Human and Computer Models of Video Understanding

15th May 2024

Workshop Report

Dr Frank Guerin

We ran a one-day workshop, with ten oral presentations from morning to late afternoon, and a poster session at lunch-time. Two of the ten presenters were invited speakers, and the other eight were from those who had submitted abstracts. We asked oral presenters to also bring posters in order to have further opportunities for discussion. We additionally had two poster-only presenters. All of our presenters came in person except for the final speaker who was unable to appear due to unforeseen circumstances, but he did provide a video of his talk, and then appeared for a live video link to answer audience questions. We had twenty five participants, all from UK institutes, with many different nationalities represented. The organisers were Frank Guerin (University of Surrey), Andrew Gilbert (University of Surrey), Quoc Vuong (Newcastle University).

Workshop aims

The core research question the workshop was concerned with was: How does the human brain understand people's activities in a video much better than existing computer systems? We invited participants from the science of human vision (psychology or brain sciences) and computer vision, focusing on understanding activities from video. To give some concrete examples: Humans can very quickly make accurate judgements about the activity happening in a video even if the quality of the video is poor, or the motions observed are ambiguous, for example, to discriminate hugging from fighting, or smoking from eating finger food. Computers cannot match human performance in these tasks, which are critical for applications in surveillance, monitoring safety and welfare in a care setting, or removing inappropriate videos from social media. We do not yet fully understand how humans perform these feats, nor how to make computer vision systems reach their performance.

Event themes

Human perception of movement features

The event opened with the invited speaker on the human vision aspect of the symposium: Prof. Frank Pollick, from the University of Glasgow. Prof. Pollick focused on the perception of human movement, also known as biological motion perception. Firstly this covered movement features and how they contribute to the categorisation of different movement styles. Secondly this examine different frameworks that have been proposed for the neural processing of movement and how these inform evaluation of our research into watching dance. Finally, Prof. Pollick considered individual differences in the perception of human movement, including studies on autism and the effect of CCTV-operator experience in judging harmful intent.

Finding and tracking objects in a 3D environment

Following this extensive introductory talk we had four speakers with shorter talks. The first two of these focused on how people find and track objects in a 3D environment. Professor Szonya Surant (Royal Holloway) opened this by looking at how our limited attention capacity means we only notice a fraction of what is in front of us, shaping our conscious experience. Creative content creators use techniques to direct attention, such as varying depth of field, which keeps parts of a scene in focus while blurring others. In a study using eye tracking in a 3D game environment, it was found that participants spent more time looking at the center of the screen when a shallow depth of field was applied during free exploration, demonstrating an effective method for guiding attention in virtual environments. Dr. Toby Perrett from the University of Bristol presented a method to mimic human spatial cognition, enabling the tracking of 3D objects that move out of sight. The method lifts partial 2D observations to 3D coordinates, matching them over time based on appearance and location, and maintaining these tracks even when objects are out of view. Testing on 100 EPIC-KITCHENS videos showed that the method accurately tracks the 3D location of objects, with 60% correctly positioned after being out of sight for two minutes.

Visual perception of humans conversing

The next two talks focused on visual perception of humans conversing. Professor Anthony Atkinson from Durham University presented two studies: The first study used brain imaging to show that specific brain regions, including the right anterior STS and parietal cortex, are activated when observing communicative interactions, even when not the focus of attention. The second study demonstrated that people are faster and more accurate at detecting interacting dyads compared to independently acting ones, indicating more efficient visual processing of social interactions. Dr. Tom Foulsham from the University of Essex reviewed studies on human eye movements when watching people in videos. Eye-tracking research shows that human faces attract attention, with fixations clustering on people in the scene. Viewers often focus on individuals who are dominant or prestigious in group conversations, anticipate changes in speakers, and are influenced by age and attractiveness in crowd scenes. The studies highlight "social attention" in dynamic situations, showing how observers select specific individuals and behavioural cues to interpret social interactions.

Lunchtime interactions and posters

Just before lunch we had brief presentations to overview the extra posters which were not full oral presentations, from Fatemeh Nazarieh and Aditya Humnabadkar. We then had a one and a half-hour lunch break and poster sessions, which provided a good deal of free time for interaction between participants, initially while seated at lunch, and thereafter at the poster boards.

Recent developments in multimodal vision-language models

Professor Shaogang Gong from Queen Mary University of London opened the afternoon talks with a discussion of challenges and progress in multimodal learning for video moment retrieval. Traditional deep learning relies on centralized, labeled big data and powerful GPUs, but privacy, energy consumption, and decentralized data ownership pose challenges to this approach. His talk focused on recent advances in using multimodal vision-language models for self-supervised learning of fine-grained video-language details without the need for extensive labelling.

From person re-identification to somatosensory cortex participation…

Following this we had the final four talks. Dr. Arindam Sikdar from Edge Hill University introduced the field of person re-identification, that is, accurately matching individuals across different camera views, a problem of great significance for surveillance. This problem poses significant challenges due to variations in pose, illumination, viewpoint, and notably, scale. He developed scale-invariant residual networks and a batch adaptive triplet loss function to enhance performance. Dr. Quoc Vuong from Newcastle University showed how people adapt to audio-visual asynchrony in speech videos. Using a continuous-judgment paradigm, participants watched videos of news reports with varying levels of asynchrony between audio and visual streams. They judged the synchrony by pressing/releasing a spacebar. Results showed that participants adapted to asynchrony over time, particularly when the visual stream led the auditory stream. The adaptation magnitude depended on the degree of asynchrony, with larger asynchrony producing more significant adaptation. This research highlights temporal adaptation in continuous speech videos, crucial for understanding auditory-visual integration. Filip Rybansky from Newcastle University explored semantic consistency in recognizing human actions in complex activities. Using 128 short videos from the Epic-Kitchens-100 dataset, an online experiment had participants describe the actions in each video using 2-3 words. Semantic consistency was measured using a sentence-BERT model to compute the similarity of responses. The study found 87 videos with high semantic consistency, indicating reliable recognizability. Further to this he presented results on masking out parts of videos for human subjects, which reveals the crucial parts of those videos, that aid human recognition. Dr. Nicholas Hedger from the University of Reading was the one speaker who could not make it in person, and appeared by video link. His talk revealed that the human extrastriate cortex is tiled with “somatosensory homunculi”, through movie-watching studies. By using connective field models to analyse responses during naturalistic viewing, it was found that higher levels of the visual hierarchy are influenced by both visual and somatosensory topographies. This demonstrates extensive multimodal tuning in the brain, indicating that somatosensory cortex participates in naturalistic vision. These findings highlight the integration of visual and body part information, supporting visually guided movements and rich sensory experiences. Initial data from a new 7T fMRI movie-watching dataset were also presented.

The two invited speakers, Profs. Pollick and Gong, and the other senior academics presenting, attended the entire workshop and provided valuable feedback to other presenters, interacting with the more junior researchers, to share their knowledge, during the breaks. Overall the workshop provided a valuable opportunity for junior attendees to get significant time with the more senior attendees, which is often not possible at a traditional conference, with a large group in attendance.

Next steps – Outcomes

The workshop enabled interaction between the two different disciplines involved: computer vision, and human perception (which primarily involves psychologists, but also neuroscientists to a lesser degree). Discussion around interdisciplinary work revealed the difficulties of publishing such work in the mainstream computer vision venues: because such work inevitably involves a detailed study on some types of video, it will not encompass the entirety of a benchmark video dataset, and computer vision venues tend to be exclusively concerned with performance measured on these benchmark datasets. The other direction is more promising: venues which publish on human perception are quite open to accepting work on computer models that model certain aspects of human abilities. Reflecting on the original research question motivating the workshop (How does the human brain understand people's activities in a video much better than existing computer systems?), we did not come out with a clear answer to this question. There is a lack of knowledge regarding the detail of how human perception accurately classifies specific categories of video and similarly a lack of knowledge of the weaknesses in computer classification of the same specific categories. This points to a direction for future study: to focus both computer vision investigation and human perception experiments on a shared set of categories, to provide detailed interpretations of errors where they arise, and thereby to compare and contrast the two. Some of the comments from workshop participants noted the dearth of work that actually connects the two fields; people normally work only in one field or the other.

Although the original question is not answered, the workshop also suggested a shift of emphasis: rather than focusing on the activity recognition or categorisation ability, it might be more insightful to focus on the learning process. We saw especially from Prof. Gong’s keynote (from the Computer Vision side) that the labelling of videos on a massive scale is problematic, and self-supervised approaches are more promising. This then puts a spotlight on the learning processes in humans, and how humans develop their recognition and categorisation abilities through exposure to moving images in natural everyday settings.

Here are some sample comments from workshop participants:

“Coming from the human-vision field, for me the workshop provided a great opportunity to learn about the latest development of machine learning and deep neural networks for action recognition. It was good to hear about other researchers’ work in human vision (and how others have been combining human and computer vision). Having a keynote speaker who is an expert in each field was very helpful. Finally, I think there was really good interactions between researchers from both human and computer vision, and this was particularly helpful for the early career attendees (PhD students and post-docs).”

“The workshop was really interesting. Apart from the presentations and posters I learned a lot from them, the focus on psychology was more than the machine learning side which from my perspective, it would have been more interesting if they were equally considered.”

“The presentations were interesting, providing works from both sides of the fields and provoked some good discussions. There is still a gap for a workshop on joined human and computer models.”

“I was looking forward to seeing more work on integrating human and computer vision rather than each in isolation. The work was very interesting nonetheless.”

“I learned lots from this academic salon, nothing bad to me.”

Acknowledgements

This workshop was funded through an Institute of Advanced Studies (IAS) Workshop Grant to Dr. Frank Guerin. We would like to express our gratitude to Mirela Dumic and Louise Jones from IAS for their support and advice throughout the process. We also got help from staff in our own school (CSEE) to organise poster boards (Dan Gardham), badges, and for knowledge regarding university systems, from Emma Wittmann, Denise Myers and Kelly Green. We are thankful to all the speakers who contributed their valuable knowledge to the discussions. Special thanks to Prof. Shaogang Gong and Prof. Frank Pollick for their research insights and inspiring talks. Dr. Frank Guerin led the organisation of the workshop along with Dr. Andrew Gilbert and Dr. Quoc Vuong. We are also grateful for the administrative support provided by Dr. Joey Lam, whose help was invaluable for the preparation and smooth running of the day-long event.