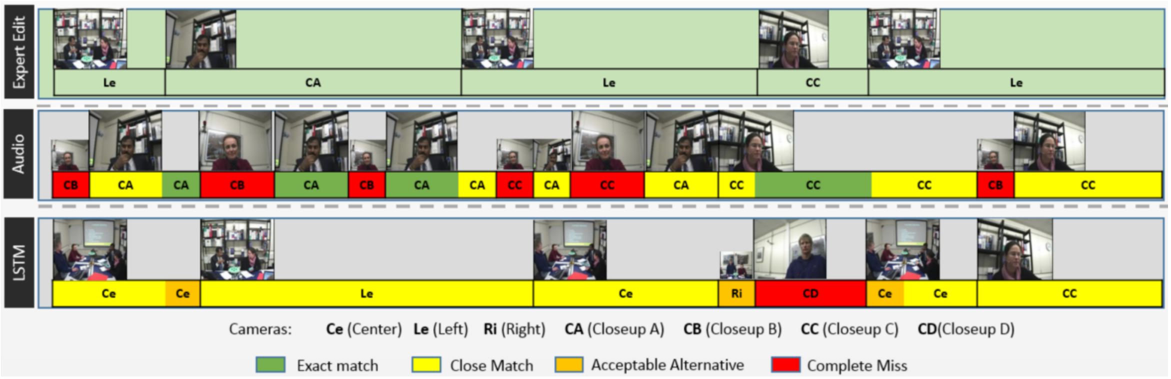


Figure 2: This figure shows a comparison of the audio-based, and LSTM-based edit to the expert version on one clip in terms of camera selection. Green indicates the same camera, yellow indicates partial targets or temporal displacement, orange an acceptable alternative, and red indicates a complete miss. In the clip (from meeting 1008d), A and C are debating the microphone feature of their product.

opposed to only fully connected layers due to the fact that editing is a decision-making process that takes into account both immediate (e.g. who just moved or started talking) and long-term (e.g. who has been talking for the past few seconds) observations, as well as previous editing decisions. The output is the joint attention score, between 0.0 and 1.0, of 6 targets in the room: each of the four participants, the whiteboard, and the presentation. Presumably, a higher score implies that the target is more likely the joint attention of the room.

The ground truth input to our model comes from a film expert who edited the three chosen meeting videos based on what they felt was the joint attention of the room, and we used Table 1 to convert this into attention scores. For example, if the expert chose the *Left* camera at time $t$ of the meeting, the ground truth score label for participants $A_t$ and $C_t$ would be 1.0, while the score 0.0 would be assigned to participants $B_t$, $D_t$, and the Whiteboard $W_t$, and Presentation $P_t$. An exponential decay function smooths out the scores before and after the shot.

We then design a simple editing program, which (1) chooses the Center camera when the $W_t$ or $P_t$ has a high score, and if not, (2) determines if a single participant $X$ scores over 0.5, in which case, the program chooses the closeup on $X$, or (3) if the score of two participants from the same side of the table have a score significant difference over the two on the other side, a medium shot that shows the two participants with the higher score is chosen, otherwise, (4) the Center camera is chosen.

To provide a baseline for comparison, an audio-based edit is generated by selecting the closeup camera that shows the person with the highest microphone input. Figure 2 shows one clip of the audio-based and LSTM-learned joint attention score based edit, compared with the expert edit.

We find that the audio edit often correctly shows who is talking, and in a simple scenario has high accuracy. However, the output is jittery and subject to noise in the audio data, resulting in more errors. It also chooses only closeup shots, which is not suitable for situations where multiple participants are in a discussion. In contrast, the LSTM-joint attention approach switches cameras in a more timely fashion, in many cases capturing both action and essential reactions of the meetings. It also identifies scenarios with interchanges or discussions between multiple participants more accurately than the audio-based approach, and uses the central camera to capture these exchanges. The LSTM learned joint attention produces a comparable edit to the human one, even with the limited amount of training data and no smoothing.

## 5 Conclusion

This work presents the idea and an initial formulation of Joint Attention with edited meeting videos to form a baseline corpus. With more advanced pose-detection algorithms, and high quality video data, we envision this work to have real-time editing applications using pre-trained LSTM models. Beyond basic criteria of pacing or shot similarity, we also hope to establish a baseline for future work on automated editing systems in various scenarios such as classrooms, film, and performing arts.

In conclusion, we have introduced a joint attention model for an automated editing tool based on LSTM. Our generated output was evaluated against a baseline audio edit, compared to the expert edit, and showed that it did reasonably well in terms of camera selection and pacing.

## References

[APS$^+$14] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. Automatic editing of footage from multiple social cameras. *ACM Trans. Graph.*, 33(4):81:1–81:11, July 2014.

[BLA12] Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. Tools for placing cuts and transitions in interview video. *ACM Trans. Graph.*, 31(4):67:1–67:8, July 2012.

[Bor85] David Bordwell. *Narrative in the Fiction Film.* University of Wisconsin Press, 1985.

[CSWS17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, Honolulu, Hawaii, USA, 2017. IEEE Xplore.

[GRLC15] Quentin Galvane, Rémi Ronfard, Christophe Lino, and Marc Christie. Continuity Editing for 3D Animation. In *AAAI Conference on Artificial Intelligence*, pages 753–761, Austin, Texas, United States, January 2015. AAAI Press.

[Hea05] Jane Heal. *Joint Attention: Communication and Other Minds: Issues in Philosophy and Psychology.* Oxford University Press, 2005.

[JY11] Arnav Jhala and R. Michael Young. *Intelligent Machinima Generation for Visual Storytelling*, pages 151–170. Springer New York, New York, NY, 2011.

[LDTA17] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. Computational video editing for dialogue-driven scenes. In *Proceedings of SIGGRAPH 2017*, 2017.

[LG15] Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, Aug 2015.

[LRGC01] Qiong Liu, Yong Rui, Anoop Gupta, and JJ Cadiz. Automating camera management for lecture room environments. In *Proceedings of SIGCHI'01*, volume 3, pages 442–449, March 2001.

[MCK$^+$05] Iain Mccowan, J Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, M Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, D Reidsma, and P Wellner. The ami meeting corpus. In *International Conference on Methods and Techniques in Behavioral Research*, page 702 pp., Wageningen, Netherlands, 01 2005. Wageningen: Noldus Information Technology.

[MLZL02] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the Tenth ACM International Conference on Multimedia*, MULTIMEDIA '02, pages 533–542, New York, NY, USA, 2002. ACM.

[ONO04] M. Ozeki, Y. Nakamura, and Y. Ohta. Video editing based on behaviors-for-attention - an approach to professional editing using a simple scheme. In *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, volume 3, pages 2215–2218 Vol.3, June 2004.

[RBB08] Abhishek Ranjan, Jeremy Birnholtz, and Ravin Balakrishnan. Improving meeting capture by applying television production principles with audio and motion detection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 227–236, New York, NY, USA, 2008. ACM.

[Smi12] Tim J. Smith. The attentional theory of cinematic continuity. *Projections*, 6(1):1–27, 2012.

[WC15] Hui-Yin Wu and Marc Christie. Stylistic Patterns for Generating Cinematographic Sequences. In *4th Workshop on Intelligent Cinematography and Editing Co-Located w/ Eurographics 2015*, pages 47–53, Zurich, Switzerland, May 2015. Eurographics Association. The definitive version is available at http://diglib.eg.org/.

[ZCSG16] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 766–782, Cham, 2016. Springer International Publishing.