Robot House Human Activity Recognition Dataset

Mohammad Hossein Bamorovat Abadi, Mohammad Reza Shahabian Alashti, Patrick Holthaus,
Catherine Menon, and Farshid Amirabdollahian

***Abstract*—Human activity recognition is one of the most challenging tasks in computer vision. State-of-the art approaches such as deep learning techniques thereby often rely on large labelled datasets of human activities. However, currently available datasets are suboptimal for learning human activities in companion robotics scenarios at home, for example, missing crucial perspectives. With this as a consideration, we present the University of Hertfordshire Robot House Human Activity Recognition Dataset (RH-HAR-1). It contains RGB videos of a human engaging in daily activities, taken from four different cameras. Importantly, this dataset contains two non-standard perspectives: a ceiling-mounted fisheye camera and a mobile robot's view. In the first instance, RH-HAR-1 covers five daily activities with a total of more than 10,000 videos.**

***Index Terms*—Human Activity Recognition, Dataset.**

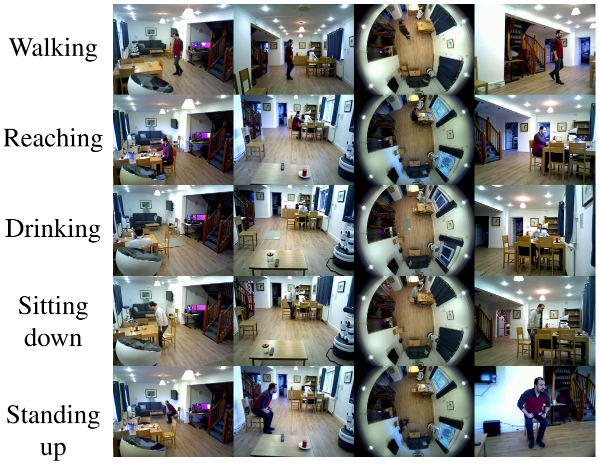
I. Introduction

In recent years, neural networks and machine learning methods have been successfully adopted for many recognition tasks in computer vision [1]. The nature of such algorithms entails that they are dependent on a high number of labelled samples depicting the relevant entity or situation. This means they are most successful when used with large datasets that are specific to the problem domain. The number of such datasets is growing rapidly [2], leading to more accurate human activity recognition models. However, most of these datasets are gathered from YouTube or outdoor environments and do not cover indoor everyday activities. As a direct consequence, these existing datasets are not ideal for human activity recognition (HAR) in the growing application domain of companion robotics and home care technologies. Therefore, we present a dataset that is suitable for human activity recognition in companion robotics scenarios. In particular, we aim to use the dataset to generate deep neural network models that are able to either use a single perspective or a fusion of multiple cameras to improve the accuracy of HAR.

II. Related Work

HAR datasets can typically be characterised based on scene properties, such as protagonist (individual or group), activity (daily activities, sports, . . .), environment (indoor or outdoor), or situation (controlled or spontaneous) and camera properties, such as data type (RGB or RGB-D), dynamics (static, moving), perspective, etc. [1], [13]. In this section, we will review the most popular RGB-based HAR video datasets and provide a brief overview of their properties in Table I.

The first publicly available datasets that contain daily activities are *KTH* [12] and *Weizmann* [11]. The low number of individual actions for each activity and the strictly controlled environment with soft background from a single perspective limit their utility in deep learning approaches. The UCF datasets, e.g. *UCF101* [8], by contrast, consist of videos that are captured from various YouTube sources without controlling the environment. *YouTube-8M* [5], of similar nature, is the largest HAR dataset so far with more than 8 Million videos in 4716 activities. The number of classes and videos in both datasets are versatile enough to be used for deep learning; however, using YouTube videos means there is no fixed view of the activities. *INRIA XMAS* [10] is the first HAR dataset that contains multiple different viewpoints, including a top-view camera in a controlled environment, while *MuHAVi* [9] is a dataset containing 8 views with 17 classes of activities. The controlled environment, lack of a dynamic perspective and the low number of videos are shortcomings of *INRIA XMAS* in our application domain. Likewise, the low number of actions (238) and controlled environment are drawbacks of *MuHAVi*. *Charades* [6] is a two-perspective dataset that includes 157 classes and 9,848 videos of daily indoor activities. *Sports-1M* [7] is one of the largest datasets, with more than one Million videos of 487 sports activities in a real-world environment that contains noisy backgrounds and a dynamic camera perspective that follows a ball or a group of people. *Moments in Time* [4] is another large recent dataset that includes more than 1 Million three-second videos labelled in 339 classes. *HACS* [3] is another new large HAR dataset with 1.5 Million videos of 200 activities. In summary, the above HAR datasets can not be adequately applied to the domain of companion robotics for the following reasons:



(a) Front (b) Back (c) Top (d) Robot

Fig. 1: Example activities of the dataset from all perspectives.

All authors are with the School of Physics, Engineering, and Computer Science, University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK. Corresponding author's email: `m.bamorovat@herts.ac.uk`

TABLE I: Overview of popular RGB-based HAR datasets and their properties.

Name	Year	Videos	Activities	Fixed Views	Environment	Situation	Dynamics	Perspective
HACS [3]	2019	1,550,000	200	-	Indoor/Outdoor	Uncontrolled	Static	Side
Moments in Time [4]	2019	1,000,000	339	-	Indoor/Outdoor	Uncontrolled	Static	Side
YouTube-8M [5]	2016	8,000,000	4,716	-	Indoor/Outdoor	Uncontrolled	Static	Side
Charades [6]	2016	9,848	157	2	Indoor/Outdoor	Controlled	Static	Side
Sports-1M [7]	2014	1,133,158	487	-	Indoor/Outdoor	Uncontrolled	Static/Moving	Side
UCF101 [8]	2012	13,320	101	-	Indoor/Outdoor	Uncontrolled	Static	Side
MuHAVi [9]	2010	238	17	8	Indoor	Controlled	Static	Side
INRIA XMAS [10]	2006	390	13	5	Indoor	Controlled	Static	Side/Top
Weizmann [11]	2005	90	10	1	Outdoor	Controlled	Static	Side
KTH [12]	2004	599	6	1	Outdoor	Controlled	Static	Side

- Daily activities: Most of the datasets are captured in mixed in-/outdoor scenarios or from random sources and are therefore do not represent repetitions of specific human daily activities. There is only cone ontrolled dataset of daily in-/outdoor activities.
- Dynamic perspective (robot view): In assistive robotics scenarios (c.f. [14]), the robot viewpoint is a crucial element. That is, the robot needs to have a good understanding of the situation and the activities a human might be engaged in while focusing on the human with its camera. With the exception of *Sports-1M*, which does not contain any daily indoor activities, there are no other datasets containing dynamic viewpoints.
- Redundancy: Companion robots may not be always engaged in direct interaction with a human but may still require information about the human's current activity to function efficiently (c.f. [14]). In these situations it might be necessary to obtain this information from an external camera. Of the above mentioned datasets, only three consider multiple perspectives.

III. RH-HAR-1 Dataset

To address the specific requirements of HAR in the assistive robotics domain and to overcome the drawbacks presented in Section II, we are currently generating the first version of the *Robot House Human Activity Recognition* dataset (*RH-HAR-1*) at the University of Hertfordshire. It consists of videos of a person in a home environment who is engaged in daily activities at different times and in various situations. Activities recorded so far are walking, drinking, sitting down, standing up and reaching for an object, cf. Figure 1. The dataset includes a dynamic perspective from a robot's point of view that is following the person plus a top-view perspective using an omnidirectional ceiling camera. In total, the activities are being recorded with four different RGB cameras from the following perspectives: I. front (static) II. back (static) III. ceiling (static, fish-eye), and IV. robot (dynamic). Each scene lasts between two and four seconds and is recorded with 30 fps. The ceiling camera is recorded at 512×486 pixels, all other cameras at 640×480. The resulting cut scenes are time-synchronised and organised by class (activity), totalling more than 10,000 short videos.

IV. Conclusion and Future Work

In this paper we have presented *RH-HAR-1*, a dataset containing video scenes of indoor daily activities. It makes use of four different synchronised perspectives including a dynamic one from a robot's viewpoint and an overview from the ceiling to address the specific challenges of activity recognition in assistive robotics scenarios. Once completed, we will make the dataset available on the University of Hertfordshire Research Archive[1]. We plan to later extend the number of activities and increase the variety of actions in each class to cover more everyday situations and increase the dataset's versatility.

References

[1] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[2] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding," in *CVPR*. IEEE, 2015, pp. 961–970.

[3] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization," in *ICCV*. IEEE, 2019, pp. 8668–8678.

[4] M. Monfort, A. Andonian, B. Zhou *et al.*, "Moments in Time Dataset: One Million Videos for Event Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 502–508, 2019.

[5] S. Abu-El-Haija, N. Kothari, J. Lee *et al.*, "Youtube-8m: A large-scale video classification benchmark," in *arXiv:1609.08675*, 2016.

[6] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *ECCV*. Springer, 2016, pp. 510–526.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," in *CVPR*. IEEE, 2014, pp. 1725–1732.

[8] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv:1212.0402*, 2012.

[9] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *AVSS*. IEEE, 2010, pp. 48–55.

[10] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[12] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *ICPR*. IEEE, 2004, pp. 32–36.

[13] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv:1806.11230*, 2018.

[14] F. Amirabdollahian, R. O. D. Akker, S. Bedaf *et al.*, "Accompany: Acceptable roboticS COMPanions for AgeiNg Years - Multidimensional Aspects of Human-System Interactions," in *HSI*. IEEE, 2013, pp. 570–577.

[1] Accessible at uhra.herts.ac.uk