EUROGRAPHICS 9th Workshop on Cinematography and Editing (2020)
M. Christie, H.-Y. Wu, T.-Y. Li and V. Gandhi (Editors)

Joint Attention for Automated Video Editing (Invited Paper)

Hui-Yin Wu[1,2], Trevor Santarra[3,4], Michael Leece[3], Rolando Vargas[3], and Arnav Jhala[1]

[1]North Carolina State University, USA; [2]Université Côte d'Azur, Inria, France;
[3]University of California Santa Cruz, USA; [4]Unity Technologies, USA

Abstract

Joint attention refers to the shared focal points of attention for occupants in a space. In this work, we introduce a computational definition of joint attention for the automated editing of meetings in multi-camera environments from the AMI corpus. Using extracted head pose and individual headset amplitude as features, we developed three editing methods: (1) a naive audio-based method that selects the camera using only the headset input, (2) a rule-based edit that selects cameras at a fixed pacing using pose data, and (3) an editing algorithm using LSTM (Long-short term memory) learned joint-attention from both pose and audio data, trained on expert edits. The methods are evaluated qualitatively against the human edit, and quantitatively in a user study with 22 participants. Results indicate that LSTM-trained joint attention produces edits that are comparable to the expert edit, offering a wider range of camera views than audio, while being more generalizable as compared to rule-based methods.

This invited paper is an extension of our previous work [WJ18], and was accepted for publication as a full paper in the proceedings of the 2020 ACM Conference for Interactive Media Experiences: https://doi.org/10.1145/3391614.3393656

Keywords: smart conferencing, automated video editing, joint attention, LSTM

Overview

Joint attention is an element of human communication where the attention of the group is drawn collectively towards focal points in an environment through non-verbal processes such as gaze, voice, and gesture [Hea05].

The capacity to make detailed records of our daily events and the explosive growth of recorded data calls for smart methods that can understand context in videos, and automatically process and present data in a meaningful way, such as for digital archiving. While intelligent camera switching technology is available to some extent, it is based primarily on audio, movement, and other low level features of the video streams. Existing work shows that LSTMs have been effective not only in image recognition and NLP tasks, but also for video summarization [ZCSG16] due to their ability to model more long ranged variable dependencies, outside of a single frame. Using LSTMs for an even more complex task such as that of video editing would be both exciting and challenging.

In this work, we present joint attention as a metric for automated video editing in corporate meeting recordings. We selected an existing corpus of meeting videos: the AMI corpus †, established by the University of Edinburgh, where 100 hours of meetings were recorded in smart meeting rooms equipped with multiple cameras, individual headsets, and microphone arrays, along with slide data, and post-meeting annotations. We consider that each camera in the meeting room can analyze the head pose of participants occupying that video. Based on extracted head pose data and audio from individual headphones, we designed and implemented three automated editing methods: a naive audio-based edit, a rule-based edit on our joint attention metric, and an LSTM method that predicts the joint attention of the meeting at each time point. These three methods use audio data, pose data, and both respectively to produce an automated edit of meetings. The extracting 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. Each method is then evaluated qualitatively using film editing metrics against the human edit, and through a user evaluation.

† http://groups.inf.ed.ac.uk/ami/corpus/

References

[Hea05] HEAL J.: *Joint Attention: Communication and Other Minds: Issues in Philosophy and Psychology*. Oxford University Press, USA, 2005. 1

[WJ18] WU H.-Y., JHALA A.: A joint attention model for automated editing. In *Proceedings of the AIIDE 2018 Joint Workshop on Intelligent Narrative Technologies and Workshop on Intelligent Cinematography and Editing* (Edmonton, Canada, 2018), CEUR-WS. 1

[ZCSG16] ZHANG K., CHAO W.-L., SHA F., GRAUMAN K.: Video summarization with long short-term memory. In *Computer Vision – ECCV 2016* (Cham, 2016), Leibe B., Matas J., Sebe N., Welling M., (Eds.), Springer International Publishing, pp. 766–782. 1

DOI: 10.2312/wiced.20201131