RH-HAR-SK: A Multi-view Dataset with Skeleton Data for Ambient Assisted Living Research

Mohamad Reza Shahabian Alashti, Mohammad Hossein Bamorovat Abadi,
Patrick Holthaus, Catherine Menon and Farshid Amirabdollahian
Robotics Research Group, School of Engineering and Computer Science
University of Hertfordshire, Hatfield, United Kingdom
Email: {m.r.shahabian , m.bamorovat, p.holthaus, c.menon, f.amirabdollahian2}@herts.ac.uk

***Abstract*—Human and activity detection has always been a vital task in Human-Robot Interaction (HRI) scenarios, such as those involving assistive robots. In particular, skeleton-based Human Activity Recognition (HAR) offers a robust and effective detection method based on human biomechanics. Recent advancements in human pose estimation have made it possible to extract skeleton positioning data accurately and quickly using affordable cameras. In interaction with a human, robots can therefore capture detailed information from a close distance and flexible perspective. However, recognition accuracy is susceptible to robot movements, where the robot often fails to capture the entire scene. To address this we propose the adoption of external cameras to improve the accuracy of activity recognition on a mobile robot. In support of this proposal, we present the dataset *RH-HAR-SK* that combines multiple camera perspectives augmented with human skeleton extraction obtained by the *HRNet* pose estimation. We apply qualitative and quantitative analysis techniques to the extracted skeleton and its joints to demonstrate the additional value of external cameras to the robot's recognition pipeline. Results show that while the robot's camera can provide optimal recognition accuracy in some specific scenarios, an external camera increases overall performance.**

***Index Terms*—Assistive Robot, Non-generative, Multi-view dataset, Skeleton-based, Activity Recognition**

I. Introduction

Assistive robots are predominantly being developed to support older people who may have difficulty with daily living [4], [14]. To be able to offer effective assistance, such robots have to monitor people's activities, for example, to help with their medication. Skeleton-based activity recognition (SAR) algorithms present a viable option in such scenarios since they can capture fine-grained details of human motion, providing accurate and nuanced information about the actions performed by an individual [21]. Moreover, the mobility of assistive robots allows them to move the camera in order to gather a high-resolution view of the human's posture and movements from a close-up perspective.

Detection accuracy is imperative in assistive robotics, since such robots often support vulnerable people and mistakes might have a serious outcome [19], [13]. However, robot cameras often suffer from a restricted field of view and can also be influenced negatively by robot and camera movements, for example, when they are mounted on the robot's head, which might be required to be moved away from the human for communicational purposes.

Combining the robot's view with external cameras allows us to capture the scene from additional perspectives, thereby increasing the overall robustness of activity recognition. Moreover, such an approach can take advantage of its situatedness, allowing recognition results from certain camera perspectives to be weighted depending on the current interaction with the human.

With this paper, we present two main contributions to human activity detection in ambient assisted living scenarios. Firstly, we present the novel dataset RH-HAR-SK comprised of human skeleton data on top of an existing video dataset [2]. The dataset contains extracted skeletons of human activities from four different perspectives and aims to provide a rich information source to train and test the performance of human activity recognition approaches in indoor scenarios. Moreover, the dataset allows for detection algorithms to rely on low-dimensional skeleton data instead of videos and therefore reduces computing resources and networking requirements, which are otherwise computationally expensive considering the multiple parallel video streams. Secondly, we demonstrate how using additional camera perspectives enhances an assistive robot's activity recognition pipeline. For that, we measured the information contained in the different views by analysing the number of missed frames and missed poses.

Results show that certain camera views provide more valuable activity recognition data than others. For example the robot's mobility helps to follow humans and capture more details of some actions. Moreover, a wider view from environment could be a complimentary. This suggests that using additional external camera views can significantly improve reliability of activity detection to allow an assistive robot to maximise its functionality and thereby increase the users' safety, comfort, and quality of life.

To present our approach, we discuss related works that apply HAR to support assistive robots in providing their functionality and introduce methods that our recognition pipeline relies upon in Sec. II. We present the new dataset and how we augmented it with additional information to enhance its versatility within the application domain in Sec. III. We evaluate the quality of each camera view in terms of missed frames and poses in Sec. IV and discuss implications for assistive robotics in Sec. V before concluding the paper in Sec. VI.

II. Related Work

In this section, a brief review of the various technologies utilized for HAR is presented, with emphasis on the significance of the development of corresponding datasets. Subsequently, an overview of pose estimation techniques is provided, and finally, a discussion of the two distinct categories of multi-view datasets and related skeleton-based works is highlighted.

A. Human activity recognition methods

Vision-based HAR methods [6], [12], [5] rely on 2-dimensional (RGB), or 3-dimensional (RGB-D) video data acquired by a wide range of devices, e.g. stereo cameras, webcams, smartphones, etc. Video material is often sourced from video streaming platforms like YouTube or social media. *Sensor-based* recognition instead, relies on additional sensors, including global positioning systems (GPS), gyroscopes, accelerometers, or magnetometers [26], [25]. Some attempts (e.g. Bharti et al. [7]) combine both approaches and fuse recognition results from multiple sensors and cameras. Our approach allows fusing recognition results using multiple cameras without relying on external sensory technology.

Vision-based activity recognition methods can operate directly on the video input (RGB or RGB-D) or on derived data such as *skeleton* information that is generated using pose extraction methods on the raw data. Methods operating on raw camera data extract features directly from image frames in the video stream and typically perform at high accuracy [12]. By contrast, our approach relies on derived data using a pose extraction method [23] to generate skeleton-based representations of human activities in a domestic environment. Such an approach has shown to be more robust than operating on raw data (RGB) against environmental clutter and varying light circumstances and could concentrate on the activity being conducted [27].

B. Human activity recognition in assistive robotics

Human activity recognition enables robots to understand and respond to human users' needs and activities. However, few studies specifically focus on the ambient assisted living (AAL). Additionally, referring to a comprehensive review [3] of assisted living technology, there is a lack of skeleton-based and multi-view HAR datasets in this field. Therefore, developing a new dataset focusing on assistive robotics will open a new horizon in this field.

C. Pose Extraction for activity recognition

Since the pose extraction method is applied at an early-stage task in the HAR pipeline, it plays a vital role in skeleton-based HAR [22]. Low or high accuracy in this section directly affects the rest of the procedure. Thus, a reliable HAR method is dependent on a high-accuracy pose extraction method. Pose extraction typically relies on either 2-dimensional (RGB) or 3-dimensional (RGB-D) input data [10], [8]. While depth data in 3-dimensional approaches allows for better recognition results, they require special sensors that are sometimes costly or unsuitable for the environment. Moreover, the storage size of such datasets increases drastically compared to RGB-based ones. Hence, publicly available datasets often provide 2-dimensional data only. To allow for later comparison to other datasets and approaches, our work relies on 2-dimensional data. Moreover, the simplicity, affordability and accessibility of RGB cameras allow us to apply a high-performance pose extraction method independent of specific hardware on a robot.

There are two general methods in two-dimensional pose estimators, *BottomUp* [8] and *TopDown* [9], [20]. The difference between the two is the sequence of finding poses and humans. The TopDown method first finds the Region of Interest (ROI), which is the human body, and then finds the poses. The provided dataset in this work also used the TopDown method. On the other hand, in the BottomUp approach, we need to find the poses, and then by grouping them, the human skeleton data will be created.

D. Generative and non-generative datasets

When it comes to data preparation techniques, generative and non-generative view invariant HAR methods are the two primary dataset groups. As implied by the name, generative approaches produce their input data from one or more actual views [24], whereas non-generative approaches acquire their data from genuine input devices like sensors and cameras. For instance, [16] is a SOTA prospective shifting approach that transforms an action into many views and is based on the angle representation in skeletons data. Their method proved reliable when dealing with incomplete data. Moreover, Generative Adversarial Networks (GAN) [11] and encoder-decoder CNN networks are popular for RGB-based approaches [28], [15]. However, there currently exist no non-generative skeleton-based HAR dataset including a robot view, and this work address this gap. Additionally, the presented dataset can provide sufficient data to create generative datasets in the future and can be adopted for the future development of assistive scenarios.

III. RH-HAR-SK Dataset

This section provides information about the *RH-HAR-SK* dataset that we created on top of the extended version of RH-HAR-01 [2] RGB data, a multi-view human activity dataset. It includes a *single person*, *trimmed video* from *four* independent cameras, two wall-mounted cameras (front-view and back-view), one mobile robot camera (robot-view), and one ceiling fish-eye camera (omni-view). Cameras were used to cover the whole area resembling an ordinary living room, and we note that the videos from different views overlap. This dataset captures fourteen daily indoor activities *[walking, bending, sitting down, standing up, cleaning, reaching, drinking, opening can, closing can, carrying object, lifting object, putting down object, stairs climbing up, stairs climbing down]* in a typical living room of a British home. The conspicuous feature is a *mobile robot* camera synchronized with three other cameras. It enables us to explore the added value of mobile observations in HRI in the context of social and assistive robotics.

In all video clips, the frame size is 640×480. As shown in Figure 1 the bounding box size varies in different frames. The variation is based on the distance of the detected human to the camera, the camera type and position, the subject's body dimension, and the number of detected poses. The *HRNet* [23] has been used to extract poses from videos. This model has been trained over the *COCO* keypoint detection dataset [17], and the *MPII* Human Pose dataset [18].

One body skeleton with 17 poses has been extracted from each frame, and the total number of video frames varies and is not fixed in each video stream and activity. Total number of synchronized videos from each camera view in all actions is 6700. Each pose includes *X* and *Y* positions in the 2D scene. In the first step, we store the extracted poses in a *JSON* file. The JSON file was transformed to the *Tensor* file to feed the ML Training mode.

All actions from different views are combined in a single five-dimensional tensor: $T = \{n, c, f, p, s\}$, where $n \in \{\mathbb{N}_0 | n < 6700\}$ denotes the *sample number*. Note: videos are synchronized, meaning each sample across the four videos from a different camera. Some of the videos are filled with zero (0) values. These refer to a video clip with missing poses; $c \in \{\mathbb{N}_0 | c < 4\}$ identifies one of the four *camera views*; $f \in \{\mathbb{N}_0 | f < 34\}$ refers to the frame number. Because the nature of the matrix does not support different dimensions, to unify it, 34 frames randomly selected and sorted as the original sequence. $p \in \{\mathbb{N}_0 | p < 17\}$ denotes the *number of extracted poses* up to a maximum of 17 identifiable poses (c.f. Tab. I); $s \in \{\mathbb{R} | s < 3\}$ combines the relative x and y position plus the *score* of this pose are in this section. The confidence score depicts the reliability level of the extracted pose. $l \in \{\mathbb{N} | l < 14\}$ is an individual tensor L with the same dimension of sample number, which shows the class labels for the actions.

A. The Input Data Size and Sampling

One of the most challenging parts of the HAR task is the video frame sampling. Every video is labelled as a single activity, and the video length is different based on action type and situation. Then, for the ML models, this variation means having a dynamic input size. Consequently, all parameters in the model should modify based on the input size. Designing this dynamic model is a significant structural challenge in AI modelling, which is still an open area for improvement. Similarly, the skeleton-based methods need to use fix size input data. However, sampling or other reduction-based methods could lose valuable data from a video stream. In this work, *ordered random sampling* method has been used, which a fix the number of frames like 34, 64 and 128, have been selected randomly from entire frames.

Figure 2 visualizes the spatial-temporal data by a 2D image. It shows the results of transforming 20 videos stream of skeleton data from walking action in robot view to 2D images. The spatial information which is extracted from each video frame is transformed into a single row, one dimension vector with 17 elements. Each element of this row can show the relevant body pose information. They could be X, Y, or the results of a specific function like the Mean square. Figure 2 shows the X value of all 17 positions. We have depicted the information of these experiments with a grayscale image to give a better understanding.

Index	Keypoint	**Index**	Keypoint
0	Nose		
1	Left eye	2	Right eye
3	Left ear	4	Right ear
5	Left shoulder	6	Right shoulder
7	Left elbow	8	Right elbow
9	Left wrist	10	Right wrist
11	Left hip	12	Right hip
13	Left knee	14	Right knee
15	Left ankle	16	Right ankle

TABLE I: Table of Keypoints index

IV. Quantitative and Qualitative Analysis

This section focuses on the *quantity* and *quality* of the extracted skeleton and its poses from the RH-HAR-SK dataset. Two general terms are considered to describe the quality of extracted skeleton from RGB images, the number of *missed frames* and the number of *missed poses*.

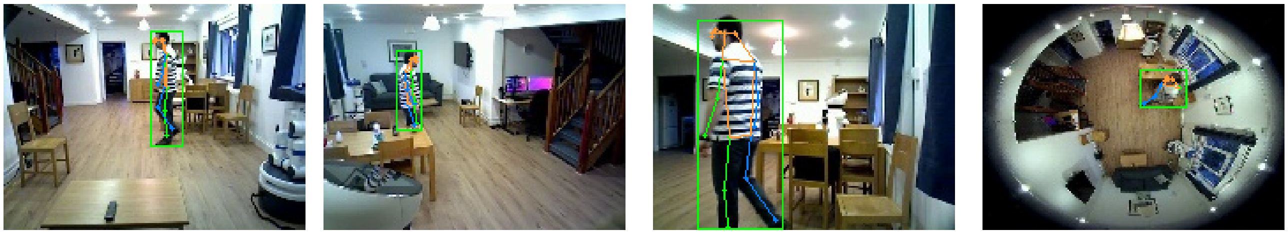
A. Missed Frames

The RGB frames on which the pose extraction methods could not find any human skeleton is considered a *missed frame*. In the RH-HAR-SK dataset, 14 actions have been captured from four synchronized camera views. The number of frames in all views is the same. Figure 4 depicts the total number of missed frames in four views and 14 actions separately. The yellow bars show the total frame number distribution in the dataset and each activity individually. The cyan bar shows the statistics of Omni-view's camera missed frames, which illustrates that the majority of actions missed the frames, higher than 45%. Meanwhile, the *walking* and *carrying object* actions by 29.6% and 36.5% have the lower missed frames in the Omni-view, respectively. At the same time, these actions have higher frames error in the robot-view with 0.9% for walking and 1.3% for carrying objects, which is negligible.

Excluding the omni-view, the highest missed frames belong to the front-view in *stairs climbing up/down* with 13.3% and 9.4%. Following that, the back-view has the same pattern in stair climbing actions by 10.2% and 4.7%.

B. Missed Poses

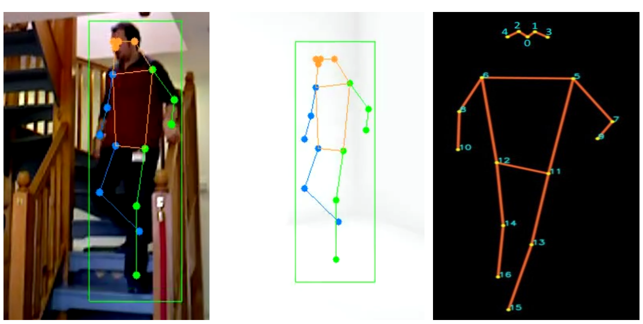
There are three parameters for each pose, *X* and *Y* values in 2D space and the *confidence* score. The confidence value refers to how much the extracted position is accurate. This value is between *0 to 1*, and we considered the values less than *0.5* as *missed poses*. Figure 5 illustrates the total number of all actions' missed poses from three views, and 17 poses separately. The total number of each pose in all activities is almost the same and hovers around 500000. The Yellow bar depicts the high-confidence poses, and the red, green, and blue show the robot, back, and front view cameras' missed poses.



(a) walking back view. **(b)** walking front view **(c)** walking robot view. **(d)** walking omni view.

Fig. 1: Synchronized skeleton output from different views of walking action.

Fig. 2: The two dimension representation of x position from 20 videos with different length in robot view from walking action



(a) **(b)** **(c)**

Fig. 3: 3a shows a subject in stair climbing down action with skeleton, 3b illustrated the skeleton only of the 3a, 3c shows the skeleton with index.

Overall, the back view has the lowest confidence in all poses, and the front-view and robot-view have the highest confidence, which changes in different joints. For the Robo-view, the highest number of missed poses belong to the lower body, with more than 50% in ankle joints and around 31% in knee joints. However, The statistics in stairs climbing up (Figure 6) and down are slightly different from all other actions. Robot camera-view shows superior results in these actions with very low missed poses. The left and right shoulders have fewer missed poses in almost all actions among all body joints.

V. Discussion

The missed frames statistics show that an omnidirectional camera is an unreliable source for body pose extraction. However, We note that it delivers good information in actions with long-range movements like walking and carrying objects. Meanwhile, there is still significant room for developing this view further, such as improving the accuracy of pose estimation by incorporating details of other views or distortion factors.

The statistics also reveal that the number of missed frames is correlated with the action type. Actions like *stair climbing up and down*, *bending*, *sitting down*, and *cleaning* that need more vertical and horizontal courses have more missed value in two fixed wall mount cameras compared to the robot view. This is because the robot head follows the human, whereas the wall-mounted cameras do not. At the same time, the robot view has moderately more missed frames due to being too close to the human or being within a cluttered environment. These manifest mainly in actions *carrying objects*, *walking*. Considering the results of both missed frames and missed poses in robot-view, we deduce that being close to the human when they are moving around quickly or for long distances can decrease recognition quality due to partially observable or not observable joints. The reason is that this view has a closer view of the human and the scene, causing missing the lower body joints.

The statistics in stairs climbing actions prove that the robot's camera movement and ability to follow the human results in fewer errors. The human has vertical movement in this action, which can be followed by a robot camera that other cameras might miss. For example, the front-view, which has the fewest missed poses on all actions on average, has the higher number of missed poses in stairs climbing up (Figure 6) and down actions.

Comparing two wall-mounted cameras with the same technical feature emphasizes the effectiveness of the viewpoint. The missed pose statistics index in Figure 5 shows that the front-view has better results regarding pose extraction quality.

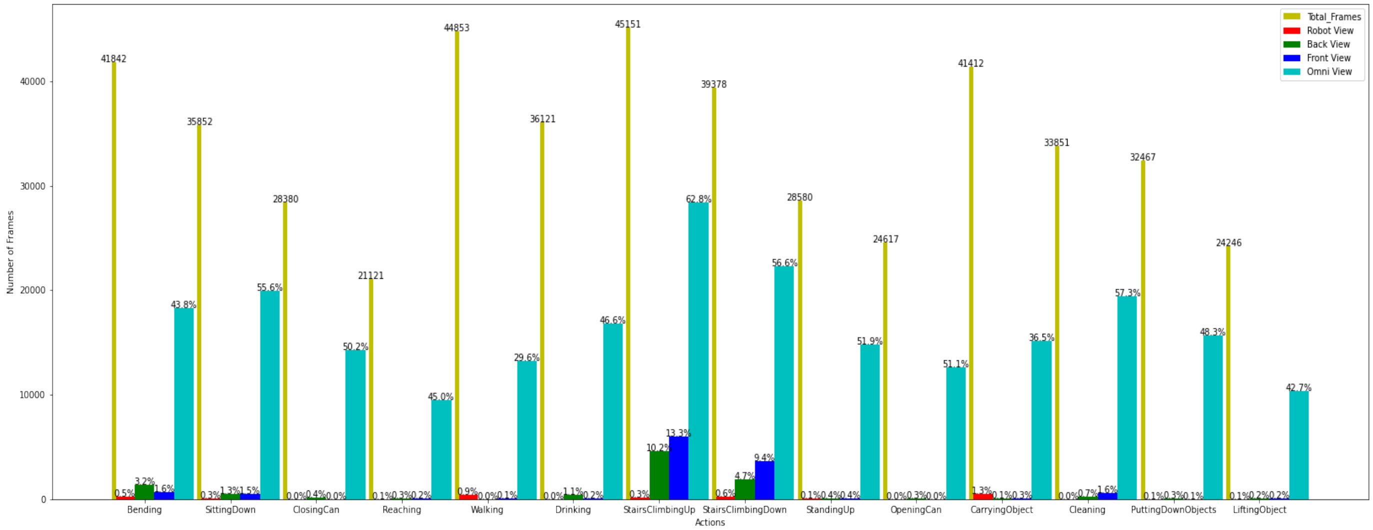


Fig. 4: Missed frame in different actions from four views

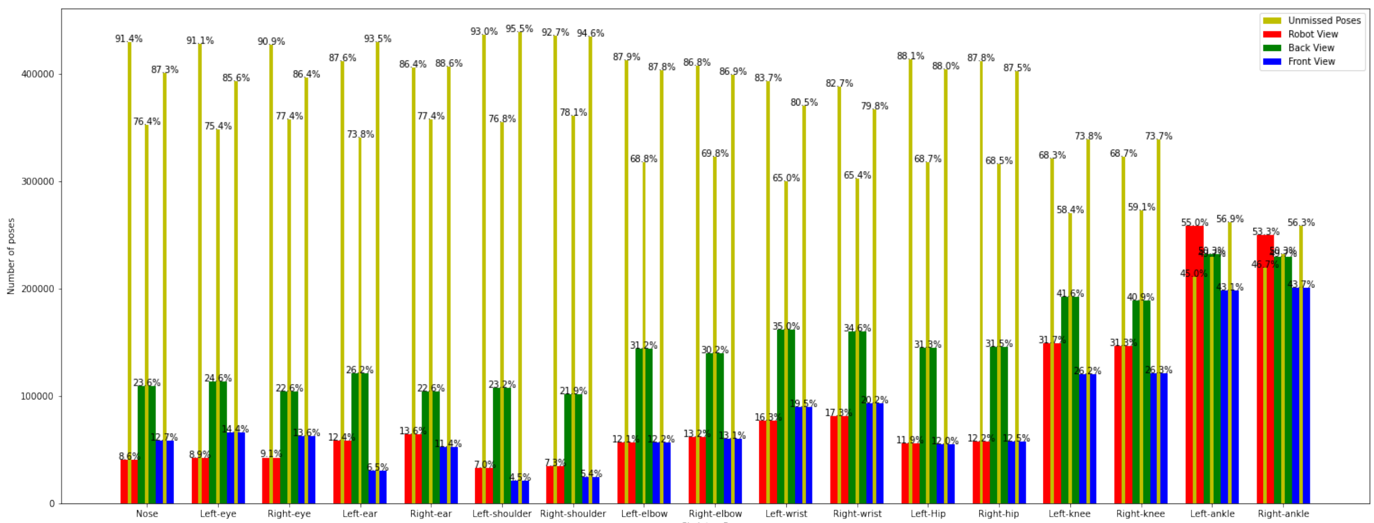


Fig. 5: The total number of all actions' missed poses in different views

On the other hand, the back view, which is also a wall-mount camera with the same technical features, results in the most missed poses in almost all actions. The only difference between these two wall-mount cameras is the altitude and view side. Reviewing the videos from these camera views in different activities suggests that the higher attitude and broader view in wall-mounted cameras can decrease the errors.

Overall, the results show that the camera position, view, activity type, and joints are highly significant in the quality of pose extraction. Theoretically, combining a robot-view camera and two other cameras can enhance skeleton extraction. Moreover, in a parallel work, we utilise this dataset to train a light-weight MV-HAR model, and the results show that adding other views has a good impact on the robot's HAR accuracy [1].

VI. Conclusion

In this paper, we have presented the novel dataset RH-HAR-SK that provides human skeleton data from multiple perspectives to facilitate human activity in ambient assisted living scenarios. We have shown how using additional cameras can enhance an assistive robot's overall activity recognition accuracy. In particular, we have shown that a broader view and higher installation height positively impact the extracted skeleton quality. In addition, we have shown that combining the robot camera with an external camera can increase accuracy when considering perspectives where humans stay within a certain area. Grafting all information into a single HRI scenario, we conclude that the proposed dataset can practically help to develop a high-level robot perception in assistive technology.

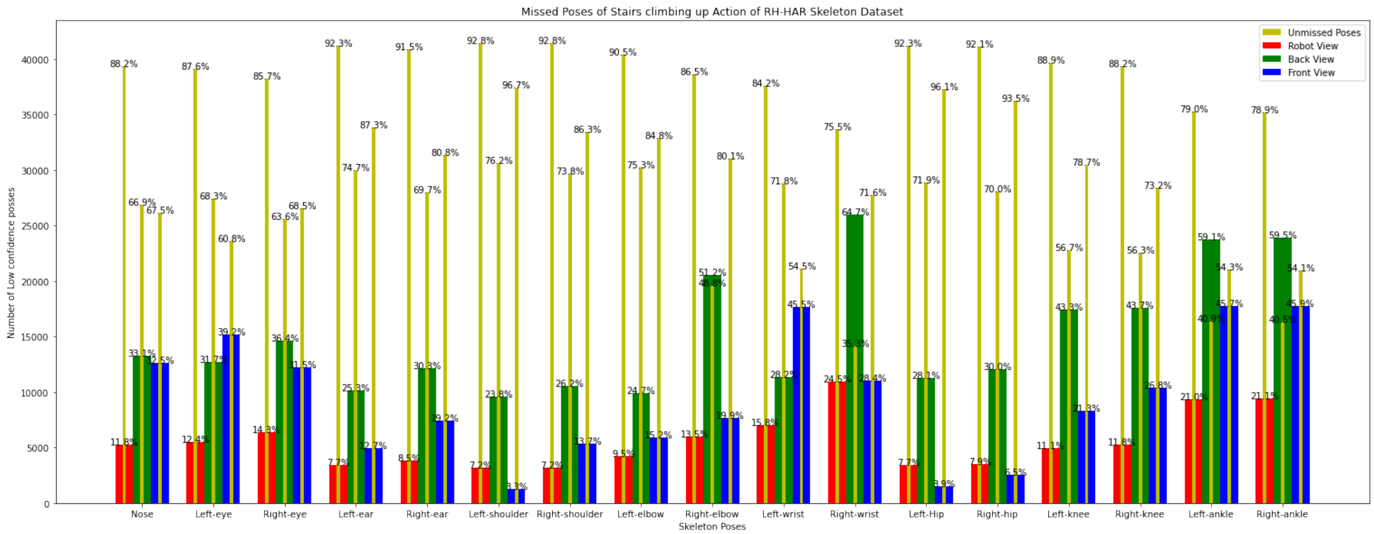


Fig. 6: The average value of Stairs climbing up action missed poses in different views

REFERENCES

[1] M.R.Shahabian .A. A lightweight multi-view skeleton-based har method for enhancing ambient assisted living scenarios. unpublished, N.D.

[2] Mohammad Hossein Bamorovat Abadi, Mohammad Reza Shahabian Alashti, Patrick Holthaus, Catherine Menon, and Farshid Amirabdollahian. Robot house human activity recognition dataset. 2021.

[3] Slavisa Aleksic, Michael Atanasov, Jean Calleja Agius, Kenneth Camilleri, Anto Cartolovni, Pau Climent-Peerez, Sara Colantonio, Stefania Cristina, Vladimir Despotovic, Hazim Kemal Ekenel, et al. State of the art of audio-and video-based solutions for aal. *arXiv preprint arXiv:2207.01487*, 2022.

[4] Farshid Amirabdollahian, Rieks op den Akker, Sandra Bedaf, Richard Bormann, Heather Draper, Vanessa Evers, Jorge Gallego Pérez, Gert Jan Gelderblom, Carolina Gutierrez Ruiz, David Hewson, et al. Assistive technology design and development for acceptable robotics companions for ageing years. *Paladyn, Journal of Behavioral Robotics*, 4(2):94–112, 2013.

[5] Muhammad Haseeb Arshad, Muhammad Bilal, and Abdullah Gani. Human activity recognition: Review, taxonomy and open challenges. *Sensors*, 22(17):6463, 2022.

[6] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79(41):30509–30555, 2020.

[7] Pratool Bharti, Debraj De, Sriram Chellappan, and Sajal K Das. Human: Complex activity recognition with multi-modal multi-positional body sensing. *IEEE Transactions on Mobile Computing*, 18(4):857–870, 2018.

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[10] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020.

[11] Jiahao Cui, Shuai Li, Qing Xia, Aimin Hao, and Hong Qin. Learning multi-view manifold for single image based modeling. *Computers & Graphics*, 82:275–285, 2019.

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[13] David Feil-Seifer, Kristine Skinner, and Maja J Matarić. Benchmarks for evaluating socially assistive robotics. *Interaction Studies*, 8(3):423–439, 2007.

[14] Moojan Ghafurian, John Muñoz, Jennifer Boger, Jesse Hoey, and Kerstin Dautenhahn. Socially interactive agents for supporting aging. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, pages 367–402. 2022.

[15] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International conference on artificial neural networks*, pages 44–51. Springer, 2011.

[16] Ruijie Hou, Yanran Li, Ningyu Zhang, Yulin Zhou, Xiaosong Yang, and Zhao Wang. Shifting perspective to see difference: A novel multi-view method for skeleton based action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4987–4995, 2022.

[17] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[18] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5137–5146, 2018.

[19] Maja J Matarić and Brian Scassellati. Socially assistive robotics. *Springer handbook of robotics*, pages 1973–1994, 2016.

[20] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4903–4911, 2017.

[21] Rainer Planinc, Alexandros Chaaraoui, Martin Kampel, and Francisco Florez-Revuelta. Computer vision for active and assisted living. 2016.

[22] Liangchen Song, Gang Yu, Junsong Yuan, and Zicheng Liu. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76:103055, 2021.

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[24] Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. Generative multi-view human action recognition. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 6212–6221, 2019.

[25] Kun Xia, Jianguang Huang, and Hanyu Wang. Lstm-cnn architecture for human activity recognition. *IEEE Access*, 8:56855–56866, 2020.

[26] Huan Yan, Yong Zhang, Yujie Wang, and Kangle Xu. Wiact: A passive wifi-based human activity recognition system. *IEEE Sensors Journal*, 20(1):296–305, 2019.

[27] Wu Zheng, Lin Li, Zhaoxiang Zhang, Yan Huang, and Liang Wang. Relational network for skeleton-based action recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 826–831. IEEE, 2019.

[28] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016.