RHM: Robot House Multi-view Human Activity Recognition Dataset

Mohammad Hossein Bamorovat Abadi[1], Mohamad Reza Shahabian Alashti[1], Patrick Holthaus[1], Catherine Menon[1], and Farshid Amirabdollahian[1]

***Abstract*—With the recent increased development of deep neural networks and dataset capabilities, the Human Action Recognition (HAR) domain is growing rapidly in terms of both the available datasets and deep models. Despite this, there are some lacks at datasets specifically covering the Robotics field and Human-Robot interaction. We prepare and introduce a new multi-view dataset to address this. The Robot House Multi-View dataset (RHM) contains four views: Front, Back, Ceiling, and Robot Views. There are 14 classes with 6701 video clips for each view, making a total of 26804 video clips for the four views. The lengths of the video clips are between 1 to 5 seconds. The videos with the same number and the same classes are synchronized in different views. In the second part of this paper, we consider how single streams afford activity recognition using established state-of-the-art models. We then assess the affordance for each of the views based on information theoretic modelling and mutual information concept. Furthermore, we benchmark the performance of different views, thus establishing the strengths and weaknesses of each view relevant to their information content and performance of the benchmark. Our results lead us to conclude that multi-view and multi-stream activity recognition has the added potential to improve activity recognition results. The RHM dataset is available at: (DOI will be provided prior to publication).**

***Index Terms*—Human Action Recognition, Human-Robot Interaction**

I. Introduction

With the growing prevalence of robots and autonomous systems in our daily lives, the domain of Human-Robot Interaction (HRI) is rapidly developing. An essential aspect of HRI is recognizing human behaviour and actions [1]. This has resulted in the emergence of the Human Action Recognition (HAR) domain, as well as the creation of numerous HAR datasets relevant to different environments. However, finding a suitable HAR dataset which contains a Robot Viewpoint has historically been challenging [1].

Human activity recognition using top-view cameras or using a dynamic Robot View has been less accurate compared to the front or Back View observation of the activity. We inquire whether the combination of viewpoints can improve detection accuracy for top-view and Robot View. To explore this, we have created a new multi-view HAR dataset, Robot House (RHM) that contains a Robot Viewpoint, a top view fish-eye camera labelled here as the Omni View, as well as two wall-mounted camera views positioned to observe front and back. These views capture the same task consisting of 14 different activities of daily living performed in front of the cameras.

[1]*School of Engineering and Computer Science,University of Hertfordshire*

In order to compare and contrast results between different viewpoints and their combination, we developed a comparison framework that allowed us to record changes in neural network models used and their resulting recognition accuracy and other performance parameters. Our methodology involved performing comparative analysis using different machine learning models, as well as Information Theoretic Analysis to identify and further characterise the relationship between multiple camera viewpoints.

In II we present a comprehensive overview of the existing HAR datasets. In section III we introduce the RHM dataset and analyse it based on Mutual Information in [2], and benchmark models in IV-B, and conclude in section V.

II. Related works

In this chapter, we review the most known RGB/D HAR datasets with a comprehensive comparison between them.

Investigating existing HAR datasets reveals that these datasets are categorised according to multiple features, including the activity's theme, camera properties, environment, subject, situation, or the activity's scenario. For example, an activity theme could be a daily, sport, industrial or surveillance activity performed by an individual or a group in an indoor or outdoor environment, controlled or uncontrolled, or in the wild. Additionally, the camera types could be RGB or RGB-D with static or dynamic positions and single or synchronized multiple views.

KTH [3] is the first RGB HAR dataset presented in 2004 with six activity classes and 599 videos.*Weizmann* [4], in 2005, has 10 classes containing 90 videos. These two datasets were prepared in an outdoor controlled environment with a static background. Daniel Weinland et. al in [5] published the first Multi-View RGB HAR dataset with five views. *INRIA XMAS* contains 390 videos with 13 activities in a controlled indoor environment. *MuHAVi* dataset [6] is published by Sanchit Singh et. al with 238 videos in 17 classes and 8 third Person fixed views (TPV) in an indoor controlled environment. H. Kuehne et. al in 2011 published a dataset with 51 classes and 6849 videos [7] termed as *MHDB51* which is a collection of static images collected mainly from movies, YouTube and Google Videos.

UCF HAR datasets are a group of datasets with different varieties of the number of classes, action types, modalities, and even views. For example, *UCF11* [8] with 11 classes and 1,160 videos, and *UCF50* [9] with 50 classes and 6,676 videos are the early versions of *UCF101* [10] with 101 classes and 13,000 videos which is one of the most famous datasets for HAR. All

of them are RGB videos prepared from YouTube clips in a diverse environment and uncontrolled situation with static and dynamic scenes. *UCF Sport* is created from sports actions in 10 classes with 150 videos [11]. *UCF-ARG* is a Multi-View dataset with 10 actions and 480 videos in each view [12]. The views are Aerial camera, Rooftop camera, and Ground camera. These views are fixed and the actions are recorded in an outdoor controlled environment. *ACT4* is a Multi-View dataset with 4 views, 14 actions, and 6,844 videos [13]. ACT4 is recorded in a controlled indoor environment. *ASLAN* is another HAR dataset with 432 classes and 10,000 videos [14]. It is trimmed from YouTube videos in an uncontrolled and diverse environment.

More recently, due to the use of neural networks and deep learning models, larger volumes of data has been needed. This has resulted in the production of some comparatively large HAR datasets, such as *Sport-1M*, the first large dataset with more than 1,000,000 videos and 487 action classes. It is Annotated on YouTube clips only with a focus on sports [15]. Also, *YouTube-8M* is another large dataset with more than 8,000,000 annotated clips in 4,800 classes with diverse environment videos [16]. **NTU** HAR datasets are two Multi-View RGB+D datasets which have been created in a controlled indoor environment with daily activities. The first version is *NTU RGB+D* with 1,000,000 annotated samples in 60 classes [17]. The second version is *NTU RGB+D 120* with 8,000,000 annotated videos in 120 classes [18].

Kinetics HAR dataset is another well-known and more usable dataset for action recognition. *Kinetics 400* was presented by Will Kay et. al in 2017 with 400 action classes and 300,000 annotated videos from YouTube clips [19]. *Kinetics 600* was published in 2018 with 600 classes and 496,000 annotated videos [20]. *Kinetics 700* was presented by Joao Carreira et. al in 2019 with 650,000 videos in 700 action classes [21]. *Ava Kinetics* is a localized human action which created from kinetics 700 with ava kinetics annotation protocol [22]. It consists of 230,000 annotated clips with 80 classes. *Kinetic_700_2020* is a 2020 edition of kinetics 700 with at least 700 videos in each class [23].

Some of the HAR datasets present another view of human actions which mostly are useful for human-object interaction. This view is from the human view which is termed as Ego or First Person (FP) View. *20BN-Something-Something* is presented with FP view of actions [24]. Raghav Goyal et. al have prepared 100,000 videos in 174 action classes. *20BN-Something-Something-V2* is published with the same view (FP) and same classes (174) but with 220,000 videos in [25]. *Charades-Ego* is another dataset which presented an FP or Ego view [26]. It provides a Multi-View HAR dataset with FP and third-person (TP) views. It contains 8000 videos and 68,500 annotated frames in 157 classes.

LEMMA is another Multi-View dataset which contain one FP and two TP views [27]. Baoxiong Jia et. al prepared LEMMA with 1,093 videos clip and 900,000 annotated frames in 641 classes. *HOMAGE* is the next Multi-View HAR dataset consisting of FP view [28]. HOMAGE presented one FP and at least one TP view for each action. Nishant Rai et. al prepared HOMAGE with 12 different sensors such as RGB, IR, microphone, acceleration, magnet, and so on. The RGB modality contains 5,700 annotated videos in 75 classes. *EPIC-KITCHENS-100* is the next Ego view HAR dataset presented in kitchen actions [29]. EPIC-KITCHENS-100 is the second version of *EPIC-KITCHENS* with 149 classes [30]. Dima Damen et. al in EPIC-KITCHENS-100 presented an EGo view (FP) dataset with 4053 classes in 700 videos and about 90000 instances.

A few of the HAR datasets use a robot to provide the dynamic view for Human Action Recognition. The first dataset with a moving robot is *LIRIS* [31]. *LIRIS* is a Multi-View HAR dataset with one Robot View and one depth TP view. It contains 10 classes in 828 videos. Another dataset which uses robots is *InHARD* [32]. Although *InHARD* used a robot for the dataset, all three (Top, Left, and Right) views are static. Since This dataset has a robot in interaction with a human, so this dataset is good for Human-Robot interaction works.

In table I, we list 42 HAR datasets with comprehensive details of each. Assessing the existing HAR datasets as described above identifies the following omissions:

- Dynamic Perspective (Robot View): There is only *LIRIS* [31] with Robot View and motion. In the human-robot interaction domain, recognizing human actions through the Robot View is crucial, and the most obvious feature of a Robot View is the motion frames. We do note that though some of the existing datasets that can be seen in the motion part of the table I as dynamic include motions in some videos, they are not in a separate part as a motion camera dataset.
- Top View (Fish eye view): For caring scenarios, using fish eye or ceiling views are common. However, we could not find a HAR dataset with ceiling views.
- Redundancy: To find the redundancy we have to check the multi-view datasets. Most of them have a different static camera at different degrees from the sides, and some with ego view. There are only *LIRIS* [31], and *InHARD* [32] with Robot View and motion.

In general, based on these conclusions, we prepared the RHM HAR dataset to cover the HAR dataset lacks.

III. Robot House Multi-View HAR dataset

The Robot House Multi-View (RHM) is a new Multi-View RGB benchmark for Human activity recognition that includes four viewpoints and focuses on human-robot interaction in the home caring domain. This dataset fully addresses the omissions identified at the end of the section II. A frame of each class and viewpoint is shown in figure 1.

A. Camera Types and Viewpoints

RMH uses robot [1] for the Robot View camera. The second unique view is a top view using a fish-eye camera, termed OmniView. There are additionally two wall-mounted cameras

[1]Fetch Robot

Dataset Name	Year	Video	An	Act	FV	En	Si	Mot	PoV	Modality	B	MV	AT	L	So	U	T	Acc
BON [33]	2022	2.6K	2.6K	18	_	Di	UC	Dy	FP	RGB	Dy	No	No	No	C	Home	Tr	No
EPIC-KITCHENS-100 [29]	2021	700	90K	4053	_	I	UC	Di	FP	RGB	Dy	No	No	No	C	Kitchen	A	Link
HOMAGE [28]	2021	5.7K	5.7K	75	2	I	UC	Di	FP/TP	12 Sensors	Dy	Yes	Yes	No	C	Home	A	Link
HA500 [34]	2021	10K	591K	500	_	Di	UC	St	TP	RGB	Dy	No	Yes	No	W	Diversity	A	Link
M-MiT [35]	2021	1M	2M	292	_	Di	UC	St	TP	RGB	Dy	No	No	Yes	W	Diversity	A	Link
MovieNet [36]	2020	1.1K	65K	80	_	Di	UC	St	TP	RGB	Dy	No	No	No	M	Diversity	A	Link
Multi-ViewPointOutdoor [37]	2020	2.3K	503K	20	3	O	UC	Di	TP	RGB	Dy	Yes	No	No	YT	Sport	A	No
HVU [38]	2020	572K	9M	3457	_	Di	UC	St	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
AViD [39]	2020	80k	80K	887	_	Di	C	St	TP	RGB	St	No	No	No	W	Diversity	A	Link
LEMMA [27]	2020	1.1K	0.9M	641	3	I	C	Di	FP/TP	RGB,D	Dy	Yes	Yes	No	C	Home	A	Link
InHARD [32]	2020	4.8K	2M	14	3	I	C	S	TP	RGB,D	Dy	Yes	No	No	C	Industrial	A	Link
FineGym [40]	2020	503	32.5K	15	_	I	UC	Di	TP	RGB	Dy	No	Yes	No	M	Sport	A	Link
Ava_Kinetic [22]	2020	500	230K	80	_	Di	UC	St	TP	RGB	Dy	No	No	Yes	YT	Diversity	A	Link
Kinetic_700_2020 [23]	2020	648K	648K	700	_	Di	UC	St	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
Jester [41]	2019	148K	5.3M	27	_	I	C	St	TP	RGB	Dy	No	Yes	No	C	Gesture	Tr	No
HACS [42]	2019	504K	1.5M	200	_	Di	UC	St	TP	RGB	Dy	No	No	Yes	YT	Diversity	A	Link
Kinetic_700 [21]	2019	650K	650K	700	_	Di	UC	St	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
NTU RGB+D 120 [18]	2019	114K	8M	120	155	I	C	St	TP	RGB,D	Dy	Yes	Yes	No	C	Daily	A	Link
MiT [43]	2019	1M	1M	339	_	Di	UC	Di	TP	RGB	Dy	No	No	No	W	Diversity	Tr	Link
20BN-sth_sth-V2 [25]	2018	220K	220K	174	_	I	UC	Di	FP	RGB	Dy	No	No	No	W	Diversity	A	No
Kinetic_600 [20]	2018	496K	496	600	_	Di	UC	Di	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
Charades-Ego [26]	2018	8K	68.5K	157	2	I	C	Di	FP/TP	RGB	Dy	Yes	Yes	Yes	C	Daily	A	Link
AVA [44]	2017	430	197K	80	_	Di	UC	St	TP	RGB	Dy	No	Yes	Yes	M	Diversity	A	Link
SLAC [45]	2017	520K	1.17M	200	_	Di	UC	Di	TP	RGB	Dy	No	No	Yes	YT	Diversity	A	No
MultiTHUMOS [46]	2017	38.6K	38.6K	65	_	Di	UC	Di	TP	RGB	Dy	No	No	No	W	Diversity	A	Link
20BN-Sth_Sth [24]	2017	100K	100K	174	_	I	UC	Dy	FP	RGB	Dy	No	Yes	No	W	Diversity	Tr	No
Kinetic_400 [19]	2017	300K	300K	400	_	Di	UC	St	TP	RGB	Dy	No	Yes	No	YT	Diversity	A	Link
M2I [47]	2017	1784	1784	22	2	I	C	St	TP	RGB,D	Dy	Yes	Yes	No	C	Diversity	Tr	No
DALY [48]	2016	8133	8133	10	_	Di	UC	St	TP	RGB	Dy	No	Yes	Yes	YT	Diversity	A	Link
YouTube-8M [16]	2016	8.2M	8.2M	4800	_	Di	UC	Di	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
NTU RGB+D [17]	2016	56K	56K	60	3	I	C	St	TP	RGB,D	Dy	Yes	Yes	No	C	Daily	Tr	Link
Charades [49]	2016	10K	10K	157	2	I	UC	St	TP	RGB	Dy	No	No	Yes	YT	Daily	Tr	Link
UTD-MHAD [50]	2015	861	861	27	5	I	C	St	TP	RGB,D	St	Yes	Yes	No	C	Daily	Tr	Link
ActivityNet [51]	2015	23K	23K	203	_	Di	UC	St	TP	RGB	Dy	No	No	No	W	Diversity	A	Link
Sport-1M [15]	2014	1M	1M	487	_	Di	UC	Di	TP	RGB	Dy	No	No	No	YT	Sport	A	Link
Berkeley MHAD [52]	2013	660	660	11	12	I	C	St	TP	RGB,D	St	Yes	Yes	No	C	Diversity	Tr	Link
Multi-View 3D Events [53]	2013	3.8K	383K	11	3	I	C	St	TP	RGB,D	Dy	Yes	Yes	No	C	Diversity	Tr	No
ASLAN [14]	2012	10K	10K	432	_	Di	UC	St	TP	RGB	Dy	No	No	No	YT	Diversity	Tr	Link
UCF101 [10]	2012	13K	13K	101	_	Di	UC	Di	TP	RGB	Dy	No	Yes	No	YT	Diversity	Tr	Link
LIRIS [31]	2012	828	828	10	2	I	C	Di	TP	RGB,D	Dy	Yes	Yes	Yes	C	Daily	Tr	Link
HMDB51 [7]	2011	6.8K	6.8K	51	_	Di	UC	Di	TP	RGB	Dy	No	No	No	YT	Daily	Tr	Link
UCF_ARG [12]	2010	480*3	480*3	10	3	O	C	St	TP	RGB	Dy	Yes	Yes	Yes	C	Daily	Tr	Link

TABLE I: Overview of popular HAR datasets and their properties, presented in descending order of year starting from 2022 and ending with 2010. (An: Number of Annotation, Act: Number of classes, FV: Number of Fixed Views, En: Environment Type (I: Indoor, O: Outdoor, Di: Diverse), Si: Situation (C: Controlled, UC: UnControlled), Mot: Camera motion capability (Dy: Dynamic, St: Static, Di: Diverse), PoV: Point of View (FP: First Person, TP: Third Person), B: Background (Dy: Dynamic, St: Static), Mu: Multi-View, At: Atomic, L: Localization, So: Source (C: Created, W: Web, M: Movie, YT: YouTube, U: Usage, T: data preparation type (Tr: Trimm, A: Annotation), Acc: Accessibility)

View Name	Motion	Position	Resolution	FR
FrontView	Static	Wall	640 * 480	30
BackView	Static	Wall	640 * 480	30
RobotView	Dynamic	Robot	640 * 480	30
OmniView	Static	Ceiling	512 * 486	30

TABLE II: RHM Viewpoints details (FR: Frame Rate)

providing a static and side view for all actions in order to provide a better comparison between the views. The Back View and Front View cameras are paced in front of each other. Table II contains more details of the cameras and viewpoints in RMH.

B. Subject

As a result of COVID-19, populating this dataset with different participants was not possible, and the actions were therefore performed by only one person.

C. Content

RHM activity classes are selected from Bedaf et. al work in [54] which features important daily activities for persons living independently. The work highlights that companion robots and ambient-assistive systems could provide a value proposition should they be able to detect these activities. The list of the activities are: Walking, Sitting Down, Standing Up, Lifting Objects, Carrying Objects, Drinking, Stairs Climbing Up, Stairs Climbing Down, Stretching, Putting Objects Down, Reaching, Opening Can, Closing Can, and Cleaning

D. Training-Validation-Testing

The RHM dataset has 14 classes in each view and the number of videos in each class and each view is between 407 to 700. The total number of videos in each view is 6,701 and for all 4 views is 26,804. The data is split into training (80%), testing (10%), and validation(10%) for each view. Table III Shows the number of videos for the training, testing, and validation in each view, and for all views. Clip length varies between 1 to 5 seconds.

E. Naming Protocol

RHM contains four folders and the name of each folder corresponds with the view name. Each splits into training, testing, and validation folders. The split folder comprises 14 folders which are class names. The clips are inside the class folders with the ordered numbering. The naming protocol is like below:

ClassName_ViewName_clipNumber.avi

For example, Drinking_RobotView_103.avi refers to clip 103 of the action class 'drinking' from the Robot Viewpoint.

F. Time Synchronising

All the clips with the same class name and number with various view names are synchronised with each other. For instance, Clip 320 in reaching action, is time-synchronised with the remaining views.

	Train	Validation	Test
Each View	4278	1076	1347
All Views	17112	4304	5388

TABLE III: Number of videos in each View/Split

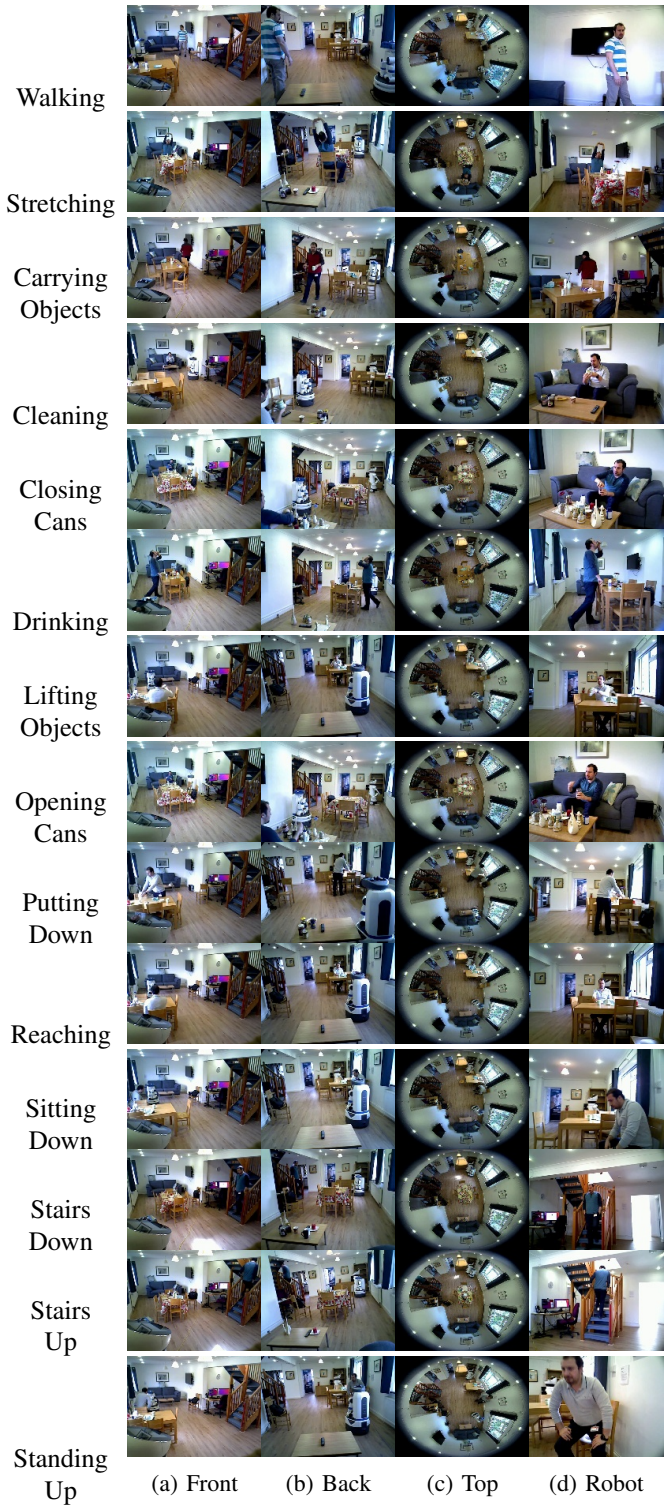

Fig. 1: Example activities of the dataset from all perspectives.

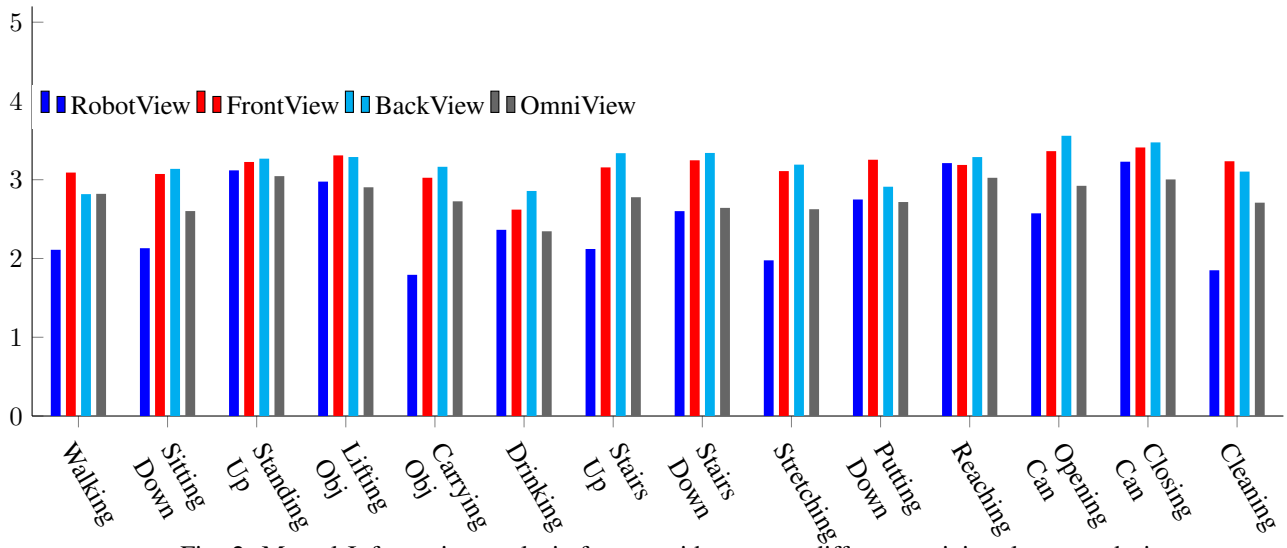

Fig. 2: Mutual Information analysis for one video across different activity classes and views

	Robot View		Front View		Back View		Omni View		Kinetic 400	
Model	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
C3D [55]	55.53	93.83	**70.3**	**97.85**	69.48	97.84	67.48	97.69	71.4	NA
R3D [56]	61.98	94.28	69.04	**97.55**	69.33	97.4	**69.71**	97.25	74.4	91
R2+1D(RGB) [56]	55.6	91.9	65.79	95.91	**66.96**	**96.58**	64.73	95.99	72	90
Slow-Fast(8*8-R50) [57]	55.15	91.61	62.28	**97.25**	**63.62**	96.43	60.65	96.51	77	92.6
Slow-Fast(8*8-R101) [57]	58.57	92.79	59.39	**96.51**	60.43	95.61	**61.76**	96.36	77.9	93.2

TABLE IV: Benchmark model on RHM dataset (No Pre-train) and Kinetic_400. Red data are the best of Top1 and the blue ones are related to the best of Top5. The underlined values indicate the highest accuracy for top1 and top5 metrics.

G. Naming Protocol

RHM contains four folders and the name of each folder corresponds with the view name. Each splits into training, testing, and validation folders. The split folder comprises 14 folders which are class names. The clips are inside the class folders with the ordered numbering. The naming protocol is like below:

ClassName_ViewName_clipNumber.avi

For example, Drinking_RobotView_103.avi refers to clip 103 of action class 'drinking' from the Robot Viewpoint.

H. Time Synchronising

All the clips with the same class name and number with various view names are synchronised with each other. For instance, Clip 320 in reaching action, is time-synchronised with the remaining views.

IV. RHM Dataset Analysis

As we embark on exploring fusion for multiple views, it is important to consider mutual information, as well single view performance based on benchmark models comparing different views, prior to two-stream fusion.

A. Mutual Information

To find out the difference between the viewpoints, we calculate Mutual information (MI) [2] analysis for a video in each class and view.

$I(X;Y)$ is a Mutual Information of two variables X and Y with joint probability distribution $P(X,Y)$ [2]:

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

We adopt MI [2] for a video with m frames as below:

$$MI(f_i, f_m) = \sum_{j=1}^{m} P(f_i, f_{i+1}) \log \frac{P(f_i, f_{i+1})}{P(f_i)P(f_{i+1})} \tag{2}$$

Which $MI(f_i, f_m)$ is the sum of all MI of two adjacent frames in a video and f_i is the first frame, and f_m is the last frame.

Since the $MI(f_i, f_m)$ adds all MI for every two adjacent frames together, so we divided the calculated MI by $m-1$ to reach the average result:

$$Ave_{mi} = \frac{1}{m-1} MI(f_i, f_m) \tag{3}$$

Which Ave_{mi} is the average of all MI of two adjacent frames in a video, m is the number of frames, and f_i is the first frame, and f_m is the last frame.

The video is selected randomly in each class. We extract the frames of the video and calculate the mutual information between two simultaneous frames and continue this method until the last two frames. We then perform the same method for the same video in another view. For instance, For walking the video is 100 for all four views. The results of performing our method to realise the difference between the same video in all views are in figure 2. High mutual information means high redundancy and low mutual information means low redundancy between frames in a video.

As it is clear, Robot Viewpoint has the lowest mutual information in all actions except reaching, especially in the actions that involve a significant movement component, e.g. walking. This result could be estimated since the camera has motion and the frames have different information. The fish-eye (Omni) view has the second lowest MI. Front and Back Views have more mutual information since they are fixed on the wall and have a fixed viewpoint.

B. Deep models Analysis

Another method for comparing the viewpoints is performing some benchmark models. We have performed C3D [55], R2+1D [56], R3D [56], and Slow-Fast [57] models.

Table IV shows the results of performing the benchmark models on the RHM dataset. We have additionally added kinetic_400 results to have better information to compare.

One notable outcome of the results relates to the robot's viewpoint. For all models, the Top1 and Top5 accuracy are the lowest for Robot View. It is clear that having motion in the Robot View is the main reason for these results. Another interesting result relates to the Omni (ceiling) view, which achieves two of the best top1 accuracy results. This is because of having a good and comprehensive viewpoint from the top for all activities. We also note that the Front View attain the all top5 best accuracy results except the R2+1D model. In terms of the top1 and top5, the wall fixed views (Front and Back Views) are the best which is understandable because they do not have a motion, and also they have a good viewpoint to cover all the action areas. In general, the Front View in the C3D model provides the highest overall accuracy results.

V. Conclusion

We introduced the Robot House Multi-View HAR Dataset (RHM), with Front (static), Back (static), Ceiling (fish-eye), and Robot (dynamic) views. RHM contains 6,701 videos in 14 classes for each view separately. The total number of clips for all views is 26,804 videos. The videos with the same number and the same classes are time-synchronised in different views. We analysed the RHM with mutual information and various benchmark models. The Robot View has the lowest level of mutual information in comparison to other views. Also, the C3D model and Front View had the best results in the benchmark analysis. Our future work will focus on the dynamic choice of complementary channels based on the affordance of views in support of an activity, guided by the mutual information. We aim to explore channel combination and multi-stream activity recongition, with the sole goal of improving activity recognition for more complex cases.

References

[1] M. H. B. Abadi, M. R. S. Alashti, P. Holthaus, C. Menon, and F. Amirabdollahian, "Robot house human activity recognition dataset," *UKRAS21*, 2021. 1

[2] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999. 1, 5

[3] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 32–36. 1

[4] 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. 1

[5] 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. 1

[6] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2010, pp. 48–55. 1

[7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563. 1, 3

[8] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1996–2003. 1

[9] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine vision and applications*, vol. 24, no. 5, pp. 971–981, 2013. 1

[10] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012. 1, 3

[11] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer vision in sports*. Springer, 2014, pp. 181–208. 2

[12] A. Nagendran, D. Harper, and M. Shah, "New system performs persistent wide-area aerial surveillance," *SPIE Newsroom*, vol. 5, pp. 20–28, 2010. 2, 3

[13] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *European Conference on Computer Vision*. Springer, 2012, pp. 52–61. 2

[14] O. Kliper-Gross, T. Hassner, and L. Wolf, "The action similarity labeling challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 615–621, 2011. 2, 3

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732. 2, 3

[16] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016. 2, 3

[17] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019. 2, 3

[18] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019. 2, 3

[19] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017. 2, 3

[20] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018. 2, 3

[21] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019. 2, 3

[22] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman, "The ava-kinetics localized human actions video dataset," *arXiv preprint arXiv:2005.00214*, 2020. 2, 3

[23] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the kinetics-700-2020 human action dataset," *arXiv preprint arXiv:2010.10864*, 2020. 2, 3

[24] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850. 2, 3

[25] F. Mahdisoltani, G. Berger, W. Gharbieh, D. Fleet, and R. Memisevic, "On the effectiveness of task granularity for transfer learning," *arXiv preprint arXiv:1804.09235*, 2018. 2, 3

[26] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Actor and observer: Joint modeling of first and third-person videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7396–7404. 2, 3

[27] B. Jia, Y. Chen, S. Huang, Y. Zhu, and S.-c. Zhu, "Lemma: A multi-view dataset for learning multi-agent multi-task activities," in *European Conference on Computer Vision*. Springer, 2020, pp. 767–786. 2, 3

[28] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J. C. Niebles, "Home action genome: Cooperative compositional action understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 184–11 193. 2, 3

[29] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100," *International Journal of Computer Vision*, vol. 130, no. 1, pp. 33–55, 2022. 2, 3

[30] D. Damen, H. Doughty, G. M. 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. 2

[31] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandréa, C.-E. Bichot, C. Garcia, and B. Sankur, "The liris human activities dataset and the icpr 2012 human activities recognition and localization competition," *LIRIS Umr*, vol. 5205, 2012. 2, 3

[32] M. Dallel, V. Havard, D. Baudry, and X. Savatier, "Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics," in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, 2020, pp. 1–6. 2, 3

[33] G. Abebe Tadesse, O. Bent, K. Weldemariam, M. A. Istiak, T. Hasan, and A. Cavallaro, "Bon: An extended public domain dataset for human activity recognition," *arXiv e-prints*, pp. arXiv–2209, 2022. 3

[34] J. Chung, C.-h. Wuu, H.-r. Yang, Y.-W. Tai, and C.-K. Tang, "Haa500: Human-centric atomic action dataset with curated videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 465–13 474. 3

[35] M. Monfort, B. Pan, K. Ramakrishnan, A. Andonian, B. A. McNamara, A. Lascelles, Q. Fan, D. Gutfreund, R. Feris, and A. Oliva, "Multi-moments in time: Learning and interpreting models for multi-action video understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[36] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, "Movienet: A holistic dataset for movie understanding," in *European Conference on Computer Vision*. Springer, 2020, pp. 709–727. 3

[37] A. G. Perera, Y. W. Law, T. T. Ogunwa, and J. Chahl, "A multiviewpoint outdoor dataset for human action recognition," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 5, pp. 405–413, 2020. 3

[38] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhagen, and L. V. Gool, "Large scale holistic video understanding," in *European Conference on Computer Vision*. Springer, 2020, pp. 593–610. 3

[39] A. Piergiovanni and M. Ryoo, "Avid dataset: Anonymized videos from diverse countries," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 711–16 721, 2020. 3

[40] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2616–2625. 3

[41] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0. 3

[42] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "Hacs: Human action clips and segments dataset for recognition and temporal localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8668–8678. 3

[43] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick *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. 3

[44] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056. 3

[45] H. Zhao, Z. Yan, H. Wang, L. Torresani, and A. Torralba, "Slac: A sparsely labeled dataset for action classification and localization," *arXiv preprint arXiv:1712.09374*, vol. 2, 2017. 3

[46] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei, "Every moment counts: Dense detailed labeling of actions in complex videos," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 375–389, 2018. 3

[47] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, and M. Kankanhalli, "Benchmarking a multimodal and multiview and interactive dataset for human action recognition," *IEEE Transactions on cybernetics*, vol. 47, no. 7, pp. 1781–1794, 2016. 3

[48] P. Weinzaepfel, X. Martin, and C. Schmid, "Human action localization with sparse spatial supervision," *arXiv preprint arXiv:1605.05197*, 2016. 3

[49] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*. Springer, 2016, pp. 510–526. 3

[50] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International conference on image processing (ICIP)*. IEEE, 2015, pp. 168–172. 3

[51] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970. 3

[52] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *2013 IEEE workshop on applications of computer vision (WACV)*. IEEE, 2013, pp. 53–60. 3

[53] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for event and object recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3272–3279. 3

[54] S. Bedaf, G. J. Gelderblom, D. S. Syrdal, H. Lehmann, H. Michel, D. Hewson, F. Amirabdollahian, K. Dautenhahn, and L. De Witte, "Which activities threaten independent living of elderly when becoming problematic: inspiration for meaningful service robot functionality," *Disability and Rehabilitation: Assistive Technology*, vol. 9, no. 6, pp. 445–452, 2014. 4

[55] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497. 5, 6

[56] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459. 5, 6

[57] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211. 5, 6