Artificial Vision Algorithms for Industry

Giovanni Maria Farinella[1,*], Antonino Furnari[1]

[1]FPV@IPLAB Group - Department of Mathematics and Computer Science, University of Catania

Abstract

Integrating artificial intelligence and computer vision on wearable devices in industrial environments can increase productivity, efficiency, and safety in the workplace. Despite the availability of wearable devices such as Microsoft HoloLens, Magic Leap, and nreal, the application of artificial intelligence algorithms on wearable devices equipped with cameras is an open research topic. To address this gap, the FPV@IPLAB group at the University of Catania has conducted research on the construction of machine learning and computer vision algorithms for portable devices. The research has focused on three main areas: localization and navigation, user-object interaction understanding, and user-object interaction anticipation. The work conducted by the FPV@IPLAB group aims to enhance the use wearable devices and to develop artificial intelligence techniques that can improve workplace efficiency and safety.

Keywords

Wearable and Mobile Systems, Egocentric Vision, Human Behavior Understanding, Human Behavior Anticipation

1. Introduction

Artificial intelligence can be used in industrial environments to increase efficiency and safety in the workplace. Wearable devices which acquire and analyze images and videos in the surrounding environment can be used to develop intelligent systems able to assist workers during their activities. In this context, wearable devices can allow to overlay virtual elements on the observed scene through augmented reality, and provide services based on artificial intelligence via to the analysis of images and videos acquired by the user. Moreover, due to their intrinsic mobility, wearable devices tend to be naturally exposed to large amounts of data specific to the user's visual experience, which can in principle provide an important source of knowledge for training and adapting machine learning and artificial intelligence algorithms. Even if the market offers today devices such as Microsoft HoloLens [1], Magic Leap [2], and Nreal [3], which are be suitable for use in industrial environments, the application of artificial intelligence algorithms in the context of wearable devices equipped with vision is still an underexplored field.

This article presents research conducted by the FPV@IPLAB group at the University of Catania on the construction of Machine Learning and Computer Vision algorithms for portable devices, with particular reference to the use of these technologies in industrial environments. In particular, research conducted in three areas relevant to the industry will be presented: *localization and navigation* based on images acquired by portable devices, *user-object interaction understanding*, and *user-object interaction anticipation* (see Figure 1). For further information on the research conducted by the IPLAB laboratory in the field of First Person Vision, please visit the web page http://iplab.dmi.unict.it/fpv/.

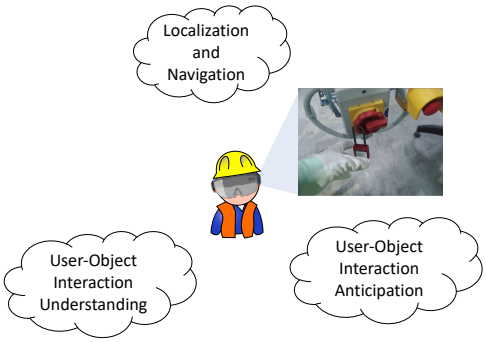

Figure 1: Main areas of research of the FPV@IPLAB group.

2. Localization and Navigation

The ability to identify the position of workers within a building, provide them with contextualized information and guide them to a destination, can be achieved through the processing of images acquired from wearable devices. While outdoor localization generally relies on GPS sys-

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

giovanni.farinella@unict.it (G. M. Farinella);
antonino.furnari@unict.it (A. Furnari)

0000-0002-6034-0432 (G. M. Farinella); 0000-0001-6911-0302 (A. Furnari)

[1]https://www.microsoft.com/en-us/hololens/

[2]https://www.magicleap.com/en-us/

[3]https://www.nreal.ai/

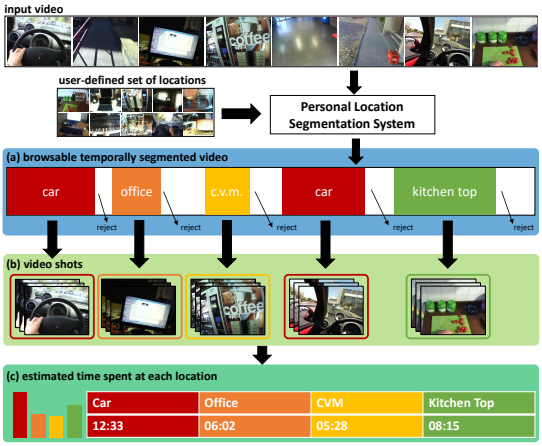

Figure 2: Diagram of the localization and temporal segmentation system for video captured by wearable devices [1].

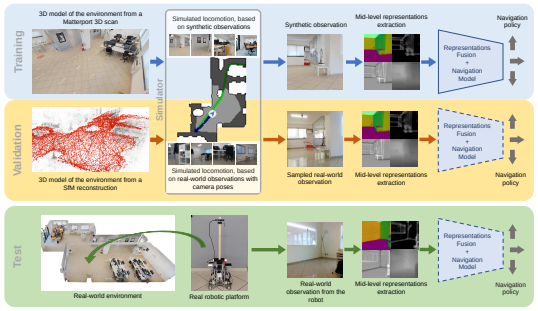

Figure 3: Approach for transferring a navigation policy to real environments [12].

tems, traditional approaches to indoor localization are based on radio-frequency technologies, such as Wi-Fi and BLE, which requires the installation of ad hoc infrastructures and cannot always guarantee a satisfactory accuracy. On the other hand, image-based localization allows to obtain more accurate results without the need for dedicated infrastructures.

2.1. Context-Based Localization

IPLAB has worked on recognizing the environment in which the user is located, called "personal location", which corresponds to specific places where the user performs certain activities, such as an office or a workbench [1, 2]. The set of relevant personal locations varies from user to user, and the developed algorithms allow to identify the environments from a set of examples provided by the user themselves, acquiring a 30-second video for each environment. The algorithm is able to recognize the personal locations of interest and discard other environments, even those not specified during system configuration. This approach allows for real-time localization and temporal segmentation of the video into coherent temporal units based on the user's context. Figure 2 illustrates the developed system, which allows to automatically index the video in order to easily navigate it, segment it into coherent clips, and estimate the time spent in each personal location. These applications can be useful in industrial contexts for work analysis and staff training. The localization system has also been used for environment recognition in a museum for visitor localization [3, 4].[4] Similar approaches have been exploited in the context of natural sites [5].

2.2. Camera pose estimation

We have also studied the problem of localization through camera position estimation. In particular, the localization of shopping carts inside a supermarket was studied using cameras mounted on the carts themselves [6, 7]. This allows the development of intelligent systems capable of guiding customers inside the store and studying their behavior to offer personalized services [8]. Localization is performed using image retrieval techniques and a metric built through deep metric learning.[5] These same technologies can be used to develop systems capable of localizing operators and guiding them inside a warehouse or other industrial environment [9]. In addition, camera pose estimation from wearable devices has also been studied considering simulated data generated from a 3D model of a real building [10]. The generation of synthetic data allows obtaining large amounts of labeled data suitable for the development of localization algorithms. The use of domain adaptation techniques also allows using synthetic data to train localization models that can work on real data [11].

2.3. Navigation

Beyond localizing workers in an industrial site, wearable systems should be able to navigate them towards a destination. A navigation system may also guide workers to follow secure paths, e.g., avoiding dangerous areas and suspended loads. Navigation algorithms can also be used to enable robots to move within the industrial environment and support the workers by escorting them or retrieving tools for them. The research activity of the FPV@IPLAB group in this area has focused on techniques for embodied visual navigation in virtual replicas of real environments [13], adaptation techniques for transfer to real scenarios [12] (see Figure 3), and human-aware robot navigation [14].

[4]Demonstration video: https://youtu.be/VYZ6Awqy1ko

[5]Demonstration video: https://youtu.be/BxbdgWxFhgc

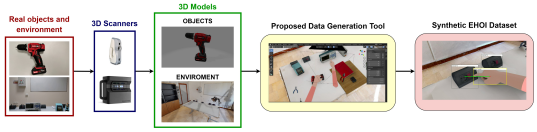

Figure 4: Protocol to generate synthetic images for human-object interaction understanding [23].

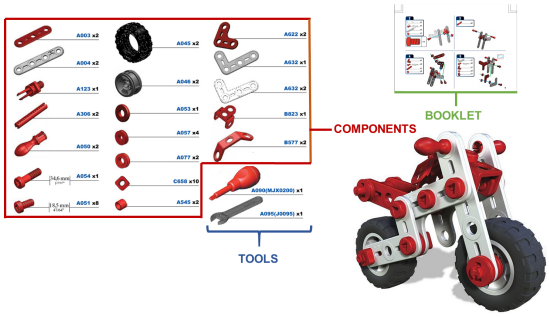

Figure 5: The MECCANO dataset [29].

3. User-Object Interaction Understanding

Recognizing the interactions between the user and the objects through wearable devices can be useful for a variety of applications in industrial scenarios, ranging from providing additional information on objects of interest through augmented reality, to monitoring user behavior and assessing the correct execution of procedures.

3.1. Object Detection and Tracking

As a first step towards user-object interaction understanding, we focused on the detection of objects seen from a first-person perspective, including synthetic-to-real domain adaptation for object detection and segmentation [15, 16, 17, 18], panoptic segmentation in industrial environments [19] and safety monitoring in construction sites [20]. Other works have considered the problem of recognizing objects in the scene and estimating which of them are currently being observed by the user [21, 22]. [6] This type of analysis allows for acquiring behavioral information on users by inferring which points of interest have been observed and for how long. Moreover, artificial intelligence algorithms can use such information to provide suggestions on the next things to see or to use.

3.2. Interaction Understanding

The FPV@IPLAB group has also focused on algorithms for understanding user-object interactions from synthetic images [23] (see Figure 4), and by leveraging the software layer provided by augmented reality devices [24]. The group is currently investigating the integration of natural language processing and object recognition to develop systems able to provide assistance to the user on the execution of specific procedures [25]. The problem of temporal segmentation of videos based on actions performed by users has also been studied in [26] and later in [27] in the form of temporal action detection. The output of these algorithms can be used as input for advanced artificial intelligence systems capable of analyzing user actions, inferring the next relevant actions, or determining any critical points in the workflow. We also investigated the impact of visual object tracking algorithms in first-person vision [28]. Visual tracking algorithms allow to keep a reference to a given object of interest in each frame of a video, which can be useful for the analysis of user-object interactions and for anticipation.

3.3. Datasets to Study User-Behavior Understanding

To facilitate the study of user-object interactions from first-person vision, the FPV@IPLAB group has contributed by collecting and labeling different datasets of egocentric videos. The MECCANO dataset [29, 30] is a multimodal dataset of egocentric videos collected in an industrial-like procedural scenario where subjects were asked to assemble a toy model of a motorbike. The dataset is provided with gaze signals, depth maps, and RGB videos acquired simultaneously with a custom headset, explicitly labeled for fundamental tasks in the context of human behavior understanding from a first-person view, such as recognizing and anticipating human-object interactions.[7] Figure 5 illustrates the parts involved in the assembly of the toy model. We collaborated in the creation of EPIC-KITCHENS [31, 32], a large dataset of first-person-view videos for action recognition, action anticipation, and object recognition. The dataset was acquired from 32 subjects in 3 different countries (Italy, UK, and Canada) and contains 55 hours of video, annotations for approximately 40,000 actions, and 500,000 objects. [8] An extension of the dataset including more videos, labels and benchmark tasks has been subsequently proposed [27]. The FPV@IPLAB group has also participated in the definition, collection, labeling and benchmarking of EGO4D [33], a large-scale egocentric video dataset that offers a vast amount of daily-life activity videos captured by 931 individuals from 74 different locations and 9 countries, accompanied by audio, 3D meshes, eye gaze, stereo

[6] Demo video: https://youtu.be/nBkYOdKYu0s

[7] Dataset: https://iplab.dmi.unict.it/MECCANO/

[8] Dataset: https://epic-kitchens.github.io/

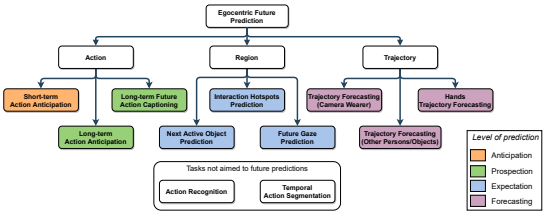

Figure 6: Taxonomy of tasks of future prediction from egocentric vision [34].

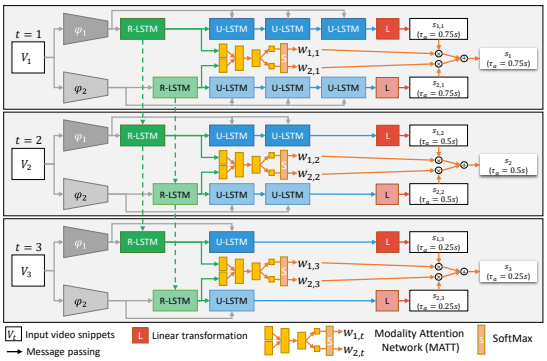

Figure 7: The Rolling-Unrolling LSTM model [38].

and synchronized videos. Together with the dataset, new benchmark challenges for understanding the first-person visual experience in the past, present and future are presented. The work has been done in collaboration with a consortium of 13 universities around the world.

4. Object Interaction Anticipation

A desirable feature for a wearable device equipped with artificial intelligence is the ability to anticipate what will happen in the scene in advance. This allows building systems that can guide the user through complex workflows and notify them if an incorrect or dangerous action is about to be taken. Our research group has investigated algorithms to predict which objects in the scene the user will interact with in the short term. We surveyed the main tasks related to the prediction of the future form egocentric videos in [34] (see Figure 6) and investigated approaches to tackle specific prediction tasks.

The FPV@IPLAB group investigated algorithms to predict which objects in the scene will be used by the user in the short term. In particular, the studies conducted in [35] have highlighted how the analysis of trajectories of objects identified from first-person videos allows obtaining information about the next objects that will be used by the user in a dynamic context. [9] We later explored the task in the context of procedural videos using object-detection based approaches in [29, 30]. The task has then been formalized as the multi-task problem of recognizing objects, and predicting future actions and time-to-contact for each of them in [33], which has also lead to the definition of the "short-term object interaction anticipation" challenge.[10]

The topic of anticipating interactions with objects has also been investigated through the definition of a challenge on egocentric action anticipation [11] related to the EPIC-KITCHENS dataset [31] and through the study of architectures and evaluation measures suitable for addressing the problem of anticipated prediction of actions from first-person videos [36]. A novel architecture to tackle the problem based on recurrent networks has been proposed in [37, 38] and extended in [39]. The developed algorithms allow predicting the set of likely next actions based on the observation of videos before they occur.[12] Applications of this approach to the domain of personal health have also been explored in [40]. The analysis of the problem has later been extended considering untrimmed input videos [41] and real-time computation constraints [42].

5. Conclusion

We presented the research conducted by the FPV@IPLAB group in the development of artificial intelligence algorithms for wearable vision devices. The problems addressed have potential applications in industrial contexts and revolve around three main themes related to *localization and navigation*, *user-object interaction understanding*, and *user-object interaction anticipation*.

References

[1] A. Furnari, S. Battiato, G. M. Farinella, Personal-location-based temporal segmentation of egocentric video for lifelogging applications, Journal of Visual Communication and Image Representation 52 (2018) 1–12.

[2] A. Furnari, G. M. Farinella, S. Battiato, Recognizing personal locations from egocentric videos, IEEE Transactions on Human-Machine Systems 47 (2017) 6–18. URL: http://iplab.dmi.unict.it/PersonalLocations/.

[3] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella, Egocentric visitors localization in cultural

[9] Video: http://iplab.dmi.unict.it/NextActiveObjectPrediction/

[10] Challenge: https://eval.ai/web/challenges/challenge-page/1623/overview

[11] Challenge: https://codalab.lisn.upsaclay.fr/competitions/707

[12] Demo video: https://youtu.be/buIEKFHTVIg

sites, Journal on Computing and Cultural Heritage (2019).

[4] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella, EGO-CH: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision, Pattern Recognition Letters (2020). URL: https://iplab.dmi.unict.it/EGO-CH/.

[5] F. L. Milotta, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella, Egocentric visitors localization in natural sites, Journal of Visual Communication and Image Representation (2019) 102664. URL: https://iplab.dmi.unict.it/EgoNature/. doi:https://doi.org/10.1016/j.jvcir.2019.102664.

[6] E. Spera, A. Furnari, S. Battiato, G. M. Farinella, Egocentric shopping cart localization, in: International Conference on Pattern Recognition, 2018. URL: http://iplab.dmi.unict.it/EgocentricShoppingCartLocalization/.

[7] E. Spera, A. Furnari, S. Battiato, G. M. Farinella, Egocart: a benchmark dataset for large-scale indoor image-based localization in retail stores, IEEE Transactions on Circuits and Systems for Video Technology 31 (2021) 1253–1267. URL: https://iplab.dmi.unict.it/EgocentricShoppingCartLocalization/.

[8] V. Santarcangelo, G. M. Farinella, A. Furnari, S. Battiato, Market basket analysis from egocentric videos, Pattern Recognition Letters 112 (2018) 83–90. URL: http://iplab.dmi.unict.it/vmba15. doi:https://doi.org/10.1016/j.patrec.2018.06.010.

[9] F. Ragusa, A. Furnari, A. Lopes, M. Moltisanti, E. Ragusa, M. Samarotto, L. Santo, N. Picone, L. Scarso, G. M. Farinella, Enigma: Egocentric navigator for industrial guidance, monitoring and anticipation, in: International Conference on Computer Vision Theory and Applications (VISAPP), 2023.

[10] S. Orlando, A. Furnari, G. M. Farinella, Egocentric visitor localization and artwork detection incultural sites using synthetic data, Pattern Recognition Letters (2020). URL: https://iplab.dmi.unict.it/SimulatedEgocentricNavigations/.

[11] D. D. Mauro, A. Furnari, G. Signorello, G. M. Farinella, Unsupervised domain adaptation for 6dof indoor localization, in: International Conference on Computer Vision Theory and Applications - VISAPP, 2021. URL: https://iplab.dmi.unict.it/EGO-CH-LOC-UDA/.

[12] M. Rosano, A. Furnari, L. Gulino, C. Santoro, G. M. Farinella, Image-based navigation in real-world environments via multiple mid-level representations: Fusion models, benchmark and efficient evaluation, CoRR abs/2202.01069 (2022). URL: https://arxiv.org/abs/2202.01069. arXiv:2202.01069.

[13] M. Rosano, A. Furnari, L. Gulino, G. M. Farinella, On embodied visual navigation in real environments through habitat, in: International Conference on Pattern Recognition (ICPR), 2020. URL: https://iplab.dmi.unict.it/EmbodiedVN/.

[14] R. Möller, A. Furnari, S. Battiato, A. Härmä, G. M. Farinella, A survey on human-aware robot navigation, Robotics and Autonomous Systems (2021).

[15] F. Ragusa, D. D. Mauro, A. Palermo, A. Furnari, G. M. Farinella, Semantic object segmentation in cultural sites using real and synthetic data, in: International Conference on Pattern Recognition (ICPR), 2020.

[16] G. Pasqualino, A. Furnari, G. Signorello, G. M. Farinella, An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites, Image and Vision Computing (2021). URL: https://iplab.dmi.unict.it/EGO-CH-OBJ-UDA/.

[17] G. Pasqualino, A. Furnari, G. M. Farinella, A multi camera unsupervised domain adaptation pipeline for object detection in cultural sites through adversarial learning and self-training, Computer Vision and Image Understanding (CVIU) (2022) 103487. URL: https://iplab.dmi.unict.it/OBJ-MDA/. doi:https://doi.org/10.1016/j.cviu.2022.103487.

[18] G. Pasqualino, A. Furnari, G. M. Farinella, Unsupervised multi-camera domain adaptation for object detection in cultural sites, in: International Conference on Image Analysis and Processing (ICIAP), 2022. URL: https://iplab.dmi.unict.it/OBJ-MDA/.

[19] C. Quattrocchi, D. D. Mauro, A. Furnari, G. M. Farinella, Panoptic segmentation in industrial environments using synthetic and real data, in: International Conference on Image Analysis and Processing (ICIAP), 2022. URL: https://iplab.dmi.unict.it/ENIGMA_SEG/.

[20] C. Quattrocchi, D. D. Mauro, A. Furnari, A. Lopes, M. Moltisanti, G. M. Farinella, Put your ppe on: A tool for synthetic data generation and related benchmark in construction site scenarios, in: International Conference on Computer Vision Theory and Applications (VISAPP), 2023.

[21] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella, Egocentric point of interest recognition in cultural sites, in: International Conf. on Computer Vision Theory and Applications, 2019.

[22] M. Mazzamuto, F. Ragusa, A. Furnari, G. M. Farinella, Weakly supervised attended object detection using gaze data as annotations, in: International Conference on Image Analysis and Processing (ICIAP), 2022.

[23] R. Leonardi, F. Ragusa, A. Furnari, G. M. Farinella, Egocentric human-object interaction detection exploiting synthetic data, in: International Conference on Image Analysis and Processing (ICIAP), 2022. URL: https://iplab.dmi.unict.it/EHOI_SYNTH/.

[24] M. Mazzamuto, F. Ragusa, A. Resta, G. M. Farinella, A. Furnari, A wearable device application for human-object interactions detection., in: International Conference on Computer Vision Theory and Applications (VISAPP), 2023.

[25] C. Bonanno, F. Ragusa, R. Leonardi, A. Furnari, G. M. Farinella, Hero: An artificial conversational assistant to support humans in industrial scenarios, in: International Conference on Signal Processing and Multimedia Applications (SIGMAP), 2022.

[26] A. Furnari, S. Battiato, G. M. Farinella, How shall we evaluate egocentric action recognition?, in: ICCV Workshops, 2017.

[27] Damen, Doughty, Farinella, Furnari, Kazakos, Ma, Moltisanti, Munro, Perrett, Price, Wray, Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100, International Journal on Computer Vision (IJCV) 130 (2022) 33–55. URL: http://epic-kitchens.github.io/2020-100.

[28] M. Dunnhofer, A. Furnari, G. M. Farinella, C. Micheloni, Visual object tracking in first person vision, International Journal of Computer Vision (IJCV) (2022). URL: https://machinelearning.uniud.it/datasets/trek150/.

[29] F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella, The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain, in: IEEE Winter Conference on Application of Computer Vision (WACV), 2021. URL: https://iplab.dmi.unict.it/MECCANO. arXiv:2010.05654.

[30] F. Ragusa, A. Furnari, G. M. Farinella, Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain, 2022. arXiv:2209.08691.

[31] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, M. Wray, Scaling egocentric vision: The epic-kitchens dataset, in: European Conference on Computer Vision, 2018.

[32] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, M. Wray, The epic-kitchens dataset: Collection, challenges and baselines, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 43 (2021) 4125–4141.

[33] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, J. Malik, Around the World in 3,000 Hours of Egocentric Video, in: IEEE/CVF International Conference on Computer Vision and Pattern Recognition, 2022.

[34] I. Rodin, A. Furnari, D. Mavroedis, G. M. Farinella, Predicting the future from first person (egocentric) vision: A survey, Computer Vision and Image Understanding 211 (2021) 103252. doi:https://doi.org/10.1016/j.cviu.2021.103252.

[35] A. Furnari, S. Battiato, K. Grauman, G. M. Farinella, Next-active-object prediction from egocentric videos, Journal of Visual Communication and Image Representation 49 (2017) 401 – 411.

[36] A. Furnari, S. Battiato, G. M. Farinella, Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation, in: ECCV Workshop on Egocentric Perception, Interaction and Computing (EPIC), 2018.

[37] A. Furnari, G. M. Farinella, What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention., in: International Conference on Computer Vision (ICCV), 2019, pp. 6252–6261.

[38] A. Furnari, G. M. Farinella, Rolling-unrolling lstms for action anticipation from first-person video, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 43 (2021) 4021–4036. doi:10.1109/TPAMI.2020.2992889.

[39] G. Camporese, P. Coscia, A. Furnari, G. M. Farinella, L. Ballan, Knowledge distillation for action anticipation via label smoothing, in: International Conference on Pattern Recognition (ICPR), 2020.

[40] I. Rodin, A. Furnari, G. M. Farinella, D. Mavroeidis, Egocentric action anticipation for personal health, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2023.

[41] I. Rodin, A. Furnari, D. Mavroedis, G. M. Farinella, Untrimmed action anticipation, in: International Conference on Image Analysis and Processing (ICIAP), 2022.

[42] A. Furnari, G. M. Farinella, Towards streaming egocentric action anticipation, in: International Conference on Pattern Recognition (ICPR), 2022.