Editorial
Special Section on Egocentric Perception

WEARABLE devices that are able to acquire and process images, video, audio, and other sensory information from a first-person perspective allow a wide variety of potential applications, such as virtual personal assistants to support users in their daily activities and tasks. In nearly all such applications, egocentric perception is a fundamental ingredient. While egocentric perception has recently attracted a lot of attention and some large-scale datasets such as EGTEA Gaze+, EPIC-KITCHENS, and Ego4D are now available for research, questions such as what to interpret, what to ignore, and how captured information can be turned into useful data for guidance or summaries are still unanswered.

This special section gathers recent advances in the field of Egocentric Perception, bringing together multiple communities including Computer Vision, Machine Learning, and Multimedia. We received 37 submissions from authors in 19 different countries. Of these, 13 were desk rejected without review by the guest editors because they were considered out of scope or of limited quality. The remaining 24 papers went through the standard review process with at least two reviewers evaluating each submission and up to three review rounds to implement major and minor revisions. In the end, 18 papers were found suitable for publication in this PAMI special section.

We would like to thank all of the authors who submitted their work to this special section, and all of the reviewers who contributed their expertise and time. The published papers address different aspects of egocentric perception, as briefly discussed in the following.

Six papers address issues related to action understanding. In particular, the paper "Symbiotic Attention for Egocentric Action Recognition with Object-centric Alignment" by Xiaohan Wang et al. proposes an approach for egocentric action recognition based on suppressing background distractors and enhancing action-relevant interactions. In "Multi-Dataset, Multitask Learning of Egocentric Vision Tasks," Georgios Kapidis et al. study multi-task learning for egocentric action recognition which leverages related datasets and tasks at training time. The paper "First- And Third-person Video Co-analysis By Learning Spatial-temporal Joint Attention" by Huangyue Yu et al. investigates cross-view video co-analysis based on a learned joint attention between egocentric and exocentric views of a given action. "MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation" by Shijie Li et al. presents a multi-stage architecture for temporal action segmentation based on temporal convolutions and progressive prediction refinements. In "Domain-Specific Priors and Meta Learning for Few-Shot First-Person Action Recognition," Huseyin Coskun et al. investigate a method for few-shot transfer learning for first-person action classification. The method relies on independently trained visual cues such as object-object interactions, hand grasps, and motion in regions related to hand locations to learn representations which transfer well to the target domain with few examples. Finally, the paper "Learning to Recognize Actions on Objects in Egocentric Video with Attention Dictionaries" by Swathikiran Sudhakaran et al. introduces a deep neural architecture which learns to pool object and scene context descriptors from frame features to perform video action recognition.

Three of the accepted papers investigate algorithms for forecasting from egocentric vision. The paper "Multiple Trajectory Prediction of Moving Agents with Memory Augmented Networks" by Francesco Marchetti et al. presents a model based on memory augmented networks to predict multiple trajectories of other agents observed from an egocentric perspective. "Forecasting Action through Contact Representations from First Person Video" by Eadom Dessalene et al. introduces a representation centered on hand-object contact for action prediction and anticipation in egocentric vision. In "Self-Regulated Learning for Egocentric Video Activity Anticipation," Zhaobo Qi et al. propose a framework for egocentric video activity anticipation which aims to emphasize novel information in the current frame with respect to previous ones and to model correlations between previous and current observations.

Four papers introduce benchmark datasets to address fundamental problems and applications related to egocentric perception. "In the Eye of the Beholder: Gaze and Actions in First Person Video" by Yin Li et al. describes the EGTEA Gaze+ dataset and addresses the task of jointly predicting human activity and gaze from egocentric vision. "JRDB: A Dataset and Benchmark of Egocentric Robot Visual Perception of Humans in Built Environments" by Roberto Martín-Martín et al. proposes a new dataset of multimodal egocentric signals collected from JackRobbot, a social mobile manipulator. In "A Large-scale Virtual Dataset and Egocentric Localization for Disaster Responses," Hae-Gon Jeon et al. present a dataset which simulates pre- and post-disaster observations from egocentric points of view with drastic changes. The dataset is designed to study different computer vision tasks such as semantic segmentation, surface normal estimation, stereo matching, optical flow estimation, and localization. The paper "EgoCom: A Multi-person Multi-modal Egocentric Communications Dataset" by Curtis G. Northcutt et al. introduces a natural conversational dataset of multimodal human communication captured from participants'

Digital Object Identifier 10.1109/TPAMI.2023.3256679

0162-8828 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

egocentric points of view. The dataset includes 38.5 hours of audio and video with word-level transcriptions and speaker labels.

Four papers study different applications of egocentric perception ranging from 3D egocentric pose estimation for virtual reality to video summarization. In "SelfPose: 3D Egocentric Pose Estimation from a Headset Mounted Camera," Denis Tome et al. tackle the problem of 3D body pose estimation from monocular images captured from an egocentric downward looking fisheye camera installed on a virtual reality device with a novel multi-branch encoder-decoder architecture. "Depth and Video Segmentation Based Visual Attention for Embodied Question Answering" by Haonan Luo proposes a visual attention mechanism for embodied question answering based on RGB observations, depth, and segmentation maps. The proposed approach extracts local semantic features with high-speed video segmentation and guides the training process with feature fusion. The paper "Leveraging Hand-Object Interactions in Assistive Egocentric Vision" by Kyungjun Lee et al. proposes a method which uses hands as contextual information for recognizing objects of interest in egocentric RGB images for assistive applications. "Generating Personalized Summaries of Day Long Egocentric Videos" by Pravin Nagar et al. presents a framework based on reinforcement learning to summarize egocentric videos, incorporating preferences such as faces, places and scene diversity.

Finally, the survey "Analysis of the hands in egocentric vision: A survey" by Andrea Bandini and José Zariffa reviews current approaches in localizing and interpreting hands in egocentric video, reviews current applications, and provides a list of datasets with hand-based annotations.

These advances in egocentric perception are exciting steps towards a future where vision-based wearable devices will be able to understand and assist humans on a daily basis. These devices will change the way we access personal computing, making us more aware of our visual experience, allowing us to extend our abilities and achieve more, overcome disabilities, and improve our safety in workplaces.

ANTONINO FURNARI
Department of Mathematics and Computer Science
University of Catania, 95125 Catania, Italy

DAVID CRANDALL
Luddy School of Informatics, Computing, and Engineering
Indiana University, Bloomington, IN 47408 USA

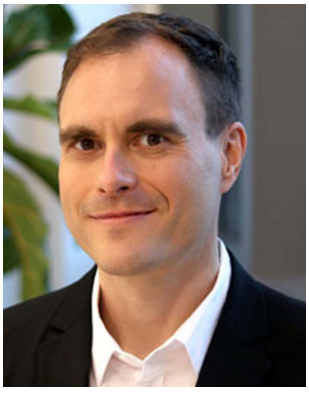
DIMA DAMEN
Department of Computer Science
University of Bristol, BS8 1UB Bristol, U.K.

KRISTEN GRAUMAN
Department of Computer Science
University of Texas at Austin,
FAIR 78712, USA

GIOVANNI MARIA FARINELLA
Department of Mathematics and Computer Science
University of Catania, 95125 Catania, Italy



Antonino Furnari is an assistant professor with the University of Catania, Italy. He works in computer vision, including egocentric vision and its applications. His research interests include focused in particular on action understanding and anticipation. He is associate editor for the Visual Computer and IET Computer Vision Journals and an area chair for ICCV 2023. He contributed to the EPIC-KITCHENS and Ego4D projects.



David Crandall is a professor of computer science and director of the Luddy Artificial Intelligence Center, Indiana University, USA. He works in computer vision, including studying the challenges, opportunities, and applications of egocentric cameras. He is an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, serves regularly as an area chair for CVPR, ICCV, and ECCV, and is a program chair of WACV 2023 and CVPR 2024. He is part of the Ego4D consortium.



Dima Damen is a professor with the University of Bristol, U.K. Her research interests include focuses on fine-grained understanding of actions and interactions. She has contributed to novel research questions including assessing skill/expertise from video sequences, domain adaptation/generalisation and multi-modal fusion using vision, audio and language. She is an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *International Journal of Computer Vision*., area chair for CVPR/ICCV/ECCV and program chair for ICCV 2021. She led the effort to collect and annotate the EPIC-KITCHENS dataset and contributed to the Ego4D project.



Kristen Grauman is a professor with the University of Texas at Austin and research director with Meta AI, USA. Her research interests include computer vision and machine learning focuses on visual recognition, video, multimodal learning, and egocentric perception. She previously served as associate editor-in-chief for *IEEE Transactions on Pattern Analysis and Machine Intelligence*, serves regularly as an area chair for CVPR/ICCV/ECCV, and was a program chair of CVPR 2015, NeurIPS 2018, and ICCV 2023. She led the Ego4D effort.



Giovanni Maria Farinella is a professor with the University of Catania, Italy. His research interests include the fields of Computer Vision and Machine Learning. He is an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *Pattern Recognition*. He founded (2006) and currently directs the International Computer Vision Summer School (ICVSS), for which he was awarded the PAMI Mark Everingham Prize in 2017. He has been program chair for ECCV 2022. He contributed to the EPIC-KITCHENS and Ego4D projects.