2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Sim-to-Real Domain Shift in Online Action Detection

Constantin Patsch[1], Wael Torjmene[1], Marsil Zakour[1], Yuankai Wu[1], Driton Salihu[1], Eckehard Steinbach[1]

***Abstract*— Human reasoning comprises the ability to understand and reason about the current action solely based on past information. To provide effective assistance in an eldercare or household environment an assistive robot or intelligent assistive system has to assess human actions correctly. Based on this presumption, the task of online action detection determines the current action solely based on the past without access to future information. During inference, the performance of the model is largely impacted by the attributes of the underlying training dataset. However, as high costs and ethical concerns are associated with the real-world data collection process, synthetically created data provides a way to mitigate these problems while providing additional data for the training process of the underlying action detection model to improve performance.**

Due to the inherent domain shift between the synthetic and real data, we introduce a new egocentric dataset called Human Kitchen Interactions (HKI) to investigate the sim-to-real gap. Our dataset contains in total 100 synthetic and real videos in which 21 different actions are executed in a kitchen environment. The synthetic data is acquired in an egocentric virtual reality (VR) setup while capturing the virtual environment in a game engine. We evaluate state-of-the-art online action detection models on our dataset and provide insights into sim-to-real domain shift. Upon acceptance, we will release our dataset and the corresponding features at https://c-patsch.github.io/HKI/.

***Index Terms*— Visual Learning, Datasets for Human Motion, Simulation and Animation**

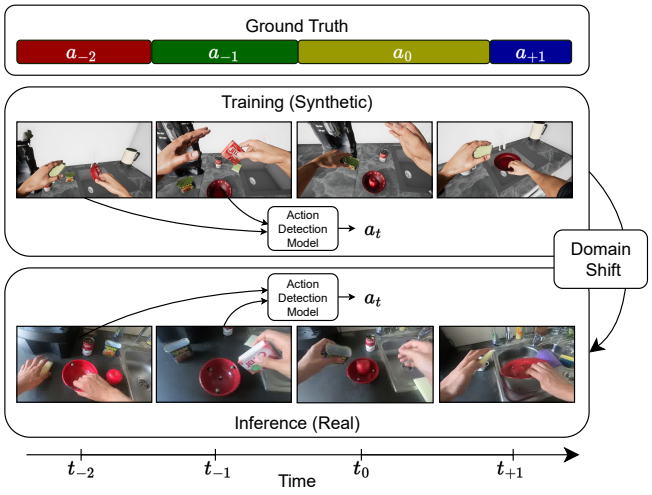


Fig. 1: Based on the synthetic video data that has been acquired from a Virtual Reality environment, an action detection model is trained in an online manner. This means that the model detects the current action solely based on past video information without knowing the future. In particular, we investigate the domain shift from synthetic to real data in case of the online action detection task.

I. Introduction

Within the field of action understanding various challenges deal with analyzing, comprehending, and interpreting human actions within different contexts. Robust action understanding has potential fields of application ranging from enhancing human-computer interaction to enabling sophisticated applications in fields such as robotics and healthcare. Compared to only classifying one action in a video in action recognition [1], [2], [3], in the case of action segmentation [4], [5], [6], and action detection [7], [8], [9] past video observations and inter-action dependencies are analyzed to infer human actions. In particular, the online action detection task, which deals with detecting the current action based on a continuous video stream, applies to real-world scenarios [10], [11]. Thus, upon detecting an action, a robotic system can potentially plan and provide tasks without receiving explicit instructions.

Regardless of the specific task, the performance of these models largely relies on the available training data. However, the collection process itself is generally costly, and depending on the real-world application, in a context like eldercare it is also privacy intrusive. Thus, several simulator-based works focus on acquiring synthetic data for the training process to mitigate these problems [12], [13], [14], [15], [16]. As a result, the training data can be enriched by additional synthetic data capturing a larger action distribution. These approaches mostly utilize commercial video games or synthetically created action animations in game engines to capture video data from a third-person perspective. Compared to previous approaches, our data acquisition is based on egocentric human-object interactions that are common in a kitchen environment. We focus on the egocentric viewpoint due to the increasing availability and usage of augmented reality smart glasses in industrial applications [17], [18] and the advantage of a consistent point of view compared to static camera positioning. By perceiving the actions from a human standpoint, a robotic system can incorporate the recognized actions into its planning while possibly also enabling the robot to learn how to interact with objects.

The simulated data is recorded with a VR headset within the Unreal Engine (UE4) game engine while utilizing the work of Martinez-Gonzalez *et al.* [19] for realistic hand-object interactions. The real-world RGB data is collected with a chest-mounted GoProHERO camera from an ego

[1]The authors are with the Department of Computer Engineering, School of Computation, Information, and Technology, Munich Institute of Robotics and Machine Intelligence (MIRMI), Chair of Media Technology, Technical University of Munich, 80333 Muenchen, Germany.

perspective. As visualized in Figure 1 we investigate how far state-of-the-art models can generalize to real-world inference when mostly being trained on the simulated data and investigate to which extent the domain shift impacts the model performance concerning the online action detection task. Thus, our contribution is three-fold:

- We provide a dataset consisting of 100 egocentric videos that are captured within a VR as well as a real-world environment that are focused on actions within a kitchen environment.
- We investigate how far state-of-the-art approaches for the online action detection task can generalize to the real-world data present during inference based on training mostly with synthetic data.
- We evaluate different state-of-the-art approaches on our dataset and provide their performance and a baseline for future work on sim-to-real approaches.

II. Related Work

A. Synthetic Data Generation

Reddy *et al.* [12] proposed the RoCoG-v2 dataset that captures robot control gestures from different views such as ground and air, and is composed of real-world and synthetic video samples captured in a third-person perspective. Similarly, Melo *et al.*[20] introduced a mixed synthetic and real-world dataset for gesture-based robotic control. Both datasets are mainly focused on the gesture/action recognition task where videos contain one individual instance of a gesture/action that is to be classified. Thus, their video samples are short and don't evolve with respect to an inherent logical structure of subsequent actions. Furthermore, there are no interactions with objects present in the datasets.

Previous work also leveraged video games as a source for simulated video data while simultaneously obtaining the ground truth information for example in driving scenarios [21], [22]. Concerning human action understanding, Roitberg *et al.* [22] proposed the Sims4Action dataset which is composed of videos that are obtained from the commercial video game *THE SIMS4*[1]. The videos are captured from a third-person perspective in various kitchen environments with different characters. The dataset consists of 10 actions, and each video contains one action. However, within the videos, the game characters often do not behave reasonably or animations appear in the gameplay that are not physically feasible or realistic. The Procedural Human Action Videos dataset (PHAV) [14] features human action videos generated in the Unity game engine from a third-person perspective, with each video containing an individual action such as 'car hit, limp, walking hug'. VirtualHome [16] is focused on providing a simulation environment in which action sequences can be composed for virtual characters, where recorded data is available from multiple static cameras in the household environment. SynADL [23] is a dataset focused on action recognition for elderly care applications, where their Eldersim platform is capable of generating realistic actions performed by various subjects within different environments. RCareWorld [13] provides a virtual environment for human-robot interactions in care homes, which is particularly focused on accurate physical representation of the human and the robot. Rahamani *et al.* [24] fit a synthetic 3D human model to a real-world action in order to render synthetic action sequences from new viewpoints. Similarly, Varol *et al.* [25] utilize 3D body reconstruction from real-world videos to obtain 3D human body motions over time, which are used to generate new synthetic data samples from various viewpoints. ManipulaTHOR [26] is a framework focused on robotic object manipulation based on egocentric visual data. They base their work on AI2-THOR [27], which provides realistic indoor scenes within which an agent can interact with various objects. However, the agent is restricted to a robot which is unsuitable for natural human action understanding.

Compared to previous work, we focus on realistic egocentric synthetic data generation with a VR setup that is visually close to the real-world environment. Furthermore compared to prior work, our dataset features realistic interactions between the human hands and the objects compared to considering coarse-grained actions or gestures without objects. Our videos also consist of subsequently dependent actions which enables the analysis of inter-action dependencies in contrast to short videos only consisting of one action. Due to the increasing interest in Human-Robot Interaction approaches based on egocentric data [28], [29], we focus on the egocentric view for our dataset. Based on egocentric action recognitions from the human perspective, a robot can plan its corresponding actions.

B. Online Action Detection

Compared to offline action detection, online action detection deals with predicting the action for the current time instant within a continuous untrimmed video stream without accessing information about future actions.

OadTR [11] is a transformer-based architecture, where the encoder determines the current action based on past frames as well as additional features from the decoder which predicts the future context. LSTR [10] is also a transformer-based approach that separately processes compressed features of the long-term past and short-term representations in order to detect the current action. Similarly, MAT [30] also separately processes the long- and short-term features while further predicting probable future actions that are feasible within the context in order to refine predictions for the current action. Instead of utilizing future predictions, the self-attention-based Colar [31] method consults category exemplars which are class-specific representations obtained from the dataset with K-Means to improve the prediction of the current action.

We focus on the online action detection task due to its real-world suitability of performing action detection on a continuous video stream. Thus, we augment the training process of the detection model with synthetic samples.

[1] www.ea.com/games/the-sims/the-sims-4

Fig. 2: Different views of our simulated kitchen environment and the corresponding views of the real-world kitchen.

Dataset	Synth+Real	Human	Sequ.	View
RoCoG-v2 [12]	✓	✓	✗	static
Sims4Action [22]	✗	✓	✓	static
PHAV [14]	✗	✓	✗	static
ManipulaTHOR [26]	✗	✗	✓	ego
SURREACT [25]	✓	✓	✗	static
SynADL [23]	✗	✓	✗	static
VirtualHome [16]	✗	✓	✓	static
HKI (ours)	✓	✓	✓	ego

TABLE I: Comparision of different synthetic datasets, where the "Synth+Real" describes whether the dataset supplies visually similar synthetic and real action videos, "Human" denotes whether the dataset is focused on human actions, "Sequ." indicates if the dataset consists of subsequent dependent actions and "View" refers to the point of view from which the data is recorded.

III. Dataset

A. Data Collection

The dataset is focused on investigating how far action understanding models are able to generalize from data obtained from a virtual environment, e.g. in the form of a digital twin, to real-world data. Thus, different from previous approaches which only provide simulated data, we also provide the corresponding real-world data that matches the synthetic data, which are reasonably close to each other in terms of the environment and the objects that are used. To retrieve visual information similar to the human point of view and to be more independent of static camera positioning we focus on egocentric data acquisition. Additionally, the egocentric view inherently generalizes well between different subjects concerning visual cues.

Within the data collection, we focus on actions in a kitchen environment, where the environment is specified concerning one specific kitchen and its respective virtual representation. The mesh of the real-world kitchen is obtained with the Polycam app on an iPad 11, where reflective surfaces are generally difficult to capture within the scan. Thus, the mesh is further supplemented with 3D objects that are available in the UE asset store. This improves the visual realism and stability of object-object interactions within the simulation, particularly with regard to the collision mesh of the kitchen appliances, which also improves the physical realism of interactions with the kitchen environment. A visual comparison between the real and the virtual kitchen is displayed in Figure 2. For the real-world data collection, we use a GoProHero7 Black mounted to the subject's chest. The synthetic data collection is based on UE4 into which we import the aforementioned mesh of the kitchen and the 3D assets. The subject uses the HTCVive VR headset and the corresponding controllers to interact with the virtual objects within the virtual kitchen environment where the physically realistic hand-object interactions are enabled by UnrealROX+ [19]. The real and virtual objects used within our dataset are based on the YCB object dataset [32] as well as object scans from ADL4D [33].

During data collection, we emphasize varying the execution of the actions with respect to hand movements and varying action-object combinations. Generally, the actions are fine-grained and visually similar, meaning that compared to actions like running and standing, changes in actions are more subtle. Furthermore, to minimize sequence-specific overfitting of the model we vary the partial ordering of the action sequences. To keep the data collection close to a real-world setting we do not enforce uniformly distributed action occurrences. We compare our dataset characteristics to the ones of previous datasets in Table I. For evaluation we provide framewise action annotations.

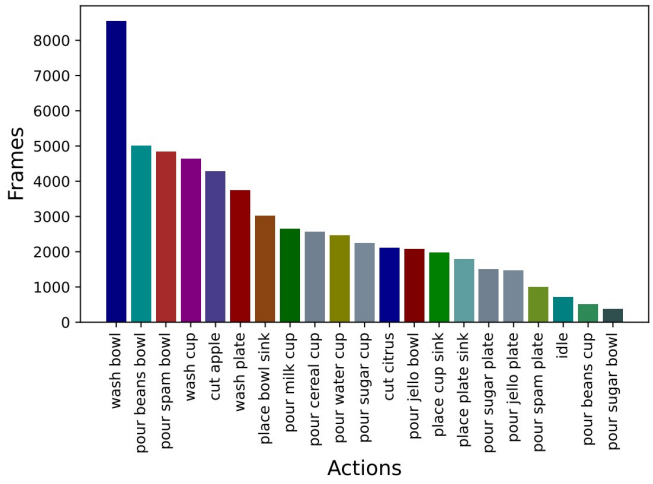


Fig. 3: Distribution over the number of frames for each action in the combined real and synthetic dataset extracted from the videos at 15 fps.

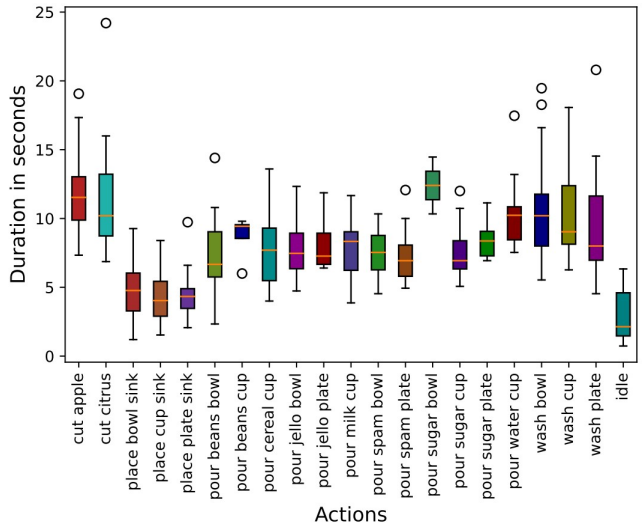


Fig. 4: Distribution over duration for each action in seconds.

B. Dataset Statistics

The overall dataset consists of 100 videos partitioned into 50 real and 50 synthetic videos. The synthetic videos are recorded at 30 fps and the real ones are recorded at 60 fps while during the feature extraction process, both are resampled to 15 fps. There are overall 21 actions which are composed of the respective verbs and nouns, where the latter correspond to the objects with which the human interacts. In total, approximately 57600 framewise features at 15 fps correspond to 491 individual action instances within all videos, where the distribution of the number of frames over individual actions is displayed in Figure 3. The distribution over the duration for each action is displayed in Figure 4. The average duration of a video is 39 seconds, where each video contains on average 5 different actions. There are 12 objects in total with which the human can interact.

IV. Evaluation

For RGB feature extraction, we use I3D and S3D models as visual backbones due to fast inference and small size suitable for practical applications, where the framewise features are extracted at 15 fps. Additionally, we calculate optical flow features in the case of I3D, where the optical flow is computed based on RAFT [34]. For training and testing, we rescale the original synthetic and real images of size 1280x720 to 256x256 and use a center crop as an input to the vision backbones. For I3D we use a stack of 16 images to compute one feature representation and for S3D a stack consists of 24 images. Both models are initialized with pre-trained Kinetics400 [35] weights. The obtained framewise features are used as input to the action detection models. We evaluate the OadTR [11], Colar [31], and MAT [30] models on the online action detection task. We chose these models due to their performance on the THUMOS14 [36] action detection dataset which ranges from 65% to 70%. We omit a separate hyperparameter tuning and adopt the parameters defined in the original approaches and keep them constant throughout the experiments. All models are trained for 30 epochs. For the following experiments, we randomly sampled 30 % of real data as a test set, which remains constant throughout all experiments, and used the remaining real data for training in the case of the *Real→Real* experiment.

Real→Real: The results in this setup indicate how far the synthetic data can be used to enhance the performance of the action detection models. During the experiment, we subsequently add synthetic data to the real training data and evaluate only concerning real data. The mean average precision (mAP) scores of Table II show that OadTR [11] performs best in the case of the I3D features and MAT [30] performs best for the S3D features. For the I3D features, we can observe that the map scores for all the considered models improve with an increasing amount of additional synthetic training data. This indicates that the synthetic data captures an additional portion of the action distribution and the variations in how they are carried out. For the S3D features, however, mainly smaller amounts of additional synthetic data improve the performance for all architectures, whereas larger amounts of additional data do not always contribute to continuously increasing performance enhancements. This indicates that particularly the synthetic optical flow data benefits the performance as the I3D features include optical flow, whereas S3D features only consider the RGB information. This generally shows that even if the amount of additional synthetic data is limited, it can still significantly impact the performance depending on its underlying action distribution.

Synth→Real: During these experiments, we start by evaluating the models on real data when only being trained on synthetic data. Additionally, we add portions of real data to the training set to investigate how far synthetic data can generalize to real data when also being supplemented with only small parts of real data. Based on the mAP scores displayed in Table III we can observe that Colar [31] generally achieves the highest performance throughout the

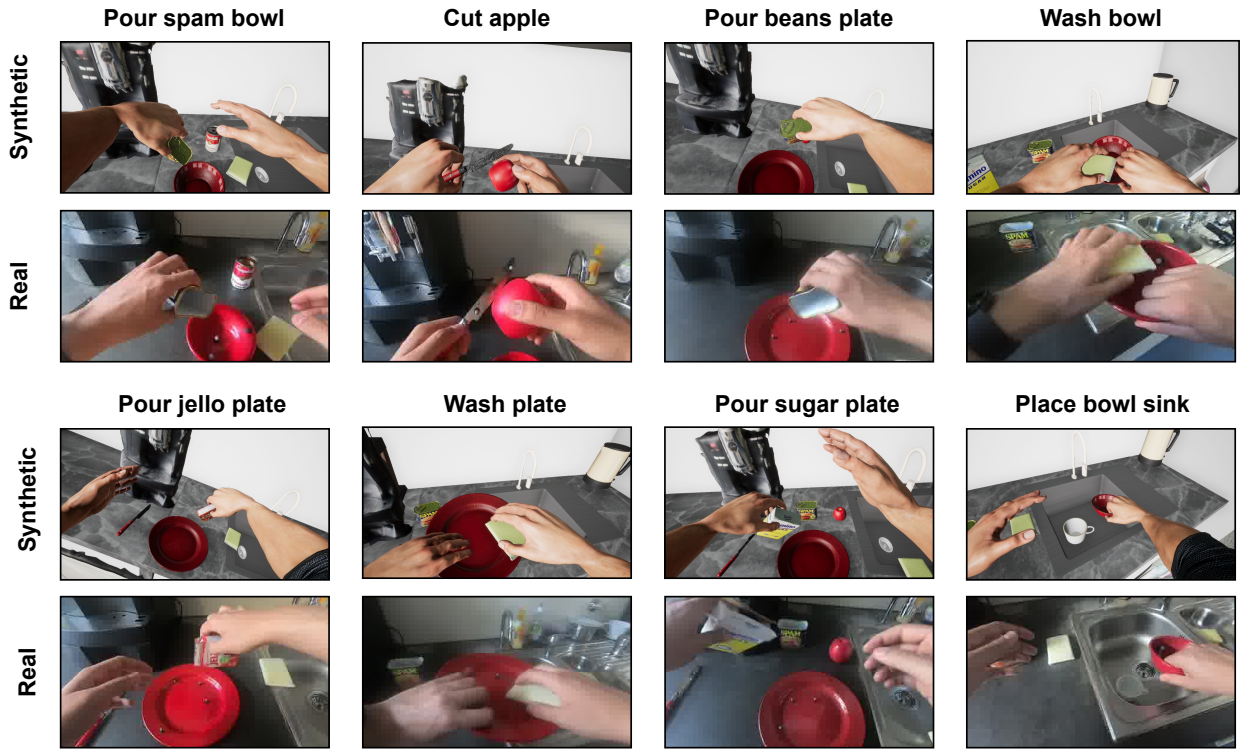


Fig. 5: Visualization of exemplary actions in the respective real and synthetic environment.

Method	Feature	Synthetic Training Data (%) 0	30	70	100
OadTR [11]	I3D	63.0	64.4	65.9	66.3
	S3D	65.3	65.5	71.1	68.5
Colar [31]	I3D	62.7	63.6	64.0	64.4
	S3D	70.4	71.0	69.0	69.3
MAT [30]	I3D	61.4	62.3	64.1	63.3
	S3D	71.8	74.9	70.7	72.7

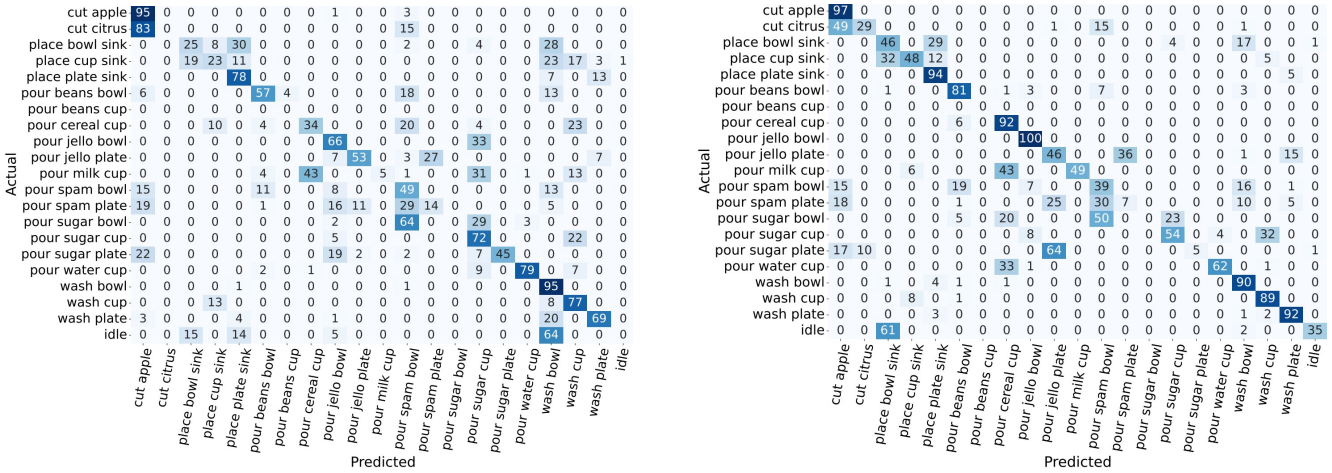
TABLE II: *Real→Real:* Evaluation of the performance of state-of-the-art models on our dataset with varying percentages of synthetic data available during training with respect to their mean average precision (mAP) score. The real data test split is kept constant throughout the experiments.

Method	Feature	Real Training Data (%) 0	10	20
MAT [30]	I3D	36.4	52.2	54.1
	S3D	35.3	51.9	54.9
OadTR [11]	I3D	39.8	51.9	55.1
	S3D	28.0	45.1	57.3
Colar [31]	I3D	40.5	53.4	56.5
	S3D	27.9	45.9	50.7

TABLE III: *Synth→Real:* Evaluation of the online action detection models trained on synthetic data with varying percentages of real data in terms of their mean average precision (mAP) score. The real data test split is kept constant throughout the experiments.

varying real data for I3D features. In the case of S3D features MAT [30] overall achieves the best performance.

With regard to the overall performance of the models, we can conclude that while training on synthetic data alone is not sufficient to achieve reliable detections, by supplementing the training process with small portions of real data, the performance can be increased significantly. This indicates that a small portion of real data in addition to training on synthetic data is crucial for bridging the sim-to-real gap for inference on real data. The model performance based on S3D features is generally lower, which indicates that the optical flow of synthetic data might particularly enhance the capabilities of the model to transfer from synthetic to real data. Thus, motion information from synthetic data in the form of optical flow seems like an important addition to RGB data for inference on real-world data. In particular, this seems to be the case for scarce real data, as the I3D performance for 0% additional real data in the first column is significantly higher than the S3D performance for all architectures. When using I3D features, we can further observe that the training process based on synthetic data with 20 percent of the overall real data performs approximately 6% lower than the case where 70 $\%$ of the real data is used to train the model which corresponds to the entries in the first column of Table II. This shows that realistic synthetic data can replace



(a) *Unseen Synth→Real* experiment in percent (%).

(b) *Real→Real* experiment in percent (%).

Fig. 6: Confusion matrices of the *Unseen Synth→Real* and *Real→Real* experiments obtained with Colar [31] based on I3D features.

a large portion of real data while achieving a competitive performance compared to solely training on a larger real database.

Unseen Synth→Real: In this experiment we change the train split that has been used in the experiment when supplementing all the synthetic with 20% real data, by explicitly removing all real video instances of the "pour sugar plate" and "pour milk cup" actions. Thus, real video samples of these actions are excluded from the training dataset and thus remain unseen. Only synthetic videos are available in the training dataset and only real videos are contained in the test dataset for these specific actions, while real videos of other actions are part of the training set. Thus, we investigate how far models can generalize from our simulated data without having real examples of that specific action available during training. We display the confusion matrices corresponding to the *Unseen Synth→Real* and *Real→Real* experiments obtained with Colar [31] based on I3D features in Figure 6. When particularly focusing on the excluded 'pour sugar plate' and 'pour milk cup' actions we can observe that while for 'pour milk cup' only 5% of the instances get correctly detected, for the 'pour sugar plate' action 45% of the instances are detected. This shows that although real video examples of actions might not be part of the training set, the model is still able to detect these action instances during inference on real-world data. Thus, we can conclude that while real data is required to achieve reliable detections, synthetic data can capture additional unseen actions that can be detected during inference on real data.

V. Conclusion

In this paper, we present HKI, a dataset for investigating the sim-to-real domain shift for human action understanding tasks. Our dataset consists of 100 synthetic and real video sequences that capture a variety of action sequences in a kitchen environment from an egocentric perspective. The evaluation is focused on the online action detection task due to its practical applicability to continuous video streams in real-world environments. We evaluate the most recent approaches based on different splits consisting of different ratios of synthetic and real data. With the *Synth→Real* experiments we showed that by adding small parts of real data to the training process based on purely synthetic data, the model can achieve competitive performances compared to the training process solely based on real data. Thus, a large part of the data acquisition process could be replaced by privacy-preserving synthetic data collection. The *Real→Real* experiments indicate that training the models on real data and supplementing the training data with synthetic data helps to improve the overall model performance throughout varying architectures. Overall, we can conclude that while real data is essential for online action detection, using synthetic data can enhance detection results and capture unseen actions during inference.

Concerning the detection performance achieved on our dataset, future work could further extend our work to real-time digital twins that accurately represent the real-world environment and other agents, which would increase scalability while ensuring privacy-preserving data acquisition.

VI. Acknowledgment

We gratefully acknowledge the funding of the Lighthouse Initiative Geriatronics by StMWi Bayern (Project X, grant no. 5140951) and LongLeif GaPa GmbH (Project Y, grant no. 5140953).

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the programme of "Souverän. Digital. Vernetzt.". Joint project 6G-life, project identification number: 16KISK002

References

[1] Wangmeng Xiang, Chao Li, Biao Wang, Xihan Wei, Xian-Sheng Hua, and Lei Zhang, "Spatiotemporal self-attention modeling with temporal patch shift for action recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 627–644.

[2] Tsung-Ming Tai, Giuseppe Fiameni, Cheng-Kuang Lee, and Oswald Lanz, "Higher-order recurrent network with space-time attention for video early action recognition," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1631–1635.

[3] Myeongjun Kim, Taehun Kim, and Daijin Kim, "Spatio-temporal slowfast self-attention network for action recognition," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2206–2210.

[4] Fangqiu Yi, Hongyu Wen, and Tingting Jiang, "Asformer: Transformer for action segmentation," in *The British Machine Vision Conference (BMVC)*, 2021.

[5] Constantin Patsch and Eckehard Steinbach, "Self-attention based action segmentation using intra-and inter-segment representations," in *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] Yazan Abu Farha and Jurgen Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3575–3584.

[7] Joao VB Soares, Avijit Shah, and Topojoy Biswas, "Temporally precise action spotting in soccer videos using dense detection anchors," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2796–2800.

[8] Yuankai Wu, Xin Su, Driton Salihu, Hao Xing, Marsil Zakour, and Constantin Patsch, "Modeling action spatiotemporal relationships using graph-based class-level attention network for long-term action detection," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 6719–6726.

[9] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai, "End-to-end temporal action detection with transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5427–5441, 2022.

[10] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto, "Long short-term transformer for online action detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1086–1099, 2021.

[11] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang, "Oadtr: Online action detection with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7565–7575.

[12] Arun V Reddy, Ketul Shah, William Paul, Rohita Mocharla, Judy Hoffman, Kapil D Katyal, Dinesh Manocha, Celso M de Melo, and Rama Chellappa, "Synthetic-to-real domain adaptation for action recognition: A dataset and baseline performances," *arXiv preprint arXiv:2303.10280*, 2023.

[13] Ruolin Ye, Wenqiang Xu, Haoyuan Fu, Rajat Kumar Jenamani, Vy Nguyen, Cewu Lu, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee, "Rcare world: A human-centric simulation world for caregiving robots," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 33–40.

[14] Cesar Roberto de Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel Lopez, "Procedural generation of videos to train deep action recognition networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4757–4767.

[15] Przemysław Korzeniowski, Szymon Płotka, Robert Brawura-Biskupski-Samaha, and Arkadiusz Sitek, "Virtual reality simulator for fetoscopic spina bifida repair surgery," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 401–406.

[16] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba, "Virtualhome: Simulating household activities via programs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8494–8502.

[17] Naghmeh Niknejad, Waidah Binti Ismail, Abbas Mardani, Huchang Liao, and Imran Ghani, "A comprehensive overview of smart wearables: The state of the art literature, recent advances, and future challenges," *Engineering Applications of Artificial Intelligence*, vol. 90, pp. 103529, 2020.

[18] Oscar Danielsson, Magnus Holm, and Anna Syberfeldt, "Augmented reality smart glasses in industrial assembly: Current status and future challenges," *Journal of Industrial Information Integration*, vol. 20, pp. 100175, 2020.

[19] Pablo Martinez-Gonzalez, Sergiu Oprea, John A. Castro-Vargas, Alberto Garcia-Garcia, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Markus Vincze, "Unrealrox+: An improved tool for acquiring synthetic data from virtual 3d environments," *CoRR*, vol. abs/2104.11776, 2021.

[20] Celso M de Melo, Brandon Rothrock, Prudhvi Gurram, Oytun Ulutan, and Bangalore S Manjunath, "Vision-based gesture recognition in human-robot teams using synthetic data," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10278–10284.

[21] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 102–118.

[22] Alina Roitberg, David Schneider, Aulia Djamal, Constantin Seibold, Simon Reiß, and Rainer Stiefelhagen, "Let's play for action: Recognizing activities of daily living by learning from life simulation video games," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8563–8569.

[23] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim, "Eldersim: A synthetic data generation platform for human action recognition in eldercare applications," *IEEE Access*, 2021.

[24] Hossein Rahmani and Ajmal Mian, "3d action recognition from novel viewpoints," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1506–1515.

[25] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman, "Synthetic humans for action recognition from unseen viewpoints," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2264–2287, 2021.

[26] Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi, "Manipulathor: A framework for visual object manipulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4497–4506.

[27] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi, "AI2-THOR: An Interactive 3D Environment for Visual AI," *arXiv*, 2017.

[28] Jianing Qiu, Lipeng Chen, Xiao Gu, Frank P-W Lo, Ya-Yen Tsai, Jiankai Sun, Jiaqi Liu, and Benny Lo, "Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8799–8806, 2022.

[29] Jianjia Xin, Lichun Wang, Kai Xu, Chao Yang, and Baocai Yin, "Learning interaction regions and motion trajectories simultaneously from egocentric demonstration videos," *IEEE Robotics and Automation Letters*, 2023.

[30] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu, "Memory-and-anticipation transformer for online action understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13824–13835.

[31] Le Yang, Junwei Han, and Dingwen Zhang, "Colar: Effective and efficient online action detection by consulting exemplars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3160–3169.

[32] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar, "Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols," *arXiv preprint arXiv:1502.03143*, 2015.

[33] Marsil Zakour, Partha Pratim Nath, Ludwig Lohmer, Emre Faik Gökçe, Martin Piccolrovazzi, Constantin Patsch, Yuankai Wu, Rahul Chaudhari, and Eckehard Steinbach, "Adl4d: Towards a contextually rich dataset for 4d activities of daily living," *arXiv preprint arXiv:2402.17758*, 2024.

[34] Zachary Teed and Jia Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[35] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[36] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," http://crcv.ucf.edu/THUMOS14/, 2014.