2023 IEEE/SICE International Symposium on System Integration (SII) | 979-8-3503-9868-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/SII55687.2023.10039226

A Survey and Comparison of Activities of Daily Living Datasets in Real-life and Virtual Spaces*

Swe Nwe Nwe Htun, Shusaku Egami, and Ken Fukuda

***Abstract*— Academics, researchers, and industrial experts have made significant efforts in human activity recognition by considering different perspectives, such as benchmark datasets, utilization of smart sensors, and development of recognition algorithms. In addition, the emerging trend of avatar technologies has drawn the attention of researchers, which assemble cognitive abilities, including activity recognition, to realize social activities and overcome boundaries of human energy, time, and environment. These could widely assist older adults, independent living care, and medical care, by the data obtained from human activities. Since human behavior in real-life is extremely complex, assembling meaningful recognition is paramount. The performance of existing recognition models depends heavily on the datasets. However, it has constraints to acquiring datasets with rich information due to the limited budgets, the specific areas that can be used for human activity, the limited number of actors, and other ethical reasons. This paper considers the critical criteria to fill the requirements for human daily activity recognition and presents a survey of the existing datasets of activities of daily living in both real-life settings and virtual spaces. It also presents new challenges and potential advances in behavior recognition technology.**

I. INTRODUCTION

Research and development regarding human activity recognition are essential parts of science and technology applications such as ambient-assisted living systems, including abnormal event detection [1], real-time action recognition [2], mobilization activity recognition in an intensive care unit (ICU) [3], motor severity analysis for Parkinson's disease [4], and others. The activity recognition task is in the last process of the pipeline of in-depth analysis, such as data collection, object detection, segmentation, feature extraction for human postures and mobility, analysis of observed data, and classification [5]. Thus, proper data collection is the key to successful recognition. Additionally, to enable communication between virtual agents and humans, absorbing human perceptual knowledge and concepts into data is required [6]. Preparing a rich dataset of human activity is an essential foundation in any of these cases. During data collection, it is necessary to observe human behavior patterns from the complex daily activities in various environments and to embed the corresponding knowledge as semantic information. As shown in Fig. 1, the robot learns a person's activities in a kitchen using collected diverse samples and analyses. As we can see in the figure, when the camera captures the video data, an essential thing that comes into play is human posture and motion. The objects appearing in the scenes are also essential for human-object interactions. Thus, the more knowledge one has in the dataset, the better analyses one can conduct to benefit from unexplored opportunities.

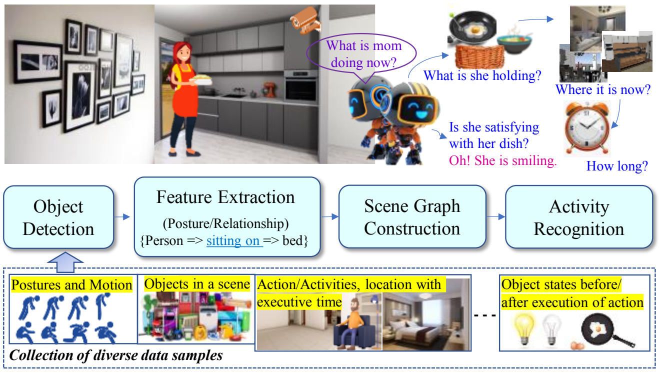

Figure 1. A recognition process using collected knowledge and analyses.

In recent years, significant research effort has been made to collect data on activities of daily living (ADLs) for activity analysis and recognition tasks. Although the ideal approach can be successfully applied to the acquired original datasets, it is unsuitable for different datasets, including unfamiliar scenes. This usually happens when the contextual information contained in the data is obtained with the specific purpose of effectively assisting recognition tasks. Because of the complex human perspective, understanding the relationship between action and activity (compositional actions) in a long time series is a critical issue. However, to the best of our knowledge, there are a few simulated datasets involving attention, spatial, and contact relationships in consecutive low-level actions to high-level activities [7]. Simulating a dataset containing rich information requires supportive resources. Most datasets are limited to the following potential conditions: 1) protection of privacy issues, 2) need for indoor furniture and appliances to create a complete scene, 3) need for recording expensive equipment and sensors to record videos, and 4) an insufficient number of participants to perform several activities.

A solution for overcoming the above limitations is virtual reality (VR) technology, which has been used in 3D games. Therefore, if such a technology can be utilized to shape human lives, the possibility of building a virtual home [8] that can reduce labor and time consumption of human energy resources for data collection has emerged. This survey work also expresses our interest in conducting research on applying virtual agents to perform complex daily activities during usual household maintenance. This survey aims to discover benchmark datasets suitable for activity recognition and

*This paper is based on results obtained from projects JPNP20006 and JPNP180013, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Swe Nwe Nwe Htun, Shusaku Egami and Ken Fukuda are with the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064 Japan (corresponding author: Ken Fukuda, phone: +81-3-3599-8049, e-mail: {swenwe.nwehtun, s-egami, ken.fukuda}@aist.go.jp).

assess the existing datasets' ability to meet real-world demands effectively. In this paper, the related research work is described in Section 2. In Section 3, the criteria for generating datasets and a comparison of currently available ADLs datasets are presented. The challenges and potential trends of simulating datasets with enriched information are described in Section 4. Finally, Section 5 concludes the paper.

II. Related Research Work

Chaquet et al. [5] proposed a 68-dataset survey, including different sets of actions, i.e., a typical action that appears in various situations recorded by spectrum cameras. Their survey presented datasets published between 2001 and 2012. These datasets were classified by considering different features such as context (free/controlled), type of interaction (human-human and human-object), type of participant (professional researcher or ordinary person), and type of view. For this purpose, different cameras (static and moving) were used. A thorough study on performing a realistic action analysis was presented. However, the simulation of an environment as close to real-life conditions as possible and the performance of various activities were not sufficiently advanced for the age range considered.

Beddiar et al. [9] also mentioned multiple datasets in their vision-based activity recognition survey. In that survey, the datasets were divided into groups by categories of actions, viewpoints, and data characteristics such as gestures, i.e., sit and bend positions, primitive actions, activities, behavior, and social interactions. However, the literature indicating how to construct an ADLs dataset similar to the real world is insufficient.

Consequently, in this survey, we provide a complete description of datasets used by the public and identify the remaining needs.

III. Datasets of Activities of Daily Living (ADLs)

Since currently available ADLs datasets must provide real-life challenges, the factors considered for preparing such a dataset are analyzed here. In this survey, we focus on the vision-based ADLs datasets category, which provides the reality of situations using RGB and depth cameras in indoor environments. According to the best of our knowledge, the available activity benchmarks exhibit some drawbacks. Therefore, a summary of the criteria required to generate ADLs datasets in a real-residential setting is presented. Some parts of the survey results are referred to by the Toyota Smarthome dataset [10].

A. Criteria to Consider in ADLs Datasets

The first criterion is context awareness. The interpretation of a scene using visual data requires context information which is derived from simultaneously looking at the background of the viewpoint (including objects, locations, etc.) while capturing the videos. Some activity datasets are focused on specific activity recognition tasks and thus are biased in the sense of daily living activity distribution. The context contained in a dataset should be bias-free if it tries to capture a real-life environment as closely as possible, such as the Home Action Genome (HOMAGE) [7], the Toyota Smarthome [10], and Charades [11]. Thus, we investigate whether the existing datasets were sufficiently collected with plentiful context. In this survey, the videos recorded in a typical residential home complete with household furniture, or the videos recorded including objects commonly used by people to simulate a typical residential setting, are considered to be "context-bias-free." In contrast, if the dataset does not provide plenty of human-object interaction and naturalness for the sake of specific object detection tasks, it is considered to be "context-biased."

The second criterion is framing techniques, which indicate how video data is captured. In this paper, we defined three levels of framing techniques. The first is the "high framing technique," which captures close-up video sequences so that the person's actions can be seen in detail. For example, a person captured in the center of the frame with several close-ups of his/her behavior during a medicine-taking process (i.e., the person holds a cup of water, takes the pills, drinks water, etc.). The distance measurement between the hands holding the medicine and the mouth might also enhance the recognition process of medication intake. The second one is the "medium framing technique." It means that it is not recorded as a close-up specific view, although the cameras are purposely installed to observe daily activities. In real-life, daily activities in uncontrolled views are recorded using inexpensive cameras installed in home environments. Such a technique is considered the "low framing technique."

The third criterion is time-lapse. A time-lapse video is a video edited to be played faster than its original speed. For instance, a system may fail to extract the foreground of a person when he/she remains in a place for a long time [1], such as a person sleeping on a bed for a long time. To overcome such cases, some datasets can speed up the video and show long hours of a real-time film in fewer hours (such as time-lapse video). Therefore, we consider that using a real-time video could be more practical in a real-life environment, even though it would be technically challenging for the detection module. Therefore, it is considered in this paper that studying the benefits of not having time-lapse data in preparing datasets plays an important role.

The fourth criterion is spontaneous (or natural) action performed by a human. Here, the criterion is assumed to have three conditions. When a participant acts exactly to a pre-directed script, the data is considered to be "low spontaneous acting." If the participants can act out according to their own will without strictly following the pre-directed script, it will be considered "medium spontaneous acting." If the participants are free to perform actions at their will in a real-home environment, the data is considered to be "high spontaneous acting."

The fifth criterion is the cross-view cross-scene setting. Integrating the information obtained from different-point-of-view multiple cameras sensing the objects can improve the robustness of activity recognition algorithms. It might be helpful to locate the cameras at places that provide a wide range of scenes for the entire room. In addition, training and testing different scenes using multiple camera views could be helpful in practice. Such information is called a cross-view cross-scene setting [12]. Thus, as a valid study, we examine whether or not a person's positions and situation appear in different parts of multiple scenes. Most datasets in this survey have attempted to use the cross-view cross-scene. Through this observation, the use of multiple camera angles is a way to create a video dataset that can visualize the person from multiple angles and determine the condition of each activity.

The final criterion is the duration of human behavior observations; it refers to long-time video recordings of activities day and night. Some research efforts have focused on observing the behavior of not only humans but also animals, such as dairy cows [13]. Behavior includes human posture, moving patterns such as walking, and many regular activities. Our valid study focuses on how long an activity takes and how many activities appear in a video. Although the currently studied datasets are not 24/7-hour continuous videos, they can indicate how long an activity takes to perform. For example, the "clean toilet" activity can be completed in about 5 seconds, while deep cleaning takes at least 3 minutes. One can also know that the "cooking" activity takes about 40 minutes. Thus, in the current study of datasets, activity videos of more than 3 minutes are considered to be "high-support" behavior data, and videos of less than that are considered to be "low-support." We believe that applying time series data is a challenging task and is representative of how a person utilizes time in daily activities.

B. *Comparison of ADLs Datasets in both Real-Life Settings and Virtual Spaces*

Considering the criteria mentioned above, we compare and describe the nature of recently released publicly available datasets. This paper compares ADL datasets obtained from real-life and virtual spaces with the same criteria. Dataset preparation is one of the main hurdles to developing and validating activity recognition frameworks. One must prepare a close-to real-life environment and hire actors, and changing the set-up condition frequently is not manageable. Several studies utilize virtual reality (VR) environments for dataset preparation [14]. VR can simulate the environment of daily living and the interaction between virtual humans and their environment by providing scripts for virtual humans. However, doing so is like giving pre-directed scripts to the participants on how to behave in real-life ADLs datasets. Natural realism will be one of the crucial challenges in simulating daily activities in VR, and that is the same with controlled real-life ADLs datasets. Therefore, this paper surveys ADL datasets from VR environments and real-life environments with the same criteria.

i. *Charades Dataset* 2016

Charades [11] is a large-scale dataset that provides samples representing daily dynamic scenes. This dataset was acquired from residential homes, including 15 types of rooms (bedroom, living room, etc.) to imitate a typical home. To represent human activities in the real world, the authors asked 267 participants to record their casual household activities at home via the Amazon Mechanical Turk (AMT), instead of creating the video shooting environment in a laboratory. First, 549 movie scripts were analyzed using term frequency to learn how frequently objects and actions represent activities. The vocabularies of actions are then distributed to create a pre-directed script, which is acted out by the workers. The obtained scenarios are further used to generate the videos. After performing the annotation process by analyzing (verb, proposition, and noun) triplets in scripts, 157 actions that most people do every day are obtained. For example, from the scenario "a person drinks milk from the fridge, and then he/she walks out of the room," actions that involve human-object interactions, such as "opening the refrigerator" and "drinking from a cup/bottle," can be obtained. Charades provided an environmental/behavioral recognition community with a way to realize the diversity, challenges, and new opportunities.

ii. *CLAD Dataset* 2017

Tayyub et al. proposed the activity CLAD dataset [15], which displays diverse scenarios of complex and long-range activity videos with rich crowdsourced annotations. In this dataset, 5 actors were carefully selected from individuals without prior knowledge of image processing to act out natural scenarios. Top-level activities were assigned to the participants, who were allowed to use their favorite objects during the scenarios. Thus, the participants were only given minimal instructions to perform the requested scenes, and the scenarios were random, according to the participants' desire. Consequently, it can be assumed that a great effort was made to acquire unbiased video data based on the information obtained from context and spontaneous action. However, the data were not free from the environmental context since the videos were recorded in a specific research lab for various ethical reasons. The dataset comprises 62 videos with 3 high-level activities, ranging from 3 to 10 minutes. Microsoft Kinect v2 sensor was used for this purpose. However, the participants' behaviors were mainly filmed to enter the frame or a specific view, and the dataset was considered a high framing technique.

iii. *DAHLIA Dataset* 2017

Similarly, the DAHLIA dataset [16] was introduced to focus on activity monitoring for older adults in realistic conditions. 44 participants were asked to act out 7 high-level semantic activities, such as cooking and having lunch, after receiving a few instructions. In making this dataset, low-level atomic actions to high-level activities were considered for recognition tasks. For example, the "cooking omelet" activity is composed of "cutting tomatoes and onions," "using an eggbeater," "putting cheese," etc. In more detail, logic describes human gestures often repeated in action, such as moving a hand forward. To shoot 51 long-range videos, 29 males and 25 females in the 23–61 years age group acted to enrich the variations for an average of 39 min per person. 3 Kinect^{TM} v2 sensors were used to fully monitor the kitchen using 15 fps for four types of streams: RGB, depth map, 3D skeleton data with 25 joints of body, and body index. Since the actions were filmed only in the kitchen, it can be assumed that the background variation was not particularly rich.

iv. *Toyota Smarthome Dataset* 2019

Das et al. proposed a dataset (called Toyota Smarthome [10]) that provides challenging information for recognizing natural and diverse activities. Older adults (60 to 80 years old) were employed to record the videos using 7 Kinect v1 cameras. To achieve a real natural scenario, these adults were not given pre-specified scripts and were only allowed to know that the video was being recorded. 18 individuals were monitored for 8 hours daily (morning to afternoon). In this way, the contextual information was made to be free. The movement of an older adult did not depend on the same place, and it was constantly changing, meaning that there were variations in the distance between the camera and the participant. In some cases, a participant's behavior is occluded, creating a challenge in the process from object detection to the final recognition stage. The dataset consists of 31 activities representing one activity per video and a total of 16,115 sample streams. Thus, performing a recognition task using the Toyota Smarthome dataset presents a difficult path, leading to an improvement in the recognition process.

v. *NTU RGB + D* 120 *Dataset* 2020

The large-scale NTU RGB + D 120 dataset [17] has emerged for 3D human activity analysis by collecting videos recorded from 106 individuals. The activities often depend on gender, age, physical condition, and cultural aspects of people living in different regions. Thus, the authors attempted to create a realistic and qualitative dataset by collecting data samples from people of different ages (10–57 years old) and broad cultural backgrounds (15 countries). In addition, the video sequences were captured to support a changeable atmosphere under 96 different environmental conditions with occlusion and illumination variations. The data samples were obtained from 155 non-restricted camera viewpoints. This dataset comprises 114,480 video samples and 8 million frames, including RGB, depth, and 3D skeleton data with 25 body joints and infrared data. A Microsoft Kinect v2 sensor was used for this purpose.

vi. *HOMAGE Dataset* 2021

HOMAGE [7] was the first benchmark dataset that included scene graph information. It comprises multi-dimensional synchronized video streams obtained from multiple perspectives (cross views) with low-level action and high-level activity interpretation. Similar to the Toyota Smarthome dataset, a series of actions in a residential house is captured by adding occlusions. The videos were sufficiently recorded for a certain amount of time and used for behavior observation. To cover different types of activities, the authors utilized a taxonomy called the American Time Use Survey, a detailed survey of lists describing the activities people do during a day. Subsequently, 27 participants were given instructions to act out actions, but they were let free to do what they wanted without imposing limitations in time and actions. Thus, the participants acted as naturally as possible. 12 sensor types were installed in different locations for third-person and ego views. According to the dataset statistics, 75 activities containing 453 atomic actions were obtained and included in 5,700 videos. The structure of the scene graph includes 86 objects and 29 relationship classes.

vii. *VirtualHome* 2018

In the datasets described above, recognizing the activities is a vital process for AI-assisted health care. However, acquiring and annotating data in a residential setting is time-consuming and tedious. To overcome this problem, the VirtualHome [8] was proposed. Similar to the Charades data, the AMT workers were first asked to create verbal scenarios that illustrate daily activities, e.g., the high-level activity "Turn the night light on" and its detail "I walk to the small-battery-operated lamp and switch it on." The next step was to convert descriptions into programs using a graphical programming language, e.g., "step1 = [WALK]<bedroom>(74), step2 = [WALK]<tablelamp>(103), step3 = [TURNTO]<tablelamp>(103), step4 = [SWITCHON]<tablelamp>(103)." Four virtual agents (2 adult males and 2 adult females) were simulated in 7 apartments. The animation was also added to randomly select a camera to watch the entire agent motion. This opens a new avenue of research to obtain cross-view data. 12 top-level activities including 75 atomic actions were available. Automated ground truth data consist of the timestamp of each action step, skeleton pose data, image segmentation data, depth, and optical flow. In that work, the program dataset was released to the public. A set of large-scale videos, which can be generated using programs with advanced features, might be available to the public later.

viii. *SIMS4ACTION Dataset* 2021

The SIMS4ACTION dataset [18] was constructed based on a commercial video game called THE SIMS 4 to create a real-life setting from virtual gaming. The interior design was simulated according to Toyota Smarthome [10] residential settings. Eight virtual agents (4 older adults, 2 adults, and 2 young adults) were created to simulate everyday activities with complex human traits, variations in physical and emotional conditions, and different ages and genders. In addition, two apartments, including three rooms (kitchen, dining room, and living room), were simulated. However, since the virtual agents are under the user's control, it is difficult to say that the context is free. Two types of camera modes were used in the virtual home interior. The first type was a fixed camera mode, which was set up to watch the entire room, and the second one was a moving camera. Finally, 10 high-level household activities were simulated, and 942 videos were obtained.

ix. *iGibson* 2021

iGibson [19] is a new simulated environment obtained from a 3D reconstruction of 6 interactive household tasks in large-scale scenes. It was intended to explore a suitable playground for learning and improving the robot's skills. This dataset was composed of 15 fully interactive scenes with 108 rooms and 570 object models using multimodal high-quality virtual sensor signals, including RGB, depth, semantic segmentation, etc. iGibson provides a super friendly web interface in which humans can interact with all details in the scenes using mouse and keyboard commands to obtain a variety of scenarios. Specifically, iGibson demonstrated the ability to maintain and update physical states such as temperature (i.e., cooking a fish in the stove at a controlled temperature), moisture level (i.e., water droplets dropping from the faucet on to the receptacle), and so on. Moreover, a generative system for the initialization of tasks and scenes was introduced. That is, a list of logical states is given as inputs, and the system can generate the physical states that satisfy all the requirements, such as providing different books on top of a table with different poses, different object models, and different scenes. Participants were asked to provide a demonstration, such as cooking onions, using a VR interface.

iGibson contained kinematic states (i.e., InsideOff, OnTopOf, under, etc.) and nonkinematic states (i.e., frozen, cooked, burned). These predicated logical states share the same underlying physical states (namely, temperature, sliced, soaked, dusty, etc.), based on continuous states.

A comparison of datasets based on the above stated criteria is shown in Table I. The number of actions, activities, subjects, video duration, modalities of datasets, and sensors used for data acquisition is compared in Table II.

IV. Challenges and Potential Trends

Although researchers from the industry and academia have made significant efforts in the field related to behavior recognition technologies, challenges still exist. According to our current studies and to the best of our knowledge, new advances and potential milestones are highlighted below.

A. *How to Define ADLs for Recognition Tasks*

Action datasets have gradually emerged over the past two decades. Although the datasets were analyzed in the previous section, there is no clear distinction between actions and activities. We could consider what kinds of scene definitions should be stored to obtain a qualitative ADLs dataset. A comparison among existing datasets is described in Table III. An atomic action could be considered to employ more than one entity (e.g., a person, an object, and their relationships). Based on this assumption, some research on the definitions of each action and object has begun by constructing ontology [20]. An activity can be composed of a series of atomic actions (compositional actions). Both the actions and objects constitute the actual performance of a task performed by a person and require time and resources to be completed. A compositional activity can consist of multiple activities with a long duration. Let us see the example of "preparing a meal," as illustrated in Fig. 2. When a person enters the kitchen and makes contact with the equipment, a series of atomic actions appears, e.g., "holding a knife, peeling tomatoes, taking a piece of bread, holding a plate and so on." Among these actions, "opening the oven, putting a dish, closing the oven, switching on the oven, switching off the oven, taking out the plate" can be regarded as an activity called "heating food." Similarly, the "making coffee" activity occurs when the person continues performing the actions of "taking a cup, pouring coffee, pouring water, taking a spoon, and mixing coffee." Therefore, "preparing a meal" is a compositional activity made up of two activities: heating food and making coffee. The consideration of events is excluded from some ADLs datasets used for recognition tasks. It can be assumed that an event might start when a condition is suddenly changed, i.e., from sitting to standing. Thus, in principle, an event is always related to one action. The timing of an event depends on the execution of an action. In other words, events are discrete and not continuous. Intensive research based on these concepts has already started to make progress [21,22].

TABLE I. Comparison of ADLs Datasets Based on Criteria Required for a Real-Life Environment

Real-World Setting						
Dataset	***Context***	***Frame Tech.***	***Time-lapse***	***Natural acting***	***Cross-view***	***Time length***
Charades [11]	Free	High	No	Low	Yes	Low
CLAD [15]	Baised	High	No	Medium	No	High
DAHLIA [16]	Free	Medium	No	Medium	Yes	High
NTU RGB+D 120 [17]	Free	High	No	Low	Yes	-*
Toyota Smart home [10]	Free	Low	No	High	Yes	Low
HOMAGE [7]	Free	Low	No	High	Yes	Low
Virtual-Home Setting						
VirtualHome [8]	Baised	Low	-*	Low	Yes	-*
Sims4Action [18]	Baised	Medium	No	Low	Yes	Low
iGibson [19]	Baised	Low	-*	Low	Yes	-*

*: no data available

TABLE III. Comparison of ADLs Datasets Based on Relationships between A Low-Level Action and A High-Level Activity

Real-World Setting				
Dataset	***Action***	***Activity***	***Episode (Composite Activity)***	***Event***
Charades [11]	Yes	No	Yes	undefined
CLAD [15]	Yes	Yes	Yes	undefined
DAHLIA [16]	Yes	Yes	No	undefined
NTU RGB+D 120 [17]	Yes	No	No	undefined
Toyota Smarthome [10]	Yes	Yes	Yes	undefined
HOMAGE [7]	Yes	Yes	Yes	undefined
Virtual-Home Setting				
VirtualHome [8]	Yes	Yes	Yes	undefined
Sims4Action [18]	Yes	Yes	No	undefined
iGibson [19]	Yes	Yes	No	undefined

TABLE II. Comparison of Datasets on the Number of Subjects Performing Activities per Video, Modalities, and Data Acquisition

Real-World Setting								
Dataset	***Subject***	***Action***	***Activity***	***Scene***	***Video***	***Duration***	***Modalities***	***Acquisition***
Charades [11]	267	157	-*	15	9,848	30 s	RGB+Flow	AMT
CLAD [15]	5	-*	3	1	62	3-10 m	RGB+Skeleton+Point Cloud	Kinect v2
DAHLIA [16]	44	-*	7	-*	51	40 m	RGB+Depth+Skeleton+Body Index	Kinect v2
NTU RGB+D 120 [17]	106	120	-*	155	114,480	-*	RGB+Depth+3DJoints+IR	Kinect v2
Toyota Smarthome [10]	18	-*	31	3	16,115	1 s-1 m	RGB+Depth+3D skeleton	Kinect v1
HOMAGE [7]	27	453	75	5	5,700	2-5 s	RGB+Audio+Scene Graph	12 sensor types
Virtual-Home Setting								
VirtualHome [8]	6	-*	548	7	-*	-*	RGB+Pose+ Normals+SS+Depth+Flow	AMT, Unity
Sims4Action [18]	8	-*	10	6	942	30 s-1m	RGB	Sims4 game
iGibson [19]	-*	-*	6	15	-*	-*	RGB+Depth+Normals+SS+Flow+LiDAR	VR Interface

*: no data available

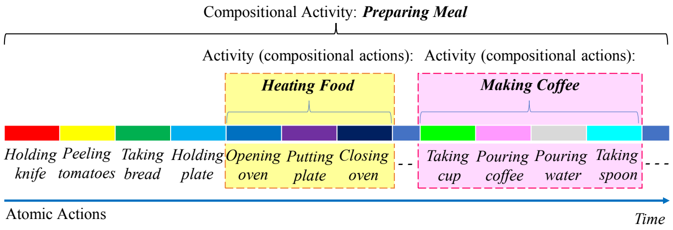

Figure 2. Relationship between atomic actions and a compositional activity.

B. *How to Care for Human Privacy*

Obtaining data to recognize activities is a challenging task due to privacy issues, especially when using a vision-based system. This is why many models have been invented but have not fully represented actual situations. Thus, simulating virtual agents in a virtual environment is a promising method for obtaining multiple training samples. Although activities could be imitated and recognized in a virtual home, there is a question of how to apply simulated virtual data to real-time recognition. It will be an incredible challenge to describe each action of an actual person in real-time using these virtual data. Recently, advanced technologies that provide accessible real-time 3D motion captures have been developed for real-live inputs. Such technologies allow users to stream their visual appearances to the recognition system by transforming a full-body animation using a real-life movement speed. This might work well in real-time detection if there are enough training samples of virtual activities.

C. *How to Include Virtual Agent Behavior*

Human postures and movements differ according to gender, age, weight, and height. In addition, disease-related stooping, muscle weakness, instability, and fear make postures quite different among older adults [23]. Therefore, imitating a body's appearance with many variations requires considerable effort during its development. In addition, each activity may have a different duration (for example, ~1 h for cooking and ~30 min for eating). Also, there is a difference in the sleeping time between naps during the daytime and ~7 h of sleep during the night. Therefore, the duration of activities must be considered to be close to that of real-life cases.

D. *How to be Aware Effect of High-Performance Tools*

High-performance cameras are used in real-life settings to obtain high-quality images that can assist the recognition process. High-speed visible cameras with frame rates of 1,000 fps at full pixel resolution have many advanced features and allow plug-and-play connectivity to the researchIR[1] software of the production company or third-party software such as MathWorks[2]. The cameras automatically identify the filters via the camera controller, thereby removing the burden of tracking. The objects can be visible even in poor light circumstances and the most difficult weather conditions. Therefore, activity recognition based on the data acquired using a high-performance camera might give different results from recognition based on the data acquired using a normal camera. Although the virtual human in virtual spaces is also imported using 3D to achieve a realistic and natural look, there are some limitations. Complex technologies are necessary to simulate the movement of the body parts of a virtual human, like a real human, while performing daily activities. Thus, some unnatural actions and appearances could be executed. Virtual human ascribes activities depend significantly on illumination, textures, occlusion, cluttered background, and viewpoint variations applied to the body and interacted objects. Since the simulation of a virtual environment is custom-made by the developer and is not the same as the real photographs we record, there is a limit to comparing the results of activity recognition using the real-world dataset. Therefore, this survey describes that it is tough to compare recognition results, although it is reasonable to explore the same criteria that assist in creating the dataset.

E. *How to Include Enriched Semantic Information*

The semantic information needs to be considered when designing a meaningful real-life daily scenario and analyzing daily living activities. Recognizing actions, compositional activities, and abnormal events do not depend only on visual appearances (i.e., postures and motion) but also requires reliable information (i.e., context, spontaneous action, and knowledge of an action related to the state, attributes of objects, and execution time of each action). In recent years, a remarkable study (VirtualHome2KG [21,22]) presented a framework for designing and augmenting knowledge graphs using simulated virtual spaces [8] to improve the analysis of daily activities. Most existing datasets do not include the following indicators: 1) spatiotemporal information and 2) hierarchy from low-level action to high-level activity. VirtualHome2KG introduced a knowledge graph representing the executed activities, primitive actions, target objects, object states, execution time, and spatial situations. When representing spatial situations, classes describe the spatial changes for the execution before/after an action. State classes indicate each object's states (e.g., open/closed). Such semantic information was investigated to track changes, such as next/previous relationships, between objects. In addition, the properties of objects are well utilized to define the attributes and affordance (such as grabbable). Moreover, the episodes for daily living were generated using the Markov Chain model to synthesize meaningful real-life scenarios.

Based on the above, it has been proved that semantic information is essential in analyzing human behavior in long episodes. In future studies, we could extend the semantic space of the recognition process by considering time, place, and occasion. From this perspective, features indicating human postures, motions, and definitions of abnormal or normal rules could be described. For instance, a person is lying in a bathtub for long hours. In such a case, the interaction between objects in a scene is not enough for the recognition process of lying posture. Knowledge is needed to determine how to distinguish between abnormal and normal. Such contributions could be given as a great piece of supplemental information to enhance monitoring systems for older adults and independent living care.

[1] https://www.flir.com/support-center/Instruments/try-out-researchir-software/

[2] https://www.mathworks.com/

V. Conclusion

This paper compares nine real and virtual ADL datasets released during the period 2016–2021. The comparison was conducted considering different categories, including context, spontaneous or natural action, time-lapse data, shooting techniques, and others to perform data collection and activity recognition processes. As more and more housing architectures using VR are gradually being investigated, 3D visualization provides people with a visual and a feeling experience. This technology expands into many fields and enables a deep investigation and understanding of cognitive abilities used for activity analysis. In this work, we have discussed five challenges that could be encountered in practice. The first is considering the relationships among actions, activities, and execution time in complex human behaviors. The second is that virtual agents could be utilized to protect human privacy regarding vision-based monitoring aspects. The third is that virtual agents should be designed according to the complex actions of human beings. The fourth is considering the current technological developments of high-performance tools that assist in data acquisition. The final recommendation considers the semantic information by deploying knowledge graphs. We believe that our survey on the simulation of complex household tasks in virtual space will provide a promising path for human-robot collaboration in the near future.

References

[1] S. N. N. Htun, T. T. Zin, and P. Tin, "Image Processing Technique and Hidden Markov Model for an Elderly Care Monitoring System," *J. Imaging*, vol. 6, no. 6, article no. 49, 13 June 2020.

[2] T. T. Zin, Y. Htet, Y. Akagi, H. Tamura, K. Kondo, S. Araki, and E. Chosa, "Real-Time Action Recognition System for Elderly People Using Stereo Depth Camera," *Sensors*, vol. 21, no. 17, article no. 5895, 1 September 2021.

[3] S. Yeung, F. Rinaldo et al., "A computer vision system for deep learning-based detection of patient mobilization activities in the ICU," *NPJ Digit. Med.*, vol. 2, no. 11, 1 Mar. 2019, doi: 10.1038/s41746-019-0087-z. PMID: 31304360; PMCID: PMC6550251.

[4] M. Lu et al., "Quantifying Parkinson's disease motor severity under uncertainty using MDS-UPDRS videos," *Medical image analysis*, vol. 73, no. 102179, 2021, doi:10.1016/j.media.2021.102179.

[5] Jose M. Chaquet, Enrique J. Carmona, Antonio Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comput. Vis. Image Underst.*, pp. 633-659, vol. 117, no. 6, 2013.

[6] Hiroshi Ishiguro, "The realisation of an avatar-symbiotic society where everyone can perform active roles without constraint," *Adv. Robot*, pp. 650-656, vol. 35, no. 11, 2021.

[7] N. Rai et al., "Home Action Genome: Cooperative Compositional Action Understanding," 2021 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (*CVPR*), pp. 11179-11188, 2021, doi: 10.1109/CVPR46437.2021.01103.

[8] X. Puig et al., "VirtualHome: Simulating Household Activities Via Programs," 2018 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (*CVPR*), pp. 8494-8502, 2018.

[9] D. R. Beddiar, B. Nini, M. Sabokrou et al., "Vision-based human activity recognition: a survey," *Multimedia Tools Appl.*, vol. 79, pp. 41–42, Nov. 2020.

[10] S. Das et al., "Toyota Smarthome: Real-World Activities of Daily Living," 2019 *IEEE/CVF Int. Conf. on Computer Vision* (*ICCV*), pp. 833-842, 2019, doi: 10.1109/ICCV.2019.00092.

[11] G.A. Sigurdsson, G. Varol, X. Wang et al., "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding," *In: Leibe, B., Matas, J., Sebe, N., Welling, M.* (*eds*) *Computer Vision – ECCV* 2016, *Lecture Notes in Computer Science*, vol. 9905, Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_31.

[12] Q. Zhang, W. Lin and A. B. Chan, "Cross-View Cross-Scene Multi-View Crowd Counting," 2021 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (*CVPR*), pp. 557-567, 2021, doi: 10.1109/CVPR46437.2021.00062.

[13] K. Sumi, S. Z. Maw, T. T Zin, P. Tin, I. Kobayashi, & Y. Horii, "Activity-Integrated Hidden Markov Model to Predict Calving Time," *Animals*, vol. 11, no. 2, 29 January 2021.

[14] J. Tayyub, M. Hawasly, D.C. Hogg, & A.G Cohn, "CLAD: A Complex and Long Activities Dataset with Rich Crowdsourced Annotations," Sept. 2017, 10.5518/249.

[15] T. Miyanishi, T. Maekawa and M. Kawanabe, "Sim2RealQA: Using Life Simulation to Solve Question Answering Real-World Events," *IEEE Access*, vol. 9, pp. 75003-75020, 14 May 2021, doi: 10.1109/ACCESS.2021.3080275.

[16] G. Vaquette, A. Orcesi, L. Lucat and C. Achard, "The DAily Home LIfe Activity Dataset: A High Semantic Activity Dataset for Online Recognition," 2017 12th *Int. Conf. on Automatic Face & Gesture Recognition* (*FG* 2017), pp. 497-504, 2017, doi: 10.1109/FG.2017.67.

[17] 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," In *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684-2701, 1 Oct. 2020.

[18] A. Roitberg, D. Schneider, A. Djamal, C. Seibold, S. Reiß and R. Stiefelhagen, "Let's Play for Action: Recognizing Activities of Daily Living by Learning from Life Simulation Video Games," 2021 IEEE/RSJ *Int. Conf. on Intelligent Robots and Systems* (IROS), pp. 8563-8569, 2021, doi: 10.1109/IROS51168.2021.9636381.

[19] Li, Chengshu et al., "iGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks," *ArXiv, abs*/2108.03272, 2021.

[20] Nishimura, Satoshi et al., "Ontologies of Action and Object in Home Environment towards Injury Prevention," *The 10th International Joint Conference on Knowledge Graphs*, 2021.

[21] S. Egami, S. Nishimura and K. Fukuda, "VirtualHome2KG: Constructing and Augmenting Knowledge Graphs of Daily Activities Using Virtual Space," *International Semantic Web Conference* (*ISWC*) *Posters, Demos, and Industry Tracks*, 2021.

[22] S. Egami, S. Nishimura and K. Fukuda, "A Framework for Constructing and Augmenting Knowledge Graphs using Virtual Space: Towards Analysis of Daily Activities," 2021 *IEEE 33rd Int. Conf. on Tools with Artificial Intelligence* (*ICTAI*), pp. 1226-1230, 2021, doi: 10.1109/ICTAI52525.2021.00194.

[23] J. T. Tavares, D. A. Biasotto-Gonzalez, N. C. Boa Sorte Silva et al., "Age-Related Changes in Postural Control in Physically Inactive Older Women," *J. of geriatric physical therapy*, vol. 42 no. 3, E81–E86. https://doi.org/10.1519/JPT.0000000000000169.