Scene Understanding and Interaction Anticipation from First Person Vision

Ivan Rodin[1], Antonino Furnari[1],
Dimitrios Mavroeidis[2], and Giovanni Maria Farinella[1]

[1] University of Catania, Catania, Italy
ivan.rodin@unict.it, furnari@dmi.unict.it, gfarinella@unict.it
[2] Philips Research, Eindhoven, Netherlands
dimitrios.mavroeidis@philips.com

Abstract. Egocentric videos can bring a lot of information about actions performed by humans, which can be beneficial for the analysis of human activity by an external agent such as a robot. This external agent is used as a personal assistant, it should be able to anticipate the user's moves, actions and intentions. Action anticipation from first-person cameras is a challenging task due to the nature of the data processed and the non-scripted, as well as person-specific character of actions. This position paper briefly reviews the current state-of-the-art approaches for action anticipation from egocentric video and proposes ideas on how to improve the quality of predictions.

Keywords: Action Anticipation, Personal Health Assistant, Egocentric Video, Neural Networks

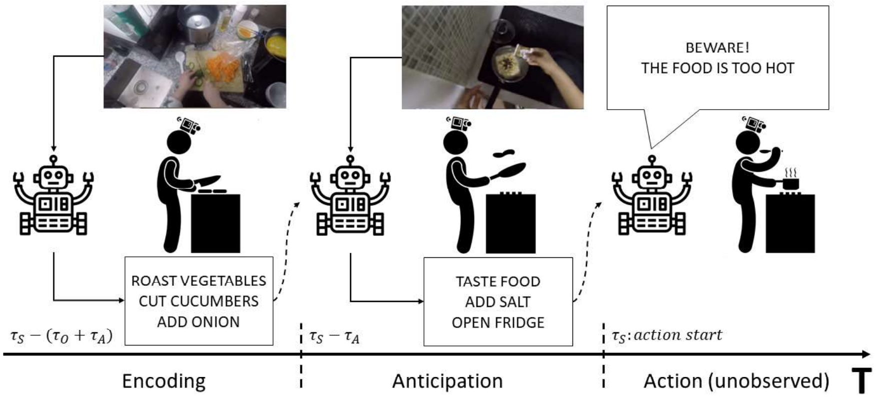

Fig. 1. Egocentric Action Anticipation with Personal Assistant.

1 Introduction

Anticipating the near future is a fundamental task for intelligent systems which need to react before an action is completed. For an external agent, such as a robot, predicting the next human's action is a crucial task in terms of smart navigation (e.g. the robot should give way to a human), assistance (e.g. the robot could bring something to a human without an explicit request) and to prevent accidents. Acquiring data from the human's point of view can boost the level of understanding of human activity and the predictive capabilities of the robot, which makes egocentric video processing convenient.

The objective of the project is to create algorithms which exploit first-person vision images/videos to support people with cognitive decline or disabilities, e.g. for memory augmentation purposes, or to make summaries of the acquired first person videos for personalised health. Moreover, the inferred information can be exploited by an external agent, e.g., a robot, to make decisions and assist the user during their daily activities.

The first aim of the project is to deeply review the state-of-the-art approaches in egocentric action anticipation, to build on them and hence improve the quality of anticipation. Then, we will focus on updating the developed anticipation model to the personal health assistant scenario. To achieve this we will consider the following sub-tasks:

- **Scene understanding.** The model should correctly retrieve the location of the user (action performer), which can be exploited as a prior for action anticipation and which should give additional information to the health-assistant.
- **Speed up the model.** The production-ready model for action anticipation should be fast enough to anticipate actions as soon as possible and transmit the prediction to external agent before the action occurs.
- **Transfer model to hardware.** The model should work in real-life environments, so the final model should be transferred to a mobile device, a robot or an operating station. Transferring the model to a device will be the final stage of the project.

The paper is structured as follows. In Section 2 we discuss the related work on action recognition and anticipation from first person view. Then, in Section 3 we define the short-term project objective (i.e., improving the quality of action anticipation) and discuss ideas on how to achieve this goal.

2 State-of-the-art

Action understanding in the context of first-person vision includes two fundamental tasks: action recognition and action anticipation.

The task of action recognition is to report what action is performed in a given video segment. This is usually done by generating a *(verb, noun)* pair describing the object the human operates on and how the object is used (e.g. cut tomato).

Action Recognition Different works have addressed the egocentric action recognition task [1–8]. Temporal Segment Networks [9] showed good performance for both third- and first-person action recognition. Other approaches leveraged on an explicit encoding of object-based features [1, 5, 7]. Learning architectures exploiting attention mechanisms showed efficiency for egocentric action recognition task [1, 3]. However, these architectures are not directly applicable to the egocentric action anticipation setting, as discussed in [10].

Action Anticipation In real-world applications, recognising the actions that already happened may be not enough. For example, in the above-mentioned scenario, the robot should predict the actions before they occur, or self-driving car should predict the next movements of the car ahead and the pedestrians in order to prevent an accident. This leads to considering the task of egocentric action anticipation, i.e., predicting the actions *before* they occur [10–16]. Recent papers on action anticipation show that combining different modalities is beneficial to improve the quality of the anticipations. The authors of [16] investigate the effect of using different egocentric modalities on the anticipation performance. The authors of [12] introduced an architecture which models and predicts the egocentric hand motion, interaction hotspots and future action. The authors of [10] proposed a model comprising 2 LSTMs: a "rolling" LSTM to summarize the past and an "unrolling" LSTM to formulate predictions about the future. They also introduced the modality-attention mechanism which learns to weight different video modalities in an adaptive fashion. In this project, we will focus on egocentric action anticipation rather than recognition.

Datasets The common datasets for egocentric action anticipation are EPIC-KITCHENS [17] and EGTEA Gaze + [2]. In our work, we will focus on EPIC-KITCHENS, which contains 39,596 action annotations, 125 verbs, and 352 nouns. During the project, we plan to collect a novel dataset containing paired first-person videos from human's and robot's points of view. For that purpose, we will use Microsoft HoloLens and the Loomo platform.

3 Improving Action Anticipation Performance

The first stage of the project will focus on the egocentric action anticipation task. To tackle the problem, we plan to build on the existing model RU-LSTM introduced in [10]. Ideas from the EPIC-KITCHENS challenge[3] will also be exploited in this project.

We will consider the video processing scheme for action anticipation proposed in [10] which consists of two main modules: video encoding and action anticipation stages. During the encoding stage, the first part of the video is processed. During the anticipation stage the model should output predictions about future actions, see Figure 1.

We plan to extend the RU-LSTM model by introducing and benchmarking different changes based on the following objectives:

[3] https://epic-kitchens.github.io/Reports/EPIC-Kitchens-Challenges-2019-Report.pdf

– **Multi-modality processing.** The model should use all available information that may be useful for action anticipation: RGB video frames, sound, motion, as well as data from the third-person vision camera (if available) [18], [10], [19]. Current approaches usually combine 2-3 modalities and they do not fully exploit all the available information within one framework.
– **Object detection.** The model should correctly detect and localize objects in egocentric videos. The model should also distinguish between objects involved (or objects that will be involved) in the interaction and those which do not play an essential role in the user action, as shown in [20]. The learning architectures should be able to consider previous objects during learning and inference. We also plan to utilise not just the labels of objects detected in video as done in [10], but also their visual appearance.
– **Hands detection and keypoints estimation.** Hands and related pose may help to identify complex hands activities, as shown in [21] and [22]. We believe, that hands keypoints data may benefit to better anticipation of events in short-term.
– **Using attention mechanisms.** We will explore the opportunity to use attention and transformer networks for decoding stage. Transformer networks showed outstanding results in various natural language processing tasks [23] and have been successfully transferred to video processing [24]. These architectures allow to perform training in self-supervised settings, which mitigates the problem of the lack of labelled data.
– **Processing various sequence lengths.** Exploring the opportunity to process long sequences of frames and utilising long-term features to improve the quality of predictions. For example, the authors of [25] show that utilizing features from past frames may improve video understanding and action recognition. We will explore how various sequence lengths affect the anticipation quality.
– **Metric.** Identifying correct metrics for evaluating action anticipation results and exploring the possibility of predicting the time to action will be analysed in the project. Papers exploring metrics and loss functions for action anticipation tasks are [26]. The authors explore various loss functions for egocentric action anticipation and introduce Verb-Noun Marginal Cross Entropy Loss. Another opportunity to improve the action anticipation quality may come from updating loss to predict time to action, as shown in [11].

4 Conclusions

The aim of the project is to develop a system which will tackle the egocentric action anticipation task in a personal health assistant setting. The model should take the video from a wearable first-person camera as input and predict future actions (verb+noun).

The short-term objective of the project is to design a model to outperform the existing state-of-the-art approaches in terms of action anticipation accuracy on EPIC-KITCHENS. Once such a model is developed, it will be adjusted to a real-time scenario to fit the industrial requirements and finally, to be transferred to an external device.

5 Acknowledgements

This research is part of the PhilHumans project[4] supported by Marie Skłodowska-Curie Innovative Training Networks - European Industrial Doctorates.

References

1. S. Sudhakaran and O. Lanz, "Attention is all we need: Nailing down object-centric attention for egocentric activity recognition," in *British Machine Vision Conference*, 2018.
2. Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 619–635.
3. S. Sudhakaran, S. Escalera, and O. Lanz, "Lsta: Long short-term attention for egocentric action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9954–9963.
4. E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5492–5501.
5. S. Singh, C. Arora, and C. Jawahar, "First person action recognition using deep learned descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2620–2628.
6. H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *2012 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2012, pp. 2847–2854.
7. M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1894–1903.
8. Y. Zhou, B. Ni, R. Hong, X. Yang, and Q. Tian, "Cascaded interactional targeting network for egocentric video analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1904–1913.
9. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision.* Springer, 2016, pp. 20–36.
10. A. Furnari and G. M. Farinella, "What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6252–6261.
11. L. Neumann, A. Zisserman, and A. Vedaldi, "Future event prediction: If and when," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.*
12. M. Liu, S. Tang, Y. Li, and J. Rehg, "Forecasting human object interaction: Joint prediction of motor attention and egocentric activity," *arXiv preprint arXiv:1911.10967*, 2019.

[4] http://www.philhumans.eu

13. Y. Abu Farha, A. Richard, and J. Gall, "When will you do what?-anticipating temporal occurrences of activities," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5343–5352.
14. Y. Zhou and T. L. Berg, "Temporal perception and prediction in ego-centric video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4498–4506.
15. M. Zhang, K. Teck Ma, J. Hwee Lim, Q. Zhao, and J. Feng, "Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4372–4381.
16. Y. Shen, B. Ni, Z. Li, and N. Zhuang, "Egocentric activity prediction via event modulated attention," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 197–212.
17. D. Damen, H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 720–736.
18. A. Cartas, J. Luque, P. Radeva, C. Segura, and M. Dimiccoli, "Seeing and hearing egocentric actions: How much can we learn?" in *Proceedings of the IEEE International Conference on Computer Vision Workshops*.
19. J. Zhao and C. G. Snoek, "Dance with flow: Two-in-one stream action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9935–9944.
20. A. Furnari, S. Battiato, K. Grauman, and G. M. Farinella, "Next-active-object prediction from egocentric videos," *Journal of Visual Communication and Image Representation*, vol. 49, pp. 401–411, 2017.
21. G. Kapidis, R. Poppe, E. van Dam, L. P. Noldus, and R. C. Veltkamp, "Egocentric hand track and object-based human action recognition," *arXiv preprint arXiv:1905.00742*, 2019.
22. B. Tekin, F. Bogo, and M. Pollefeys, "H+ o: Unified egocentric recognition of 3d hand-object poses and interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4511–4520.
23. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
24. C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7464–7473.
25. C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick, "Long-term feature banks for detailed video understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 284–293.
26. A. Furnari, S. Battiato, and G. Maria Farinella, "Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation," in *Proceedings of the European Conference on Computer Vision (ECCV)*.