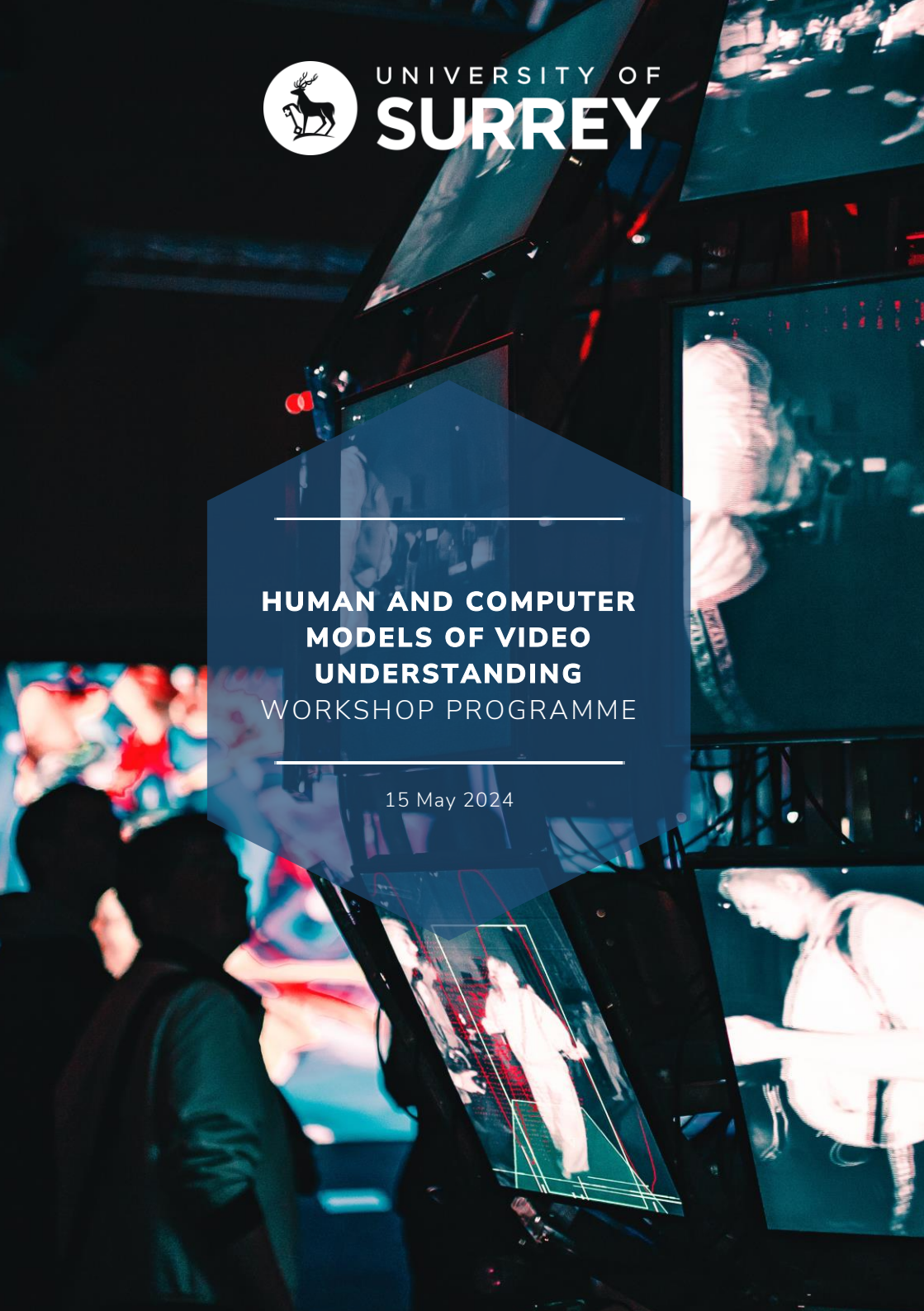


HUMAN AND COMPUTER MODELS OF VIDEO UNDERSTANDING

WORKSHOP PROGRAMME

15 May 2024

OUR SPONSOR

IAS Institute of Advanced Studies

The Institute of Advanced Studies (IAS) at the University of Surrey sponsors workshops and Fellowships at the 'cutting edge' of science, engineering, social science and the humanities. Through this scheme the Institute fosters interdisciplinary collaborations and encourages a flow of international scholars to visit, enjoy their stay at Surrey and leave behind excellent ideas and innovations.

ias.surrey.ac.uk

INTRODUCTION

How does the human brain understand people's activities in a video much better than existing computer systems? 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.

Our leading question is a research challenge for both human vision scientists and computer vision scientists to collaborate on. While there have been significant advances in both fields, there has been little cross-fertilisation of ideas between the separate groups. On one side human vision scientists have made advances in understanding the brain's approach of separating processing into a hierarchical form and motion stream corresponding to the structure of visual cortices, and have made many recent advances in understanding what people attend to, using eye-tracking technology. From the computer vision side rapid advances have been made in the last decade with the explosion of deep learning techniques, which can capture features that humans might attend to in video, and capture temporal relationships. However, it is still the case that the best computer systems to recognise activities in videos fall very far short of human performance.

The two communities can benefit by sharing ideas, and incorporating insights from each other's research. But we believe that much more can be gained from actually working together. Specifically we mean that the techniques used by psychologists in targeted experiments, for example to elucidate the objects and relationships that humans look at in a video can serve the needs of a computer vision project that seeks to improve recognition of a particular activity. In the other direction, the techniques used by computer vision researchers to detect particular features, and capture spatio-temporal relationships among elements can serve the needs of a human vision project that seeks to test hypotheses about a particular model of human vision, via a working model.

Workshop Chair:
Dr Frank Guerin, University of Surrey

Organising committee:
Dr Andrew Gilbert, University of Surrey and Dr Quoc Vuong, Newcastle University

Administrative support:
Louise Jones, Institute of Advanced Studies and Dr Joey Lam, University of Surrey

PROGRAMME

WEDNESDAY 15 MAY

Leggett Building, Manor Park Campus

(BST)

09.30 – 10.00 Registration Tea and Coffee

10.00 – 11.00 **Invited Speaker:**
"A Systems View of Perceiving Biological Motion: from Features to Brain Circuits" - Professor Frank Pollick

11.00 – 12.20 Oral Presentation Session:

"The Effects of Depth of Field on Attention whilst Exploring a Virtual Environment" *S. Durant, D. Gulhan, H. Kaur Suri, K. Vekariah, J. Haider-Smith* - Szonya Durant

"Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind" *T. Perrett, D. Damen, A. Kanazawa, S. Goel, C. Plizzari* - Toby Perrett

"On the Incidental and Deliberate Visual Processing of Communicative Interactions" *A. Atkinson, Q. Vuong* - Anthony Atkinson

"Looking at People in Videos: Evidence from Human Eye Movements" *T. Foulsham* - Tom Foulsham

12.20 – 12.30 Poster Announcements:

"StableTalk: Advancing Audio-to-Talking Face Generation with Stable Diffusion and Vision Transformer" - Fatemeh Nazarieh

"Advancements in 3D Plant Phenotyping: Precise Part Segmentation and Trait Measurement Through Video-Derived Point Clouds" - Reena Reena

"Driving Through Graphs: A Networked Perspective on Scene Representation" - Aditya Humnabadkar

12.30 – 13.00 Sandwich/Buffet Lunch in the Venue

13.00 – 13.30 Poster Session 1:
Szonya Durant, Toby Perrett, Anthony Atkinson, Tom Foulsham and Fatemeh Nazarieh

13.30 – 14.00 Poster Session 2:
Reena Reena, Aditya Humnabadkar, Arindam Sikdar, Quoc Vuong and Filip Rybansky

14.00 – 15.00 **Invited Speaker:**
Multimodal Learning in Video Moment Retrieval - Professor Shaogang Gong

15.00 – 15.30 Afternoon Coffee

15.30 – 17.00 Oral Presentation Session:

"Scale-Invariant Batch-Adaptive Residual Learning for Person Re-Identification" *A. Sikdar, A. S. Chowdhury* - Arindam Sikdar

"Measuring Temporal Adaptation in Videos of Speech" *Q. Vuong, M. Laing, V. Bansal, A. Rees* - Quoc Vuong

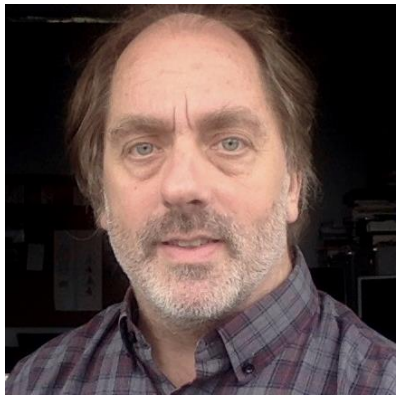
"Semantic Consistency in Identifying Human Actions" *F. Rybansky, S. Rahmani, A. Gilbert, F. Guerin, Q. Vuong* - Filip Rybansky

"Movie-Watching Reveals that Human Extrastriate Cortex is Tiled with Somatosensory Homunculi" *N. Hedger, T. Knapen* - Nicholas Hedger

17.00 Close

INVITED SPEAKERS

Professor Frank Pollick



Frank Pollick is a Professor of Psychology in the School of Psychology and Neuroscience at the University of Glasgow. He serves as the Director of Innovation, Engagement and Enterprise for the School, and is co-author of the textbook Cognitive Psychology published by McGraw Hill Education. He has a diverse set of research interests that include how we perceive human movement and how this ability varies with expertise, disease and brain development. He also studies how people interact with technology; neuroergonomics and the use of realtime fMRI neurofeedback to understand cognition. He contributes editorial service to the International Journal of Humanoid Robots and Technology, Mind, and Behaviour.

A Systems View of Perceiving Biological Motion: from Features to Brain Circuits

Professor Frank Pollick, School of Psychology and Neuroscience, University of Glasgow

This talk will present research results from our lab that have explored the perception of human movement – also known as biological motion perception. In the first part of the talk, I will discuss movement features and how they contribute to the categorisation of different movement styles. Next, I will examine the different frameworks that have been proposed for the neural processing of movement and how these inform evaluation of our research into watching dance. Finally, I will address how individual differences might impact the perception of human movement by reviewing our studies on autism and the effect of CCTV-operator experience in judging harmful intent. Throughout the talk I will highlight ways in which quantitative measures of human movement have been related to perception and underlying brain activity.

Professor Sean Gong

Professor Sean Gong FREng is a computer vision and machine learning scientist. He pioneered person re-identification and video behaviour analysis for law enforcement. Prof Gong is elected a Fellow of the Royal Academy of Engineering, and served on the steering panel of the UK government Chief Scientific Adviser's Science Review on Security. He has made unique contributions to the engineering of AI video analytics for law enforcement and the security industry and was awarded an Institution for Engineering and Technology Achievement Medal for Vision Engineering for outstanding achievement and superior performance in contributing to public safety. A commercial system built based on his research won an Aerospace Defence Security Innovation Award and a Global Frost & Sullivan Award for Technical Innovation for Law Enforcement Video Forensics Technology. Gong is Professor of Visual Computation and Director of the Computer Vision Laboratory at Queen Mary University of London, a Fellow of ELLIS (European Laboratory for Learning and Intelligent Systems), a Fellow of AAIA (Asia-Pacific Artificial Intelligence Association), a Turing Fellow of the Alan Turing Institute, a Fellow of the Institution of Electrical Engineers, and a member of the UK Computing Research Committee. He founded Vision Semantics and served as the Chief Scientist of three start-ups, two of which have been acquired by NASDAQ listed companies. He is a Distinguished Scientist of Veritone. He received his DPhil from Oxford University.

Multimodal Learning in Video Moment Retrieval

Professor Shaogang Gong, Queen Mary University of London, Queen Mary Computer Vision Laboratory

Deep learning has revolutionised AI machine learning techniques in computer vision over the past decade largely due to the availability of centralised big data with exhaustive labelling and cheap computing power from Nvidia's GPUs. However, privacy concerns from data protection and environmental concerns on energy consumption together with an increasing demand for decentralised user-ownership of localised unlabelled data pose fundamental challenges to the established wisdom of deep learning on centralised big data from scratch with exhaustive labelling available for model training. In this talk, I will present challenges and recent progress on exploring multimodal vision-language models for self-supervised learning of fine-grained video-language dynamic details without fine-grained labelling in model training for video moment retrieval.

ORAL PRESENTATIONS

The Effects of Depth of Field on Attention whilst Exploring a Virtual Environment

Professor Szonya Durant, Royal Holloway, University of London

We do not notice everything in front of us, due to our limited attention capacity. What we attend to forms our conscious experience and is what we retain over time. Thus, creative content creators must strive to direct your attention in different media, from cinema to computer games. To do this they have developed various techniques that involve either directly using centrally presented cues such as arrows or instructions to move attention or rely on image features or so- called "bottom-up" cues that involve manipulating the salience of the parts of an image. Shifting attention usually involves moving our central vision around a screen, but this problem becomes more pronounced in virtual environments where users are free to explore by moving in any direction through it. This can be seen in first- person view screen- based computer video games. Such an experience allows the user to choose how they sample their environment. Often the designer of the environment wishes the user to interact and view certain parts of the scene. In this study we test out a subtle manipulation of visual attention through varying depth of field. Varying depth of field is a cinematic technique that can be implemented in virtual worlds and involves keeping parts of the scene in focus whilst blurring other parts of the scene. We use eye tracking to investigate this technique in a 3D game environment, rendered on a monitor screen. Participants navigated through the environment using keyboard keys and began by freely exploring in the first part and in the second part were instructed to find a target object. We manipulated whether the frames were rendered fully in focus (termed a deep depth of field) or whether a shallow depth of field was applied (where the outer edges of the scene appear blurred. We measured where on the screen participants looked. We divided the screen into 3x3 equal sized regions and calculated the proportion of the time participants spent looking in the central square. On average across all trials participants spent 67% of their fixation time on the central area of the screen. This means that they preferred to navigate by looking in the direction they were heading in. We found that there was a significant difference when freely exploring the scene – participants spent more time looking in the centre of the screen when a shallow depth of field was applied than with a deep depth of field. This was no longer the case during the search task. We demonstrate how these techniques might be effective for manipulating attention by keeping user's eyes looking straight ahead when they are freely exploring a virtual environment.

Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind

Dr Toby Perrett, University of Bristol

As humans move around, performing their daily tasks, they are able to recall where they have positioned objects in their environment, even if these objects are currently out of sight. In this paper, we aim to mimic this spatial cognition ability. We thus formulate the task of Out of Sight, Not Out of Mind - 3D tracking active objects using observations captured through an egocentric camera. We introduce Lift, Match and Keep (LMK), a method which lifts partial 2D observations to 3D world coordinates, matches them over time using visual appearance, 3D location and interactions to form object tracks, and keeps these object tracks even when they go out-of-view of the camera - hence keeping in mind what is out of sight.

We test LMK on 100 long videos from EPIC-KITCHENS. Our results demonstrate that spatial cognition is critical for correctly locating objects over short and long time scales. E.g., for one long egocentric video, we estimate the 3D location of 50 active objects. Of these, 60% can be correctly positioned in 3D after 2 minutes of leaving the camera view.

On the Incidental and Deliberate Visual Processing of Communicative Interactions

Professor Anthony Atkinson, Durham University

The interpretation of social interactions between people is important in many daily situations. In this talk, we will present the results of 2 studies examining the visual perception of other people interacting. The first study used functional brain imaging to investigate the brain regions involved in the incidental visual processing of social interactions; that is, the processing of the body movements outside the observers' focus of attention. The second study used a visual search paradigm to test whether people are better able to find interacting than non-interacting people in a crowd. In the first study, we measured brain activation while participants (N = 33) were presented with point-light dyads portraying communicative interactions or individual actions. These types of stimuli allowed us to investigate the role of motion in processing social interactions by removing form cues. Participants in our study discriminated the brightness of two crosses also on the screen, thus excluding the body movements from the participant's task-related focus of attention. To investigate brain regions that may process the spatial and temporal relationships between the point-light displays, we either reversed the facing direction of one agent or spatially scrambled the local motion of the points. Incidental processing of communicative interactions elicited activation in right anterior STS only when the two agents were facing each other. Controlling for differences in local motion by subtracting brain activation to scrambled versions of the point-light displays revealed significant activation in parietal cortex for communicative interactions, as well as left amygdala and brain stem/cerebellum. Our results complement previous studies and suggest that additional brain regions may be recruited to incidentally process the spatial and temporal contingencies that distinguish people acting together from people acting individually.
Our second study focussed on deliberate visual processing of communicative

Continued ▶

interactions in the observer's focus of attention. Participants viewed arrays of the same point-light dyads used in our first study, but here they searched for an interacting dyad amongst a set of independently acting dyads, or for an independently acting dyad amongst a set of interacting dyads, by judging whether a target dyad was present or absent (targets were present on half the trials). In each of two experiments (N=32 and N=49), participants were faster and more accurate to detect the presence of interacting than independently acting target dyads. Moreover, visual search for interacting target dyads was more efficient than for independently acting target dyads, as indicated by shallower search slopes (increase in response time with increasing number of distractors) for the former as for the latter. In the second experiment, we measured the eye movements of the participants using an eye tracker. The analyses of the eye tracking data are ongoing. Based on the results from our first study and on search performance, we expect that fixation duration on communicative-dyad targets will be shorter than on independent-dyad targets, because less attentional focus (as measured by fixation duration) is needed to process social interactions.

Looking at People in Videos: Evidence from Human Eye Movements

Dr Tom Foulsham, University of Essex

A large body of research demonstrates that human faces attract attention. This is also true in video, with eye-tracking studies showing that fixations are clustered on the people in the scene. I will review a number of studies from psychology investigating where observers look when watching people in video, and how this is related to their understanding and judgements of the scene.

For example, participants are quite consistent in their judgements of people having a conversation, even from a brief video clip ("thin-slicing"). We have shown that when watching such clips the individuals who receive the most attention are those who are rated as the most dominant or prestigious in the group. This is one case where participants must make a decision about who to fixate at each point in time, and this appears to vary due to both the behaviour of the actor and the characteristics of the observer. In crowd scenes featuring multiple people, certain targets get fixated more often than others and this is affected by age and attractiveness.

When observing pre-recorded conversation, participants spend most of the time fixating the person speaking, but they can also anticipate the change in speaker. This may be partly due to reading the cues provided by interacting participants (such as their gaze). We can manipulate these cues to study signalling in a naturalistic setting, and we have also shown nuanced differences in observers high in autistic and ADHD-related traits. Importantly, we also find a close correspondence between fixations on pre-recorded videos and the gaze displayed by participants in a real face-to-face interaction, which suggests a high level of ecological validity.

Taken together, these studies show "social attention" operating in complex and dynamic situations. This involves not just looking at other people, but selecting specific individuals, and specific behavioural cues, in order to interpret the scene.

Scale-Invariant Batch-Adaptive Residual Learning for Person Re-Identification

Dr Arindam Sikdar, Edge Hill University

In the field of person re-identification (re-ID), accurately matching individuals across different camera views poses significant challenges due to variations in pose, illumination, viewpoint, and notably, scale. Traditional methods in re-ID have focused on robust feature descriptor generation and sophisticated metric learning, yet they often fall short in addressing scale variations effectively. In this work, we introduce a novel approach to scale-invariant person re-ID through the development of our scale-invariant residual networks coupled with an innovative batch adaptive triplet loss function for enhanced deep metric learning. The first network, termed Scale-Invariant Triplet Network (SI-TriNet), leverages pre-trained weights to form a deeper architecture, while the second, Scale-Invariant Siamese Resnet-32 (SISR-32), is a shallower structure trained from scratch. These networks are adept at handling scale variations, a common yet challenging aspect in re-ID tasks, by employing scale-invariant (SI) convolution techniques that ensure robust feature detection across multiple scales. This is complemented by our proposed batch adaptive triplet loss function that refines the metric learning process, dynamically prioritizing learning from harder positive samples to improve the model's discriminatory capacity. Extensive evaluation on benchmark datasets Market-1501 and CUHK03 demonstrates the superiority of our proposed methods over existing state-of-the-art approaches. Notably, SI-TriNet and SISR-32 show significant improvements in both mean accuracy metrics, affirming the efficacy of our scale-invariant architectures and the novel loss function in addressing the complexities of person re-ID. This study not only advances the understanding of scale-invariant feature learning in deep networks but also sets a new benchmark in the person re-ID domain, promising more accurate and scalable solutions for real-world surveillance and security applications.

Measuring Temporal Adaptation in Videos of Speech

Dr Quoc Vuong, Newcastle University

People are increasingly consuming video media through the internet. This can lead to a mis-match between the auditory and visual streams due to internet connectivity. For instance in a video of a news anchor reporting a story, there can be time lag between the spoken words and the corresponding movements of the anchor's mouth and lips (as well as body gestures). This asynchrony between the auditory and visual streams can also arise due to various physical, bio-physical and neural mechanisms, but people are often not aware of these differences. There is accumulating evidence that people adapt to auditory-visual asynchrony at different time scales and for different stimulus categories. However, previous studies often used very simple auditory-visual stimuli (e.g., flashing lights paired with brief tones) or they used short videos of a few seconds. In the current study, we investigated the temporal adaption of continuous speech presented in longer videos. Speech is one of the strongest case for auditory-visual integration, as demonstrated by multi-sensory illusions like the McGurk-McDonald and

Continued ▶

Ventriloquist effects. To measure temporal adaption of speech videos, we developed a continuous-judgment paradigm in which participants continuously judge over several tens of seconds whether an auditory-visual speech stimulus is synchronous or not. The stimuli consisted of 40 videos (duration: M = 63.3s, SD = 10.3s). For each video, we filmed a close-up (upper body) of one male and one female speaker reporting a news story transcribed from a real news clip (e.g., about the Brexit vote outcome or about Boris Johnson's resignation). Each speaker reported 20 news stories. We then created seven asynchronous versions of each video by shifting the relative stimulus onset asynchrony (SOA) between the auditory and visual streams between -240ms (auditory stream leading) to +240ms (visual stream leading) in 80ms steps. This included SOA = 0ms (i.e., the original synchronous video). The first 5-10s of all videos were synchronous. For each participant in the continuous-judgment task, we randomly selected 10 videos at each SOA (70 total). Participants continuously judged the synchrony of each video by pressing/releasing the spacebar throughout the duration of the video (response sampling rate = 33ms). The mean proportion perceived synchrony across the duration of the videos were calculated from participants' continuous responses after the initial synchronous period. For the auditory-leading videos (SOAs 0ms), participants initially showed a drop in proportion perceived synchrony but this proportion increased over time, suggesting that they were adapting to the asynchrony over time. The magnitude of temporal adaptation depended on the SOA, with the largest SOA producing the largest adaptation. Consistent with previous studies, our findings suggest that temporal adaptation occurs for long, continuous speech videos but only when the visual stream leads the auditory stream.

Semantic Consistency in Identifying Human Actions

Filip Rybansky, Newcastle University

People quickly recognise human actions carried out in everyday activities. There is evidence that Minimal Recognisable Configurations (MIRCs) contain a combination of spatial and temporal visual features critical for reliable recognition. For complex activities, observers may have different descriptions varied in their semantic similarity (e.g., washing dishes vs cleaning dishes), potentially complicating the investigation of MIRCs in action recognition. Therefore, we measured the semantic consistency for 128 short videos of complex actions from the Epic-Kitchens-100 dataset (Damen et al., 2022), selected based on poor classification performance by our state-of-the-art computer vision network MOFO (Ahmadian et al., 2023). In an online experiment, participants viewed each video and identified the performed action by typing a description using 2-3 words (capturing action and object). Each video was classified by at least 30 participants (N=76 total). Semantic consistency of the responses was determined using a custom pipeline involving the sentence-BERT language model, which generated embedding vectors representing semantic properties of the responses. We then used adjusted pair-wise cosine similarities between response vectors to compute a ground truth description for each video, a response with the greatest semantic neighbourhood density (e.g., pouring oil, closing shelf). The greater the semantic neighbourhood density was for a ground truth candidate, the more semantically consistent were responses for the associated video. We uncovered 87 videos where semantic consistency confirmed their reliable recognisability, i.e. where cosine-similarity between the ground truth candidate and at least 70% of responses was above a similarity threshold of 0.65. We will use a subsample of these videos to investigate the role of MIRCs in human action recognition, e.g., gradually degrading the spatial and temporal information in videos and measuring the impact on action recognition. The derived semantic space and MIRCs will be used to revise MOFO into a more biologically consistent and better performing model.

Continued ►

Movie-Watching Reveals that Human Extrastriate Cortex is Tiled with Somatosensory Homunculi

Dr Nicholas Hedger, University of Reading

Our rich, embodied visual experiences of the world involve integrating information from multiple sensory modalities - yet how the brain brings together multiple sensory reference frames to generate such experiences remains unclear. Recently, it has been demonstrated that BOLD fluctuations throughout the brain can be explained as a function of the activation pattern on the primary visual cortex (V1) topographic map. This class of 'connective field' models allow us to project V1's map of visual space into the rest of the brain and discover previously unknown visual organization. Here, we extend this powerful principle to incorporate both visual and somatosensory topographies by explaining BOLD responses during naturalistic movie-watching as a function of two spatial patterns (Connective fields) on the surfaces of V1 and S1. We show that responses in the higher levels of the visual hierarchy are characterized by multimodal topographic connectivity: these responses can be explained as a function of spatially specific activation patterns on both the retinotopic and somatosensory homunculus topographies, indicating that somatosensory cortex participates in naturalistic vision. These novel multimodal tuning profiles are in line with known visual category selectivity, for example for faces and manipulable objects. Our findings demonstrate a scale and granularity of multisensory tuning far more extensive than previously assumed. When inspecting their topographic tuning in S1, we find a full band extrastriate visual cortex from retrosplenial, laterally to the fusiform gyrus, is tiled with somatosensory homunculi. These results demonstrate the intimate integration of information about visual coordinates and body parts in the brain that likely supports visually guided movements and our rich, embodied experience of the world. Finally, we present initial data from a new, densely sampled 7T fMRI movie-watching dataset optimised to shed light on the brain basis of human action understanding.

POSTER ANNOUNCEMENTS

StableTalk: Advancing Audio-to-Talking Face Generation with Stable Diffusion And Vision Transformer
Fatemeh Nazarieh, University of Surrey

Audio-to-talking face generation stands at the forefront of advancements in generative AI. It bridges the gap between audio and visual representations by generating synchronized and realistic talking faces. This significantly improves human-computer interaction and content accessibility for diverse audiences. Despite substantial research in this area, critical challenges such as the lack of realistic facial animations, inaccurate audio-lip synchronization, and intensive computational demands continue to restrict the practicality of the talking face generation methods applications. To address these issues, we introduce a novel approach leveraging the emerging capabilities of Stable diffusion models and vision Transformers for Talking face generation (StableTalk). By incorporating the Re-attention mechanism and adversarial loss into StableTalk, we have markedly enhanced the audio-lip alignment and the consistency of facial animations across frames. More importantly, we have optimized computational efficiency by refining operations within the latent space and dynamically adjusting the visual focus based on the given conditions. Our experimental results demonstrate that StableTalk surpasses existing methods in terms of image quality, audio-lip synchronization, and computational efficiency.

Advancements in 3D Plant Phenotyping: Precise Part Segmentation and Trait Measurement Through Video-Derived Point Clouds
Reena Reena, Edge Hill University

This research represents a groundbreaking approach in plant phenotyping by harnessing 3D point clouds generated from video data. Focusing on the comprehensive characterization of plant traits, this method enhances the precision and depth of phenotypic analysis, crucial for advancements in genetics, breeding, and agricultural practices.

Advanced Video Data Capture and Processing for Detailed Segmentation

High-Fidelity Video Acquisition: Capturing detailed video footage of plants under varying environmental conditions forms the foundation of this method. The use of high-resolution cameras allows for capturing minute details crucial for accurate part segmentation.

Rigorous Preprocessing for Optimal Data Quality: Following capture, the video data undergoes meticulous preprocessing. Stabilization, noise filtering, and color correction are performed to ensure that the subsequent segmentation algorithms can accurately identify different parts of the plant.

Segmentation and 3D Point Cloud Generation: The application of state-of-the-art image processing algorithms segments the plant parts within each video frame. Subsequently, photogrammetry and depth estimation techniques create detailed 3D point clouds, effectively capturing the geometry of individual plant components.

Continued ▶

Part Segmentation and Trait Measurement for Enhanced Phenotyping

Precise Plant Part Segmentation: This methodology enables the accurate segmentation of individual plant parts, such as leaves, stems, and flowers, within the 3D space. This precise segmentation is crucial for assessing complex plant traits and understanding plant structure in its entirety.

Comprehensive Trait Measurement: The 3D point clouds facilitate comprehensive measurements of plant traits. This includes quantifying leaf area, stem thickness, flower size, and even more subtle features like leaf venation patterns, providing a multi-dimensional view of plant phenotypic traits.

Temporal Tracking for Dynamic Trait Analysis: An integral advantage of using video data is the ability to track and measure these traits over time. This dynamic analysis allows for monitoring growth patterns, developmental changes, and responses to environmental stimuli in a way that static images cannot achieve.

Conclusion: A Breakthrough in Plant Phenotyping and Agricultural Research

This research significantly enhances the capability for detailed plant part segmentation and trait measurement, setting a new standard in plant phenotyping. The level of detail and accuracy afforded by this method offers invaluable insights for agricultural technology, plant genetics, and breeding programs. It represents a critical step forward in our ability to understand and optimize plant characteristics, with far-reaching implications for food production and ecological sustainability.

"Driving Through Graphs": A Networked Perspective on Scene Representation
Aditya Humnabadkar, Edge Hill University

In the evolving landscape of traffic management and autonomous driving technology, the analysis of traffic scenes from video data stands as a crucial challenge. Traditional approaches often rely on complex, high-dimensional image analysis, necessitating significant computational resources and sophisticated algorithms. Recognizing the limitations of these methods, our research introduces a novel, streamlined approach centered around a graph-based framework for understanding traffic dynamics.
Central to our methodology is the exploration of complex scene analysis through the lens of object-object interaction within traffic scenes. This interaction dynamics is adeptly captured through our specially designed graph structures, which are further analyzed and interpreted using Graph Neural Networks (GNNs) as a foundational element. By employing GNNs, our framework delves into the intricate dynamics of traffic environments. We

Continued ▶

focus on the high-level interactions and behaviours within traffic scenes, distilling the essential patterns of movement and relationships among elements such as vehicles and pedestrians.

To validate the effectiveness of our framework, we conducted extensive testing using two prominent datasets: the METEOR Dataset and the INTERACTION Dataset. Our methodology demonstrated exceptional performance, achieving an accuracy of 62.03% on the METEOR Dataset and an impressive 98.50% on the INTERACTION Dataset. These results underscore the capability of our graph-based approach to accurately interpret and analyze the dynamics of traffic scenes.

Through this rigorous evaluation, our research not only showcases the significant advantages of incorporating graph neural networks for traffic scene analysis but also highlights the power of our novel approach in abstracting and understanding the complex patterns of movement and interactions within traffic environments. Our work sets a new benchmark in the field, offering a promising direction for future advancements in traffic management and autonomous vehicle technologies.

FACULTY OF ENGINEERING AND PHYSICAL SCIENCE

University of Surrey
Guildford, GU2 7XH, UK

surrey.ac.uk