AI4ME
AI for Personalised Media Experiences

AI4ME PUBLICATIONS TO DATE SEPTEMBER 2024

ForecasterFlexOBM: A multi-view audio-visual dataset for flexible object-based media production

Davide Berghi, Craig Cieciura, Farshad Einabadi, Maxine Glancy, Oliver Charles Camilleri, Philip Anthony Foster, Asmar Nadeem, Faegheh Sardari, Jinzheng Zhao, Marco Volino. Proceedings from 2024 IEEE International Conference on Multimedia and Expo (Niagra Falls, Ontario, Canada). Access paper here.

Link to dataset web page here.

Leveraging machine learning techniques, in the context of object-based media production, could enable provision of personalized media experiences to diverse audiences. To fine-tune and evaluate techniques for personalization applications, as well as more broadly, datasets which bridge the gap between research and production are needed. We introduce and publicly release such a dataset, themed around a UK weather forecast and shot against a blue-screen background, of three professional actors/presenters – one male and one female (English) and one female (British Sign Language). Scenes include both production and research-oriented examples, with a range of dialogue, motions, and actions. Capture techniques consisted of a synchronized 4K resolution 16-camera array, production-typical microphones plus professional audio mix, a 16-channel microphone array with collocated Grasshopper3 camera, and a photogrammetry array. We demonstrate applications relevant to virtual production and creation of personalized media including neural radiance fields, shadow casting, action/event detection, speaker source tracking and video captioning.

Authoring Inter-Compatible Flexible Audio for Mass Personalization.

Craig Cieciura, Elettra Bargiacchi and Philip J B Jackson. Accepted to the 157th Audio Engineering Society Convention in October this year. Not yet published.

Bestiari: a hypnagogic experience created by combining complementary state-of-the-art spatial sound technologies, Catalan Pavilion, Venice Art Biennale 2024

Craig Cieciura, Anthony Myatt, Carlos Casas, Armand Lesecq, Filipa Ramos and Philip J B Jackson. Accepted to the 157th Audio Engineering Society Convention in October this year. Not yet published.

CoLeaF: A Contrastive-Collaborative Learning Framework for Weakly Supervised Audio-Visual Video Parsing

Faegheh Sardari, Armin Mustafa, Philip JB Jackson, Adrian Hilton. Proceedings of the ECCV 2024. Access paper here.

Weakly supervised audio-visual video parsing (AVVP) methods aim to detect audible-only, visible-only, and audible-visible events using only video-level labels. Existing approaches tackle this by leveraging unimodal and cross-modal contexts. However, we argue that while cross-modal learning is beneficial for detecting audible-visible events, in the weakly supervised scenario, it negatively impacts unaligned audible or visible events by introducing irrelevant modality information. In this paper, we propose CoLeaF, a novel learning framework that optimizes the integration of cross-modal context in the embedding space such that the network explicitly learns to combine cross-modal information for audible-visible events while filtering them out for unaligned events. Additionally, as videos often involve complex class relationships, modelling them improves performance. However, this introduces extra computational costs into the network. Our framework is designed to leverage cross-class relationships during training without incurring additional computations at inference. Furthermore, we propose new metrics to better evaluate a method's capabilities in performing AVVP. Our extensive experiments demonstrate that CoLeaF significantly improves the state-of-the-art results by an average of 1.9% and 2.4% F-score on the LLP and UnAV-100 datasets, respectively.

ForecasterFlexOBM: A multi-view audio-visual dataset for flexible object-based media production.

Berghi, D., Cieciura, C., Einabadi, F., Glancy, M., Camilleri, O. C., Foster, P. A., Nadeem, A., Sardari, F., Zhao, J., & Volino, M. ICME 2024. Access paper here

Leveraging machine learning techniques, in the context of object-based media production, could enable provision of personalized media experiences to diverse audiences. To fine-tune and evaluate techniques for personalization applications, as well as more broadly, datasets which bridge the gap between research and production are needed. We introduce and publicly release such a dataset, themed around a UK weather forecast and shot against a blue-screen background, of three professional actors/presenters – one male and one female (English) and one female (British Sign Language). Scenes include both production and research-oriented examples, with a range of dialogue, motions, and actions. Capture techniques consisted of a synchronized 4K

resolution 16-camera array, production-typical microphones plus professional audio mix, a 16-channel microphone array with collocated Grasshopper3 camera, and a photogrammetry array. We demonstrate applications relevant to virtual production and creation of personalized media including neural radiance fields, shadow casting, action/event detection, speaker source tracking and video captioning.

SEM-POS: Grammatically and Semantically Correct Video Captioning.

Asmar Nadeem, Adrian Hilton, Robert Dawes, Graham Thomas, Armin Mustafa; Proceedings of the IEEE/CVF. Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 2606-2616 Access paper here

Generating grammatically and semantically correct captions in video captioning is a challenging task. The captions generated from the existing methods are either word-by-word that do not align with grammatical structure or miss key information from the input videos. To address these issues, we introduce a novel global-local fusion network, with a Global-Local Fusion Block (GLFB) that encodes and fuses features from different parts of speech (POS) components with visual-spatial features. We use novel combinations of different POS components - 'determinant + subject', 'auxiliary verb', 'verb', and 'determinant + object' for supervision of the POS blocks - Det + Subject, Aux Verb, Verb, and Det + Object respectively. The novel global-local fusion network together with POS blocks helps align the visual features with language description to generate grammatically and semantically correct captions. Extensive qualitative and quantitative experiments on benchmark MSVD and MSRVTT datasets demonstrate that the proposed approach generates more grammatically and semantically correct captions compared to the existing methods, achieving the new state-of-the-art. Ablations on the POS blocks and the GLFB demonstrate the impact of the contributions on the proposed method.

CAD-contextual multi-modal alignment for dynamic AVQA

Asmar Nadeem, Adrian Hilton, Robert Dawes, Graham Thomas, Armin Mustafa. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024. Access paper here.

In the context of Audio Visual Question Answering (AVQA) tasks, the audio and visual modalities could be learnt on three levels: 1) Spatial, 2) Temporal, and 3) Semantic. Existing AVQA methods suffer from two major shortcomings; the audio-visual (AV) information passing through the network isn't aligned on

Spatial and Temporal levels; and, inter-modal (audio and visual) Semantic information is often not balanced within a context; this results in poor performance. In this paper, we propose a novel end-to-end Contextual Multi-modal Alignment (CAD) network that addresses the challenges in AVQA methods by i) introducing a parameter-free stochastic Contextual block that ensures robust audio and visual alignment on the Spatial level; ii) proposing a pre-training technique for dynamic audio and visual alignment on Temporal level in a self-supervised setting, and iii) introducing a cross-attention mechanism to balance audio and visual information on Semantic level. The proposed novel CAD network improves the overall performance over the state-of-the-art methods on average by 9.4% on the MUSIC-AVQA dataset. We also demonstrate that our proposed contributions to AVQA can be added to the existing methods to improve their performance without additional complexity requirements.

PAT: Position-Aware Transformer for Dense Multi-Label Action Detection

Sardari F, Mustafa A, Jackson P, Hilton A. Proceedings of the ICCV 2023 Workshop on AI for Creative Video Editing and Understanding. Access paper here

We present PAT, a transformer-based network that learns complex temporal co-occurrence action dependencies in a video by exploiting multi-scale temporal features. In existing methods, the self-attention mechanism in transformers loses the temporal positional information, which is essential for robust action detection. To address this issue, we (i) embed relative positional encoding in the self-attention mechanism and (ii) exploit multi-scale temporal relationships by designing a novel non hierarchical network, in contrast to the recent transformer-based approaches that use a hierarchical structure. We argue that joining the self-attention mechanism with multiple sub-sampling processes in the hierarchical approaches results in increased loss of positional information. We evaluate the performance of our proposed approach on two challenging dense multi-label benchmark datasets, and show that PAT improves the current state-of-the-art result by 1.1% and 0.6% mAP on the Charades and MultiTHUMOS datasets, respectively, thereby achieving the new state-of-the-art mAP at 26.5% and 44.6%, respectively. We also perform extensive ablation studies to examine the impact of the different components of our proposed network.

Learning Projective Shadow Textures for Neural Rendering of Human Cast Shadows from Silhouettes

Einabadi F, Guillemaut JY, Hilton A. Proceedings of the Eurographics Symposium on Rendering (EGSR) 2023. Access paper here.

This contribution introduces a two-step, novel neural rendering framework to learn the transformation from a 2D human silhouette mask to the corresponding cast shadows on background scene geometries. In the first step, the proposed neural renderer learns a binary shadow texture (canonical shadow) from the 2D foreground subject, for each point light source, independent of the background scene geometry. Next, the generated binary shadows are texture-mapped to transparent virtual shadow map planes which are seamlessly used in a traditional rendering pipeline to project hard or soft shadows for arbitrary scenes and light sources of different sizes. The neural renderer is trained with shadow images rendered from a fast, scalable, synthetic data generation framework. We introduce the 3D Virtual Human Shadow (3DVHshadow) dataset as a public benchmark for training and evaluation of human shadow generation. Evaluation on the 3DVHshadow test set and real 2D silhouette images of people demonstrates the proposed framework achieves comparable performance to traditional geometry-based renderers without any requirement for knowledge or computationally intensive, explicit estimation of the 3D human shape. We also show the benefit of learning intermediate canonical shadow textures, compared to learning to generate shadows directly in camera image space. Further experiments are provided to evaluate the effect of having multiple light sources in the scene, model performance with regard to the relative camera-light 2D angular distance, potential aliasing artefacts related to output image resolution, and effect of light sources' dimensions on shadow softness.

Deep Neural Models for Illumination Estimation and Relighting

Einabadi F, Guillemaut JY, Hilton A. Proceedings of the Computer Graphics Forum 2021. Access Papers here.

Scene relighting and estimating illumination of a real scene for insertion of virtual objects in a mixed-reality scenario are well-studied challenges in the computer vision and graphics fields. Classical inverse rendering approaches aim to decompose a scene into its orthogonal constituting elements, namely scene geometry, illumination and surface materials, which can later be used for augmented reality or to render new images under novel lighting or viewpoints. Recently, the application of deep neural computing to illumination estimation, relighting and inverse rendering has shown promising results. This contribution aims to bring together in a coherent manner current advances in this conjunction. We examine in detail the attributes of the proposed approaches, presented in three categories: scene illumination estimation, relighting with reflectance-aware scene-specific representations and finally relighting as image-to-image transformations. Each category is concluded with a discussion on the main characteristics of the current methods and possible future trends. We also provide an overview of current publicly available datasets for neural lighting applications.

Audio Inputs for Active Speaker Detection and Localization Via Microphone Array

Berghi D, Jackson P.J.B. Proceedings from the 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA. Access papers here.

This study considers the problem of detecting and locating an active talker's horizontal position from multichannel audio captured by a microphone array. We refer to this as active speaker detection and localization (ASDL). Our goal was to investigate the performance of spatial acoustic features extracted from the multichannel audio as the input of a convolutional recurrent neural network (CRNN), in relation to the number of channels employed and additive noise. To this end, experiments were conducted to compare the generalized cross-correlation with phase transform (GCC-PHAT), the spatial cue-augmented log-spectrogram (SALSA) features, and a recently-proposed beamforming method, evaluating their robustness to various noise intensities. The array aperture and sampling density were tested by taking subsets from the 16-microphone array. Results and tests of statistical significance demonstrate the microphones' contribution to performance on the TragicTalkers dataset, which offers opportunities to investigate audio-visual approaches in the future.

Fusion of Audio and Visual Embeddings for Sound Event Localization and Detection

Berghi D, Wu P, Zhao J, Wang W, Jackson P.J.B. Proceedings from the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, 2024. Access papers here.

Sound event localization and detection (SELD) combines two subtasks: sound event detection (SED) and direction of arrival (DOA) estimation. SELD is usually tackled as an audio-only problem, but visual information has been recently included. Few audio-visual (AV)-SELD works have been published and most employ vision via face/object bounding boxes, or human pose keypoints. In contrast, we explore the integration of audio and visual feature embeddings extracted with pre-trained deep networks. For the visual modality, we tested ResNet50 and Inflated 3D ConvNet (I3D). Our comparison of AV fusion methods includes the AV-Conformer and Cross-Modal Attentive Fusion (CMAF) model. Our best models outperform the DCASE 2023 Task3 audio-only and AV baselines by a wide margin on the development set of the STARSS23 dataset, making them competitive amongst state-of-the-art results of the AV challenge, without model ensembling, heavy data augmentation, or prediction post-processing. Such techniques and further pre-training could be applied as next steps to improve performance.

AI for Personalised Media Experiences

Leveraging Visual Supervision for Array-Based Active Speaker Detection and Localization

Berghi D, Jackson P.J.B. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32 2024. Access papers here.

Conventional audio-visual approaches for active speaker detection (ASD) typically rely on visually pre-extracted face tracks and the corresponding single-channel audio to find the speaker in a video. Therefore, they tend to fail every time the face of the speaker is not visible. We demonstrate that a simple audio convolutional recurrent neural network (CRNN) trained with spatial input features extracted from multichannel audio can perform simultaneous horizontal active speaker detection and localization (ASDL), independently of the visual modality. To address the time and cost of generating ground truth labels to train such a system, we propose a new self-supervised training pipeline that embraces a "student-teacher" learning approach. A conventional pre-trained active speaker detector is adopted as a "teacher" network to provide the position of the speakers as pseudo-labels. The multichannel audio "student" network is trained to generate the same results. At inference, the student network can generalize and locate also the occluded speakers that the teacher network is not able to detect visually, yielding considerable improvements in recall rate. Experiments on the TragicTalkers dataset show that an audio network trained with the proposed self-supervised learning approach can exceed the performance of the typical audio-visual methods and produce results competitive with the costly conventional supervised training. We demonstrate that improvements can be achieved when minimal manual supervision is introduced in the learning pipeline. Further gains may be sought with larger training sets and integrating vision with the multichannel audio system.

3D Virtual Human Shadow (3DVHshadow)

Einabadi F, Guillemaut JY, Hilton A. In: Learning Projective Shadow Textures for Neural Rendering of Human Cast Shadows from Silhouettes, Publisher: Centre for Vision, Speech and Signal Processing (CVSSP), 2023. Access paper here.

3DVHshadow contains images of diverse synthetic humans generated to evaluate the performance of cast hard shadow algorithms for humans. Each dataset entry includes (a) a rendering of the subject from the camera view point, (b) its binary segmentation mask, and (c) its binary cast shadow mask on a planar surface -- in total 3 images. The respective rendering metadata such as point light source position, camera pose, camera calibration, etc. is also provided alongside the images. Please refer to the corresponding publication for details of the dataset generation.

Audio Visual Speaker Localization from EgoCentric Views

Zhao J, Xu Y, Qian X, et al. arXiv preprint, 2023. Access paper here.

The use of audio and visual modality for speaker localization has been well studied in the literature by exploiting their complementary characteristics. However, most previous works employ the setting of static sensors mounted at fixed positions. Unlike them, in this work, we explore the ego-centric setting, where the heterogeneous sensors are embodied and could be moving with a human to facilitate speaker localization. Compared to the static scenario, the ego-centric setting is more realistic for smart-home applications e.g., a service robot. However, this also brings new challenges such as blurred images, frequent speaker disappearance from the field of view of the wearer, and occlusions. In this paper, we study egocentric audio-visual speaker DOA estimation and deal with the challenges mentioned above. Specifically, we propose a transformer-based audio-visual fusion method to estimate the relative DOA of the speaker to the wearer, and design a training strategy to mitigate the problem of the speaker disappearing from the camera's view. We also develop a new dataset for simulating the out-of-view scenarios, by creating a scene with a camera wearer walking around while a speaker is moving at the same time. The experimental results show that our proposed method offers promising performance in this new dataset in terms of tracking accuracy. Finally, we adapt the proposed method for the multi-speaker scenario. Experiments on EasyCom show the effectiveness of the proposed model for multiple speakers in real scenarios, which achieves state-of-the-art results in the sphere active speaker detection task and the wearer activity prediction task. The simulated dataset and related code are available at this https URL.

DTF-AT: Decoupled Time-Frequency Audio Transformer for Event Classification

Alex T, Ahmed S, Mustafa A, Rana M, Jackson P.J.B. Proceedings from the 38th Annual AAAI Conference on Artificial Intelligence 2024. Access paper here.

Convolutional neural networks (CNNs) and Transformer-based networks have recently enjoyed significant attention for various audio classification and tagging tasks following their wide adoption in the computer vision domain. Despite the difference in information distribution between audio spectrograms and natural images, there has been limited exploration of effective information retrieval from spectrograms using domain-specific layers tailored for the audio domain. In this paper, we leverage the power of the Multi-Axis Vision Transformer (MaxViT) to create DTF-AT (Decoupled Time-Frequency Audio Transformer) that facilitates interactions across time, frequency, spatial, and channel dimensions. The proposed DTF-AT architecture is rigorously evaluated across diverse audio and speech classification tasks, consistently establishing new benchmarks for state-of-the-art (SOTA) performance. Notably, on the

challenging AudioSet 2M classification task, our approach demonstrates a substantial improvement of 4.4% when the model is trained from scratch and 3.2% when the model is initialised from ImageNet-1K pretrained weights. In addition, we present comprehensive ablation studies to investigate the impact and efficacy of our proposed approach. The codebase and pretrained weights are available on https://github.com/ta012/DTFAT.git

MAX-AST: Combining Convolution, Local and Global Self-attentions for Audio Event Classification

Alex T, Ahmed S, Mustafa A, Rana M, Jackson P.J.B. Proceedings from ICASSP 2024. Access papers here.

In the domain of audio transformer architectures, prior research has extensively investigated isotropic architectures that capture the global context through full self-attention and hierarchical architectures that progressively transition from local to global context utilising hierarchical structures with convolutions or window-based attention. However, the idea of imbuing each individual block with both local and global contexts, thereby creating a hybrid transformer block, remains relatively under-explored in the field.To facilitate this exploration, we introduce Multi Axis Audio Spectrogram Transformer (Max-AST), an adaptation of MaxViT to the audio domain. Our approach leverages convolution, local window-attention, and global grid-attention in all the transformer blocks. The proposed model excels in efficiency compared to prior methods and consistently outperforms state-of-the-art techniques, achieving significant gains of up to 2.6% on the AudioSet full set. Further, we performed detailed ablations to analyse the impact of each of these components on audio feature learning. The source code is available at https://github.com/ta012/MaxAST.git

QoE Assessment for Multi-Video Object Based Media

Lyko T, Elkhatib Y, Sparks M, Ramdhany R, Race N. Proceedings from 14th International Conference on Quality of Multimedia Experience (QoMEX), Lippstadt, Germany 2022. Access paper here.

Recent multimedia experiences using techniques such as DASH allow the streaming delivery to be adapted to suit network context. Object Based Media (OBM) provides even more flexibility as distinct media objects are streamed and combined based on user preferences, allowing the experience to be personalised for the user. As adaptation can lead to degradation, modelling and measuring Quality of Experience (QoE) are crucial to ensure a perceptibly-optimal user experience. QoE models proposed for DASH include quality-related factors from single video-object streams and hence, are unsuitable for

multi-video OBM experiences. In this paper, we propose an objective method to quantify QoE for video-based OBM experiences. Our model provides different strategies to aggregate individual object QoE contributions for different OBM experience genres. We apply our model to a case study and contrast it with the QoE levels obtained using a standard QoE model for DASH.

Differential QoE in Picture-in-Picture Gaming Videos: A Subjective Study

Lyko T, Elkhatib Y, Sparks M, Ramdhany R, Race N. Proceedings from 15th International Conference on Quality of Multimedia Experience (QoMEX), 2023. Access paper here.

Video streaming continues to be the largest service delivered on the internet. This includes gaming videos, delivered both on-demand and live, where gaming footage is usually accompanied by a video of the player overlaid on top of the gameplay - resulting in Picture-In-Picture (PiP) content. Currently, PiP content is usually combined into a single video before being delivered to the client via technologies such as HTTP Adaptive Streaming (HAS). In this study, we investigated the QoE importance of gameplay and player elements in PiP gaming videos by varying the video quality of these elements individually. We conducted a subjective study, testing nine quality permutations based on three quality levels across three pieces of content from different gaming genres, with 30 participants recruited using an ethical crowdsourcing platform. We found that gameplay was significantly more important in terms of overall QoE, while the player element made a difference in only a few cases.

Self-Adaptive Systems Challenges in Delivering Object-Based Media

Dean P, Porter B. Not yet published. ACSOS 2024 Main Track. No link at present.

“I want to be independent. I want to make informed choices.”: An Exploratory Interview Study of the Effects of Personalisation of Digital Media Services on the Fulfilment of Human Values.

M Evans, Kerlin L, Parkes J, Burlington T. Proceedings from ACM International Conference on Interactive Media Experiences (IMX'22) 2022. Access paper here.

AI for Personalised Media Experiences

From the landing page of a shopping website, to a tailored layout on a video streaming app, digital media experiences are becoming increasingly personalised, and none of us have the same experience as each other. We report on a series of in-depth interviews, with UK media users from 19 to 68 years old, exploring their awareness, feelings, expectations and concerns about the digital media being personalised 'for them', and the language that they use when talking about it. Our repeatable, extensible methodology develops insights aligned to a framework of fundamental human values.

Producing Personalised Object-Based Audio-Visual Experiences: an Ethnographic Study

Cieciura C, Jackson P.J.B., Glancy M. Proceedings from 2023 ACM International Conf on interactive media Experiences. Access paper here.

Developments in object-based media and IP-based delivery offer an opportunity to create superior audience experiences through personalisation. Towards the aim of making personalised experiences regularly available across the breadth of audio-visual media, we conducted a study to understand how personalised experiences are being created. This consisted of interviews with producers of six representative case studies, followed by a thematic analysis. We describe the workflows and report on the producers' experiences and obstacles faced. We found that the metadata models, enabling personalisation, were developed independently for each experience, restricting interoperability of personalisation affordances provided to users. Furthermore, the available tools were not effectively integrated into preferred workflows, substantially increasing role responsibilities and production time. To ameliorate these issues, we propose the development of a unifying metadata framework and novel production tools. These tools should be integrated into existing workflows; improve efficiency using AI; and enable producers to serve more diverse audiences.

MG-TNet: Multi-Granular Transformer Network for Action Detection

Sardari, F, Armin Mustafa, Philip Jackson, Adrian Hilton. Proceedings from CVPR 2023. no link at present.

The Decision Space of Adaptive Caching for Object-Based Media

Dean P, Porter B. Not yet published.

Learning Self-Shadowing for Clothed Human Bodies

Einabadi F, Guillemaut JY, Hilton A. Proceedings from Eurographics Symposium on Rendering (EGSR) 2024. Access paper here.

This paper proposes to learn self-shadowing on full-body, clothed human postures from monocular colour image input, by supervising a deep neural model. The proposed approach implicitly learns the articulated body shape in order to generate self-shadow maps without seeking to reconstruct explicitly or estimate parametric 3D body geometry. Furthermore, it is generalisable to different people without per-subject pre-training, and has fast inference timings. The proposed neural model is trained on self-shadow maps rendered from 3D scans of real people for various light directions. Inference of shadow maps for a given illumination is performed from only 2D image input. Quantitative and qualitative experiments demonstrate comparable results to the state of the art whilst being monocular and achieving a considerably faster inference time. We provide ablations of our methodology and further show how the inferred self-shadow maps can benefit monocular full-body human relighting.