A Study of User Intent in Immersive Smart Spaces

Kelsey Rook*, Brendan Witt†, Reynold Bailey‡, Joe Geigel‡, Peizhao Hu‡, and Ammina Kothari‡
*Andrews University
†University of Maryland, Baltimore County
‡Rochester Institute of Technology
rook@andrews.edu, bwitt1@umbc.edu, {rjbvcs, jmg1590, hxpvcs, abkgpt}@rit.edu

***Abstract*—Smart spaces are typically augmented with devices capable of sensing various inputs and reacting to them. Data from these devices can be used to support system adaptation, reducing user intervention; however, mapping sensor data to user intent is difficult without a large amount of human-labeled data. We leverage the capabilities of head-mounted immersive technologies to actively capture users' visual attention in a unobtrusive manner. Our contributions are three-fold: (1) we developed a novel prototype that enables studies of user intent in an immersive environment, (2) we conducted a proof-of-concept experiment to capture internal and external state data from smart devices together with head orientation information from participants to approximate their gaze, and (3) we report on both quantitative and qualitative evaluations of the data logs and pre-/post-study survey data using machine learning and statistical analysis techniques. Our results motivate the use of direct user input (e.g. gaze inferred by head orientation) in smart home environments to infer user intent allowing us to train better activity recognition algorithms. In addition, this initial study paves a new way to conduct repeatable experimentation of smart space technologies at a lower cost.**

I. Introduction

Smart spaces are areas augmented by devices that can interface with and regulate appliances in the home, such as heating and ventilation systems, lights, voice assistants such as Amazon Alexa, and televisions. They focus on automating tasks done in the space in order to increase productivity and reduce user intervention. While automation based on explicit user interaction with devices is simple, figuring out user intent implicitly is more difficult and requires experimentation and the training of complex algorithms to determine what smart-spaces' occupants are doing. Activity recognition and prediction, which use sensor data from and observations about an environment to understand the general activity or task a user is currently performing and anticipate future actions, has become an active area of research since the birth of ubiquitous computing [1], with the goal of more naturally and efficiently predicting what resources/information/device settings etc. are needed for various tasks [2].

Without the understanding of user intent, activity recognition relies solely on data collected from smart home objects. Sensor data is used to infer activities being performed, but sensors can be used for multiple purposes and sensor data is often noisy, making activity recognition difficult [3]. Some approaches have gone beyond activity recognition to attempt activity prediction, with the intent of recognizing what activities are being performed currently and then predicting what activities users will perform next based on past events or sensor data. Some of these activity prediction algorithms simply predict a next activity given a sequence of events that has previously occurred based on causality programmed by the developer, while others also take the timing of events into account, and can provide an estimate of when the next activity is likely to occur [4]. Many of these activity recognition and prediction algorithms require training in the form of supervised machine learning with manually labeled data [3].

In addition, studying activity recognition algorithms can be cumbersome, as setting up the experimentation environments is time-consuming, costly, and these environments can be difficult to augment appropriately. For example, replaying events accurately requires a complex and time-synchronized distributed system by itself. The use of mixed reality not only paves a new way for repeatable experimentation but also provides a unobtrusive method for obtaining user visual attention and helping us to infer user intent naturally.

In this paper, we present a study and experimentation on using mixed reality to infer user intent and improve activity recognition. The experimentation consists of various physical and virtual smart home devices which create a space that is modular and can be replayed effectively. Participants immerse into this space through head-mounted augmented reality gear (e.g., Microsoft HoloLens) which offers additional sensing capabilities, such as head orientation, head position, and hand gestures. To measure the effectiveness of this new method for inferring user intent, we conduct a user study involving 30 participants. Participants are asked to perform two daily tasks within the immersive smart space. We log all data from both physical and virtual smart devices and HoloLens. Data from sensing devices and a participation questionnaire (pre/post experimentation) is systematically analyzed with a goal to answer the following questions:

1) *Can the use of head orientation and attention data collected from a mixed reality headset complement data from smart devices to enable meaningful recognition of user activity?*
2) *How does a user's previous experience with smart home devices and/or their previous experience with augmented reality influence their ability to complete tasks in a mixed reality space?*
3) *Does the use of head orientation data alone allow for better recognition of user activity than the states of devices in the user's environment?*

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

The use of mixed reality as an attempt to understand and infer user intent is new and viable, due to the rapid adoption of mixed reality wearable devices making direct extraction of user visual attention possible in an unobtrusive manner. We summarize our main contributions as follows:

- A user study on how mixed reality can improve activity recognition and the understanding of user intent. We also share the lessons learned from constructing the experimentation system and user study.
- The development of the first mixed reality experimentation system that enables studies of user intent in an immersive and repeatable environment.
- Quantitative and qualitative evaluations of data logs and survey data to measure the effectiveness of this new method.

The remainder of this paper is organized as follows: Section II gives an overview of related work in activity recognition and mixed reality solutions. Section III describes the setup of our mixed reality environment. Section IV provides details about our user study, and results are presented in Section V. Sections VI focuses on the lessons learned and discusses possible future directions of our work. Finally, Section VII concludes the paper.

II. Related Work

Prior work on activity recognition and prediction is highly applicable to smart home environments. In fact, physical smart homes have been used as a suitable test space for researchers studying activity recognition and prediction. The CASAS project [5], for example, focuses on the collection and sharing of smart home sensor data which is used to train recognition and prediction machine learning algorithms. Virtual smart home environments are also beginning to emerge due to their convenience; for example, OpenSHS provides a cross-platform 3D smart home simulator for dataset generation [6], and SIMACT is designed for research in the area of activity recognition [7]. Virtual Reality has also been used as a test-bed to design a home environment quickly and inexpensively, in order to collect data on which to create an activity recognition system [8], and has been combined with gesture sensors to conduct occupational therapy exercises and to collect important medical data [9]. Researchers have also developed software to visualize and control smart home environments in virtual reality [10]. Augmented reality provides the further benefit of physical presence while still affording the flexibility of including virtual devices to augment the space [11].

While collecting data from smart devices is common in the field of pervasive computing, the collection of first-person data is still a novel idea within the context of activity recognition. Although most smart home datasets involve only information from sensors placed around the home, such as thermostats and motion sensors, several researchers have tried collecting data from the user as well. For example, researchers have attempted to use cameras attached to the subject to observe actions up close or across a large area, allowing an algorithm to learn from this video feed over time [12], [13].

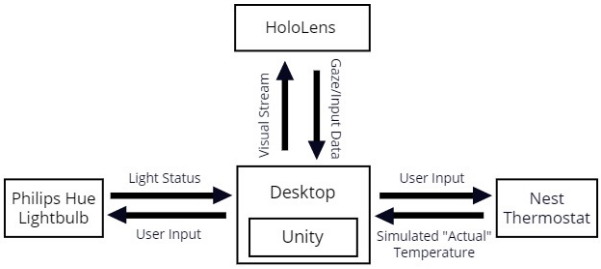

Fig. 1: System architecture for data collection system

Various methods have been proposed to collect sensor and environment data in order to perform activity recognition and prediction. Early work in the computer vision community has used video samples of various activities to train activity recognition algorithms [14]. Further developments in the pervasive computing community have explored using on-body sensors to collect sensor and environment data needed to perform activity recognition, and using RFID tags on objects to recognize what parts of the environment a user utilizes [15]. Recently, attempts have been made to create unsupervised approaches to activity recognition, in which systems can be trained without human labeling [15].

III. Immersive Smart Spaces through AR

This section focuses on the setup of the environment and methods used to collect activity-related data in an augmented reality (AR) smart home. Fig. 1 illustrates the basic architecture of our framework which is designed to be modular so that additional devices can easily be incorporated. Our application uses the Unity 3D graphics platform to render the scene in real-time based on head orientation and position information from the HoloLens. The HoloLens' front facing camera also captures the user's hand gestures which are used to update the rendering of the AR environment and control specific physical devices.

For our simulated home environment, we attempted to create an experience that was as close to a real-life smart environment as possible while considering the limitations of the HoloLens. Our environment is comprised of the following physical and virtual objects:

- physical objects
  - counter surface
  - microwave and bags of popcorn
  - lamp with Philips Hue bulb
- virtual objects
  - microwave UI superimposed on the physical microwave
  - stove top
  - uncooked steak, pan, and plate
  - cupboard
  - television
  - window
  - Nest thermostat linked to a real Nest Home account

– virtual light switch that controls the physical Philips Hue bulb
– control panel allowing user to see the status of the window, stove, temperature, etc.

Views of this simulation can be seen in Figures 3, 4, and 5. For grasping virtual objects, the HoloLens' built-in air tap, clicking motion (as illustrated in Figure 2) was utilized. This click-and-drag motion, which follows the user's hand motion rather than head orientation, added a more natural element to object interaction. For objects that are typically static, such as the television and thermostat, we added other interaction mechanisms, including the ability to turn on the television and adjust the room temperature using the Nest thermostat. The cupboard doors and window could also be opened and closed using a single tap rather than the dragging motion used for movable objects. As the HoloLens does not provide gaze information, we use the center of the HoloLens's field of view as an approximation for the user's area of focus. This approach is in line with the work of other researchers which has demonstrated that head pose and orientation can be used for estimating visual attention [16].

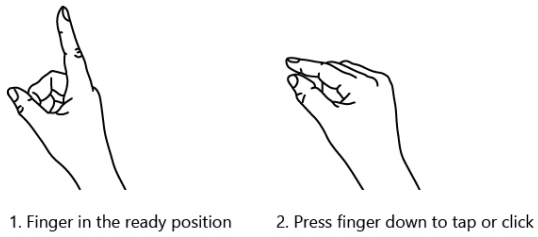

Fig. 2: Microsoft HoloLens's Air Tap gesture

With our current setup, data from the HoloLens' camera and sensors along with data from the other devices in the environment are streamed over a dedicated network to a desktop server where it is processed. The user's head orientation is computed from this data and the Unity application sends updated graphics back to the HoloLens in real-time. The data from all devices is also synchronized and logged at a rate of 2Hz for further analysis. In the future, for increased portability, it would be more efficient to deploy an application containing our smart home environment directly from the HoloLens.

IV. User Study

We recruited a total of 30 participants in our user study. The majority of our participants were college students, with 73.3% under the age of 24 and 26.7% between ages 25 and 35. 46.7% of participants had a high school diploma, 20% had a bachelors degree, 30% had a masters degree, and one participant had no degree. 46.7% of participants identified as female and 53.3% as male. 63.3% of our participants rated themselves as somewhat knowledgeable about technology and 23.3% rated themselves as very knowledgeable, and no participants stated that they were mostly unfamiliar with or that they had no experience with technology. Half of our participants had no

Fig. 3: View of smart home from AR perspective

smart home devices and 13.4% said that they had little to no knowledge of smart devices, but 40% stated that they were somewhat knowledgeable and 13.3% stated that they were very knowledgeable with smart home devices.

Participants were first asked to read and sign a consent form. They were then given a pre-experiment questionnaire asking for demographic information and participant's experience with and knowledge of general technology, AR, and smart home devices. On completion, we set them up in our test environment, taught them the "select" hand gesture, and instructed them to experiment in the environment until they felt comfortable with the HoloLens' various functions.

The main body of the experiment involved immersing the participants in our simulated kitchen environment. They were briefed on the functionality of all of the virtual objects so that data collection wouldn't have to be interrupted mid-activity to answer questions, and then asked to act out the completion of two tasks. The first task was cooking a virtual steak, which involved turning on a burner, placing a pan on the burner, placing the steak on top of the pan, and waiting for ten seconds. The second task was preparing popcorn, which involved putting a cardboard popcorn box into the physical microwave and then virtually clicking the "popcorn" button to trigger a 45-second timer. When the popcorn was finished the microwave would beep, and the subject taking the popcorn box out of the microwave would signify the end of that task. Participants were instructed to use any device in the room to go about the task as they would normally at home, and we tried not to suggest that they perform any specific sequence of actions as not to influence their behavior. After the two tasks were completed, we instructed participants to explore their space if they wanted to before ending the simulation.

Next, we administered a post-experiment questionnaire, which included three questions on a Likert Scale asking about the user's in-activity experience including how helpful they thought the objects were, whether or not their opinions of smart devices had changed, and whether or not they would be more likely to purchase and use a smart device. These

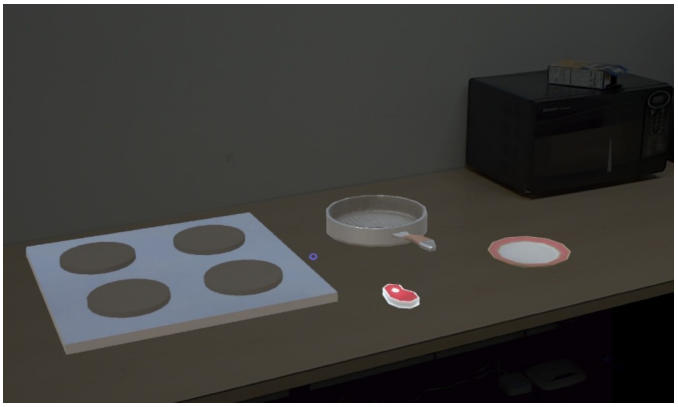

Fig. 4: Another view of smart home from AR perspective

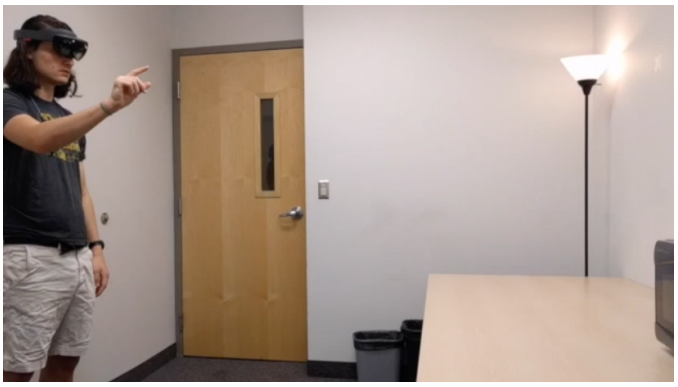

Fig. 5: View of user with HoloLens from outside perspective

were followed by two short-answer sections asking the user to suggest possible applications for our framework and to provide any additional comments about the technology (AR and smart devices) used in the experiment.

V. Evaluation

A. Classification results

Because of the current simplicity of our model, we found it best to use a hidden Markov model (HMM) to evaluate our data. We started with two states representing the task directives, cooking steak and microwaving popcorn, and then added a third exploration state to account for any time not spent directly on the two tasks.

While our dataset was large enough to be significant, it was also too small to create multiple dedicated testing and training sets. To generate more significant results with the data we had, we used 10-fold cross-validation, splitting up our data into 10 sets and randomly resampling the data to evaluate our model.

For testing purposes, we hand-labelled all of our data, which would be impossible for larger datasets, but this made quantifying the accuracy of our model much simpler. In future experiments, it would be more beneficial for hidden states to be labeled during the experiment by the researcher. It should also be noted that the data collected from objects in the room, such as the temperature, stove status, and microwave status, is not collected as continuous data, but rather as discrete sensor events as triggered by the user. So, to reflect the discrete nature of the data, we have referred to the dataset collected from these objects not as object states, but as object events. We hoped that the sensor data collected from the environment and headset would generate data natural enough to compare with data collected from completely physical smart spaces when put through activity recognition algorithms.

The HMM trained and tested on object event data performed relatively well, with an average of 66% precision, referring to the ability of the model to identify only relevant activities, and an average of 68% recall, meaning that it correctly found 68% of relevant activities. This HMM tended to choose the exploring task at times when the participant was making popcorn or steak, often classifying tasks not related to exploring as exploration tasks. This resulted in a precision of 74% for exploration as seen in Table I. The event HMM's preference for the exploring task can be explained by the user's tendency to use devices in the room not as tools to help them cook or make popcorn, but out of curiosity. We expect this issue to become less prominent as familiarity with the environment increases and as improvements are made to the visual realism of the virtual objects.

The HMM trained and tested on attention data performed, in general, worse than the object event model, with an extremely strong bias toward the cooking steak task. The model we tested our attention data on has an average of 42% precision, and 55% recall, as seen in Table III. Precision for the cooking steak task for this model was 93% while precision for the microwaving popcorn and exploring tasks were only 13% and 20% respectively, as shown in Table II. The model's preference towards the steak task can be explained by the active time it takes to cook the steak. During the popcorn task the participant can turn on the microwave, look elsewhere for 45 seconds, and then look back at the microwave very briefly; however, the steak task required the participant's more constant attention. Since our model has only 3 hidden states, this effect is much more exaggerated, as the task of cooking steak accounts for 1/3 of possible hidden states.

Fig. 6 shows examples of machine-labeled data compared to our baseline human annotated data for three participants. The Figures illustrate the accuracy of both the attention-trained and event-trained HMMs.

TABLE I: Precision and recall of results from event data

	Precision	**Recall**
Cooking Steak	.52	.59
Microwaving Popcorn	.73	.76
Exploring	.74	.69

TABLE II: Precision and recall of results from attention data

	Precision	**Recall**
Cooking Steak	.93	.52
Microwaving Popcorn	.13	.53
Exploring	.20	.61

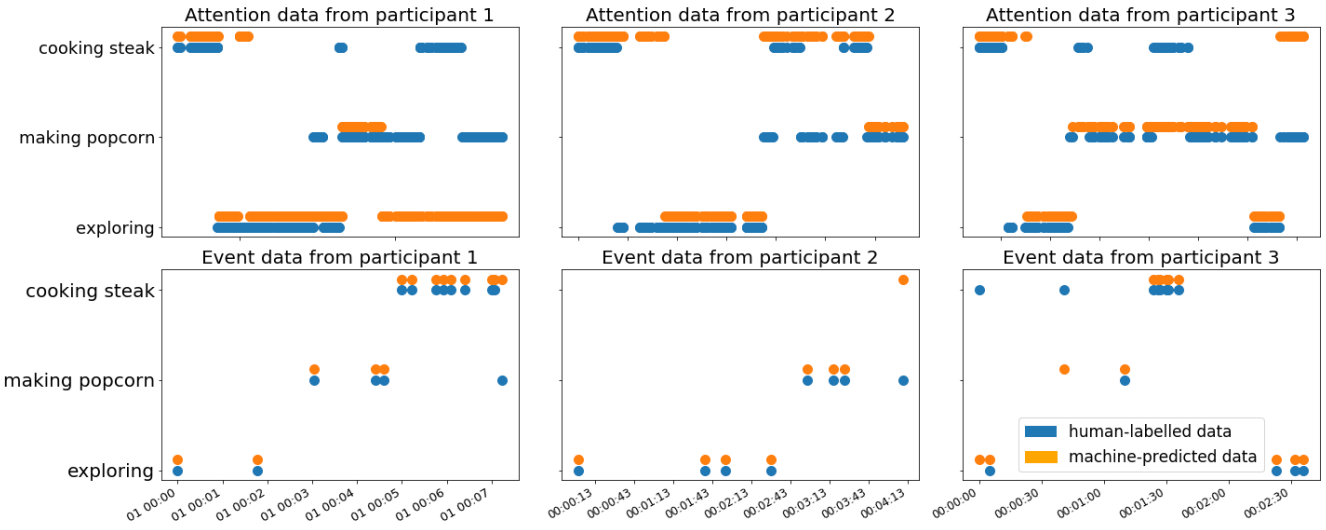

Fig. 6: Inferred user attention data and device's event data of three participants.

TABLE III: Averaged precision and recall from event and attention data

	Precision	Recall
Attention	.42	.55
Events	.66	.68

B. *User feedback analysis*

Questionnaires that participants filled out before and after were concatenated, and statistical analysis and Chi-squared tests were done on various responses throughout the combined form using IBM SPSS. Of our 30 participants, our demographic information shows that:

- 16 were male and 14 were female
- 76% were under age 24, and the remaining participants were ages 25-35
- 83.3% were students, 10% had full-time jobs, and 6.7% had part-time jobs
- 46.7% had high school diplomas, 3.3% had no degree, 20% were at the Bachelors level, and 30% were at the Masters level

According to Chi-squared tests, although there is not a wide range of ages included in our study, we found no significant differences between the gender or age of participants and their familiarity with smart home devices.

Most of our participants thought the space and the devices in it were helpful to them when it came to doing the two household tasks, with 53.3% reporting that they were somewhat helpful, and 16.7% reporting that they were very helpful. In response to the question "As a result of their inclusion and your time with them, are you any more interested in these technologies and their use around the home?", 50% responded that they were very interested, and 40% responded that they were somewhat interested. According to Chi-squared tests, we found no significant differences between participants' familiarity with smart home devices, and how helpful they viewed the smart home devices in the space to be.

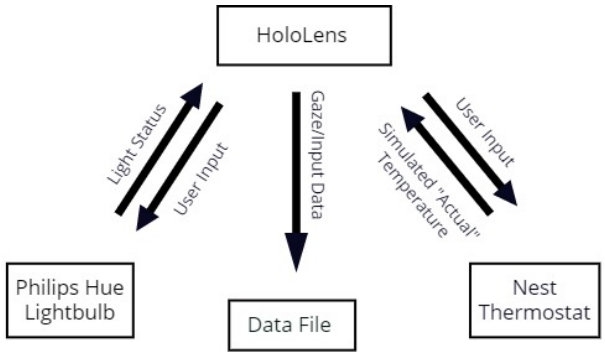

Fig. 7: Proposed experiment setup for data collection

VI. Lessons Learned

The implementation of our activity recognition models show that unimodal device state analysis is more accurate than unimodal head orientation/attention analysis with data provided by the HoloLens, but as head orientation data is collected continuously, it can be more useful for detecting user intent when augmenting object events that occur at widely-spaced intervals. While neither method has proven to be very accurate or precise, this can be remedied by implementing a better activity recognition system that combines both head orientation and device state modes.

In the future, for portability, it would be more efficient to deploy an application containing our smart home environment directly from the HoloLens, as shown in Figure 7. This would allow the environment to run without directly communicating with a computer running Unity, and increase the ease of setup of our smart environment even more.

Works utilizing the framework of our project would output more sophisticated data if a larger variety of tasks were to

be done in the space, and users could go about tasks in the order they pleased. We believe that including more than our 2 existing tasks will make the collected data better aligned with real-life use, and will increase the significance of what our hidden Markov models predict. Including more smart devices and virtual objects would help accomplish this by allowing users to complete more every-day tasks, reducing the bias towards one task that has been shown in our activity recognition analysis. Future works could also improve the congruence between different tasks done in the space, so that the vast majority of tactile tasks are based on physical rather than virtual objects. The tasks done in this study were half physical and half virtual, and we believe that user comfort will improve if virtual interactions are limited to interactions with smart devices, keeping the direct tasks at hand to be physical. We would also want to allow the user to move objects that are interactable but confined to a plane of motion, such as the cupboard doors and window, through a click-and-drag motion, rather than with a single click. Putting the user in a more realistic space would also reduce the level of exploration, making the flow between different tasks more realistic. For example, the thermostat placed in our environment didn't actually affect the room's temperature and the light didn't significantly increase the brightness of the room. As a result, most participants utilized these objects not to aid in their tasks but out of curiosity, making our data less realistic. Because of this, we should attempt to make interactions with any object in our smart home environment as realistic and connected with reality as possible, while still minimizing the time and expense of setting up physical smart home devices. Future works should also seek to collect a much larger and more diverse dataset, as our 30 participant dataset was too small to create dedicated training and testing sets. All of our participants were 35 or younger and the majority were college students, making it impossible to draw conclusions on how different age groups or people with differing occupations would react to the space, so future studies determining the ease of use and navigation of this type of space with those populations would be insightful.

VII. Conclusion

In this study we developed a mixed reality environment that we used to represent an activity recognition framework that attempts to report what task a user is doing in our virtual space. Initial activity recognition models that we have been able to develop with participant data gathered from the space show that we are able to recognize tasks done in the space from event and head orientation data to a degree comparable to models based on physical smart home data, and in fact, head orientation collected from headset sensors, while not as accurate as object state information, provides a faster mechanism for recognizing the user's activity. In addition, most users noted that they were not hindered by the unfamiliarity of mixed reality technology that fueled this experience, as our questionnaire analysis data shows. Our smart home environment successfully implements a more portable smart environment, and has the potential to collect data that is as useful as physical smart home data, with the addition of attention data taken directly from headset sensors.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Award No. IIS-1559889. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. Weiser, "The computer for the 21st century," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 3, no. 3, pp. 3–11, Jul. 1999. [Online]. Available: http://doi.acm.org/10.1145/329124.329126

[2] T. Van Kasteren, A. Noulas, G. Englebienne, and B. Kröse, "Accurate activity recognition in a home setting," in *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 2008, pp. 1–9.

[3] B. Minor, J. R. Doppa, and D. J. Cook, "Data-driven activity prediction," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 15*, p. 805814, Aug 2015.

[4] B. Minor and D. J. Cook, "Forecasting occurrences of activities," *Pervasive and Mobile Computing*, vol. 38, p. 7791, 2017.

[5] D. Cook, M. Schmitter-Edgecombe, A. Crandall, C. Sanders, and B. Thomas, "Collecting and disseminating smart home sensor data in the casas project," *Proceedings of the CHI workshop on developing shared home behavior datasets to advance HCI and ubiquitous computing research*, pp. 1–7, April 2009.

[6] N. Alshammari, T. Alshammari, M. Sedky, J. Champion, and C. Bauer, "Openshs: Open smart home simulator," *Sensors*, vol. 17, no. 5, p. 1003, Feb 2017.

[7] K. Bouchard, A. Ajroud, B. Bouchard, and A. Bouzouane, "Simact: A 3d open source smart home simulator for activity recognition," *Advances in Computer Science and Information Technology Lecture Notes in Computer Science*, p. 524533, 2010.

[8] D. Surie, F. Lagriffoul, T. Pederson, and D. Sjlie, "Activity recognition based on intra and extra manipulation of everyday objects," *International Symposium on Ubiquitous Computing Systems*, p. 196210, Nov 2007.

[9] A. M. Qamar, A. R. Khan, S. O. Husain, M. A. Rahman, and S. Baslamah, "A multi-sensory gesture-based occupational therapy environment for controlling home appliances," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 671–674.

[10] L. Borodulkin, H. Ruser, and H.-R. Trankler, "3d virtual" smart home" user interface," in *Virtual and Intelligent Measurement Systems, 2002. VIMS'02. 2002 IEEE International Symposium on*. IEEE, 2002, pp. 111–115.

[11] D. W. Seo, H. Kim, J. S. Kim, and J. Y. Lee, "Hybrid reality-based user experience and evaluation of a context-aware smart home," *Computers in Industry*, vol. 76, pp. 11–23, 2016.

[12] N. Rhinehart and K. M. Kitani, "First-person activity forecasting with online inverse reinforcement learning," *2017 IEEE International Conference on Computer Vision (ICCV)*, p. 36963705, Oct 2017.

[13] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," *2007 IEEE 11th International Conference on Computer Vision*, p. 18, Oct 2007.

[14] Y. Yacoob and M. Black, "Parameterized modeling and recognition of activities," *Computer Vision and Image Understanding*, vol. 73, no. 2, p. 232247, 1999.

[15] T. Gu, S. Chen, X. Tao, and J. Lu, "An unsupervised approach to activity recognition and segmentation based on object-use fingerprints," *Data Knowledge Engineering*, vol. 69, no. 6, p. 533544, 2010.

[16] C. Segura, C. Canton-Ferrer, A. Abad, J. R. Casas, and J. Hernando, "Multimodal head orientation towards attention tracking in smartrooms," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 2, April 2007, pp. II–681–II–684.