16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Human-Object Maps for Daily Activity Recognition

Haruya Ishikawa, Yuchi Ishikawa, Shuichi Akizuki, and Yoshimitsu Aoki
{hishikawa, yishikawa} @aoki-medialab.jp, {akizuki, aoki} @elec.keio.ac.jp
Keio University
Kanagawa, Japan

Abstract

In the field of action recognition, when and where an interaction between a human and an object happens has the potential to be valid information in enhancing action recognition accuracy. Especially, in daily life where each activities are performed in longer time frame, conventional short term action recognition may fail to generalize do to the variety of shorter actions that could take place during the activity. In this paper, we propose a novel representation of human object interaction called Human-Object Maps (HOMs) for recognition of long term daily activities. HOMs are 2D probability maps that represents spatio-temporal information of human object interaction in a given scene. We analyzed the effectiveness of HOMs as well as features relating to the time of the day in daily activity recognition. Since there are no publicly available daily activity dataset that depicts daily routines needed for our task, we have created a new dataset that contains long term activities. Using this dataset, we confirm that our method enhances the prediction accuracy of the conventional 3D ResNeXt action recognition method from 86.31% to 97.89%.

1 Introduction

How can humans and robots coexist daily in an environment? We have seen the upwards trend of smart home technologies during the past decade ranging from simple IoT light bulbs to mobile vacuum machines [1]. This popularity is predicted to increase in the following years with more devices and robotic assistants being developed. For smart home technologies to become better integrated in our daily lives, these devices should be able to learn from our daily routines and provide services to us accordingly. The question arises in *how* to make these machines learn from our daily routines.

Daily routines can be thought as activities that are often times executed around a certain time frame and may require the person to be at a certain location and interacting with specific objects. Each activities are not short, but usually encompass several minutes or even a few hours. For example, an employee may have a daily routine of getting to office at 9am, cleaning up the work space until 10am, working until noon, and so on. There are many activities a person may do each day and these activities could vary depending on different occasions, which makes tracking and recognizing daily activities a burdensome task. Also, how one might perform an activity may vary from person to person, introducing another challenge of how to generalize an activity. Due to these technical difficulties, there have not been many academic studies done in the area of understanding daily routines. However, a person may exhibit a specific behavior or use a particular object in each of the scenarios that may become a hint to recognizing those long term activities.

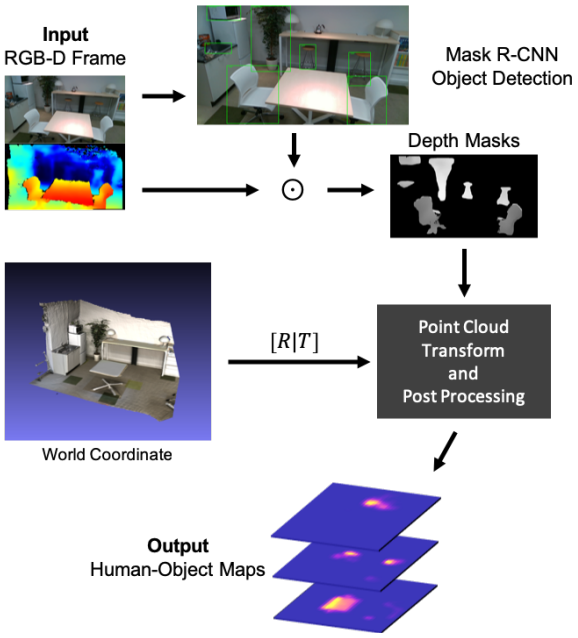


Figure 1: An overview of creating Human-Object Maps. Given a RGB-D image, depth masks of objects are estimated using Mask R-CNN and taking a Hadamard product between the depth frame. The depth masks are converted into point cloud data and is transposed to world coordinate. Finally, the point clouds are post-processed to create a Human-Object Map.

In recent years, there has been significant works concentrated on computer vision algorithms which can predict very short human actions that on the order of a few seconds. In contrast, we focus on predicting daily activities that span over a longer period of time. We feel that these emerging action recognition techniques

could be used for longer activity recognition, but often times, these activities tend to include several smaller sub-actions which may hinder the prediction accuracy. Therefore, there is a need to represent daily routines into features that will enhance the recognition longer activities.

Human and object interaction has the potential to be valid information to recognize long term activities. In long term activities, it might be better to incorporate the information about how human and objects interacted rather than just using appearance information. This is because, due to having various sub-actions in long term activities, using interaction information may generalize well to the noise caused by appearance information.

In our work, we tackle the problem of recognizing daily activities in a single room. In addition to performing activity recognition on RGB video clips using conventional action recognition method, we introduce the use of human and object probability maps called Human-Object Maps (HOMs) for this prediction task. HOMs contains useful features about where someone might have been during the activity, as well as what objects they interacted with. These features also shows how human or objects moved in the given video data, portraying spatio-temporal information. Also, we have used time of the day feature for our prediction task since time information is readily available with the video clips. For our analysis, we compared the baseline method of using only video clips with our method of using HOMs and time features with the video clips. Since there are no publicly available daily routine datasets, we have created a dataset which contains annotated RGB-D videos of daily activities captured from a fixed camera. We validate our approach by showing that the use of HOMs and time features with video clips increases prediction accuracy. We further evaluated the effects of using HOMs and time features by comparing predictions that only used HOMs and predictions that used video clips with time features. Our work have made the following contributions:

- Proposed a novel approach of representing human and object interaction using probability maps called Human-Object Maps (HOMs) for representing daily routines.
- Introduced activity recognition algorithm that used encoded HOMs features as well as time in addition to conventional action recognition algorithm.
- Created a daily activity dataset which contains real life video clips of long term activities and validated our approach of using HOMs and time features using this dataset. Improvements are seen by using our proposed method.

2 Related Works

2.1 Action Recognition

Recently, we have seen ongoing improvements in the field of action recognition including recognition of daily activities. Most of these emerging techniques incorporates machine learning to model a wide range of human activities [2, 3, 4, 5]. To learn these models, it requires sufficient amounts of annotated data, which is often difficult to create from scratch. There are publicly available large datasets such as the ActivityNet and Kinetics dataset [6, 7], but in our work, these datasets are not used due to the fact that each action in these datasets are video clips that span around a few seconds. Thus, we created our own set of dataset for daily life activities containing video clips that target activities that are done in longer time spans.

Although, many vision based advancements were made, vision alone may not be sufficient since it is unclear whether the robot will recognize the activity based on human behavior, but recognize it based on the background [8]. In recent years, there has been strides in incorporating vision data with other modals such as location and time [2, 9, 10]. In our work, we feel the use of location and time is critical in recognition of long term activities, and have incorporated these attributes.

S. Bokhari et al. [11] have conducted studies using activities that are more long term for activity forecasting. The target activities are similar to our work, but is focused on human trajectory forecasting rather than prediction task.

2.2 Action Map

A few novel works have been done on the representation of human actions called *Action Maps* [12, 13]. These maps are 2D (or 3D) probability maps which contains information on where an action did or could occur. Map representation seems very useful in the application of mobile robots since these maps contain location information. In our work, we followed this mapping technique, but instead of mapping the actions, we mapped the probability of human and object positions.

2.3 Human Object Interaction

Recent works in human object interaction (HOI) seem promising for understanding activities based on how human interacts with the surrounding environments [14, 15, 16]. For example, G. Gkioxari et al. introduced a method that detects <human, verb, object> triplets from image using only appearance features [14]. We feel it is possible to directly infer the long term actions by keeping track of these smaller human and object interaction, but it would be cumbersome to track wide range of small interactions. Therefore, in

our work, we took a step back and mapped 2D locations of humans and objects using HOMs. With the use of HOMs, we feel that neural networks would generalize and recognize key interactions between human and object and even between object and object.

3 Approach

3.1 Overview

The goal of this work is to predict daily routine activities. In a conventional action recognition problem setting, given RGB video frames, our task is to predict which action was portrayed throughout those frames. In our work, we tackle the same problem with long term daily activities using HOMs and time features using a dataset we have created for this task, which is explained in more details in section 4.1. Samples of what activities we focused on recognizing is shown in Fig. 4.

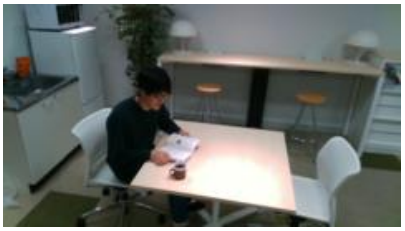
In this section, we will explain our approach for this prediction task. In section 3.2, we will explain HOMs in detail including how they are created. In section 3.3, we will explain the network architectures for the proposed method along with baseline and comparison methods.

3.2 Human-Object Maps

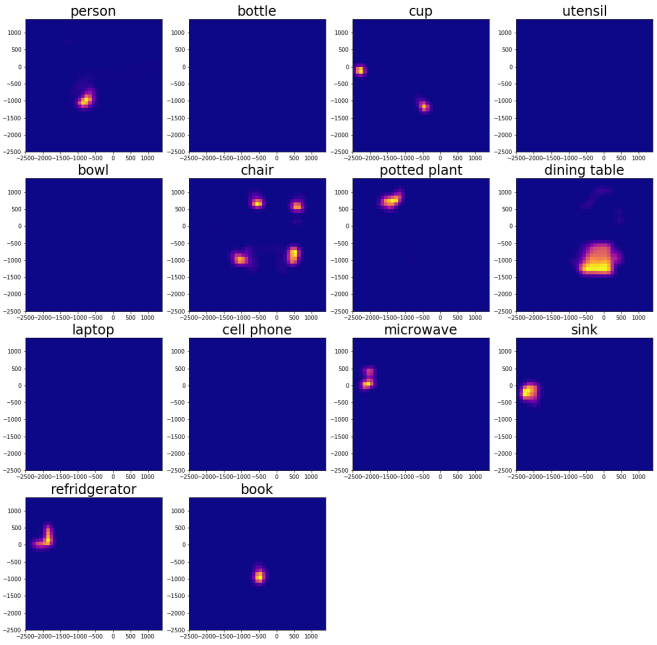
Human-Object Maps (HOMs) are probability maps of human and objects created from RGB-D video frames. These maps represent spatio-temporal information about humans and objects as well as the interaction between them. As shown in Fig. 2b, there are total of 14 maps, which the labels are: Person, Bottle, Cup, Utensil, Bowl, Chair, Potted Plant, Dining Table, Laptop, Cell Phone, Microwave, Sink, Refrigerator, and Book.

For each RGB and depth frame, we executed the following steps, which is also shown in Fig. 1:

1. From the RGB image, we use pretrained Mask R-CNN for object detection [17].
2. For each of the detected object masks, we take a Hadamard product between the depth image and the mask and convert the masks into point cloud representation in the camera coordinate.
3. Then the point clouds are converted to world coordinate using transform matrix which were calculated beforehand.
4. Those point clouds are then reduced into a 2D map by projecting the points looking down from the ceiling.
5. The points in the 2D maps are converted into a 2D histogram of small $10\times10[\text{cm}^2]$ grids (as shown in Fig. 2b).



(a) Sample frame.



(b) Human-Object Maps.

Figure 2: An example of a RGB frame and Human-Object Maps of 'Coffee Break' is shown in (a) and (b) respectively. Using these maps, human-object interaction information can be easily obtained, such as the relationship between chairs and human which can represent 'sitting'.

6. Finally, by taking the averages, the 2D histogram is converted into 39×39 pixel image.

Given the number of frames, we took running averages of the HOMs.

Since there are heatmaps for every objects, the number of objects to track must be carefully selected before. If more objects are tracked, the more computation resources for map creation must be set aside.

3.3 Network Model for Activity Recognition

An overview of the network model we have used is shown in Fig. 3. In our proposed method, we use RGB-D video frames as input. The RGB frames are used as the input frames for 3D ResNeXt feature extraction module. The 3D ResNeXt feature extraction

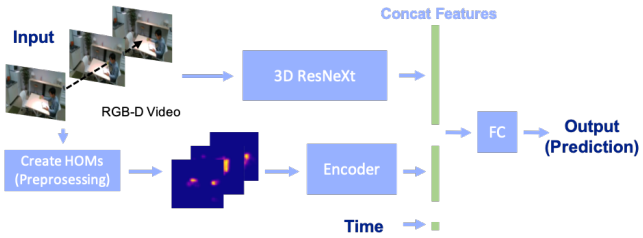


Figure 3: An overview of our network model. 3D ResNeXt takes RGB video frames as input and extracts viable features. The HOM encoder network outputs feature vectors from created HOMs. The feature vectors from all of the modules, including time feature, are concatenated into a single feature vector which is passed though the fully connected layers to output a prediction.

module has a depth of 101 and was pretrained using action recognition datasets (ActivityNet and Kinetics [6, 7]). The motivation for using 3D ResNeXt as our feature extraction network was that in recent works, this network architecture has the highest accuracy in action recognition methods that use only RGB frames as inputs [5]. We also use the RGB and depth frame to create HOMs, which are then encoded into feature vector using 3 layers of convolutional neural networks. As for the time of the day feature, we convert time, which is given as ISO format, into continuous value normalized between 0 and 1. The feature vectors from 3D ResNeXt, HOM encoder, and time conversion are then concatenated into one feature vector and is passed though several fully connected layers (FCs) to obtain our prediction.

We turned off the HOM encoder and time conversion for the baseline method of only using video RGB frames. For our comparison methods of using video frames with HOMs, using video frames with time, and using only HOMs, we have turned off the network respectively and only concatenated the needed features.

4 Experiments

4.1 Dataset

For our work, since there are no publicly available datasets for the task of recognizing daily routine activities, we have created our own dataset to evaluate our approach. We used a Intel® RealSense™ D415 sensor to take RGB-D video frames of the daily activities [18]. The activities we have tracked and the number of clips as well as the number of total frames are given in Table. 1. As for our method of gathering data, we took video clips of daily life in our lab from 9:00 to 20:00. Afterwards, we selected several segments from the video clips and annotated what activity the clip portrays and the time of which the activity occurred.

Table 1: Summary of the daily activity dataset. For each Activity, we show the number of clips and the total number of frames.

Activity	Clips	Frames
Coffee Break	15	9,752
Cooking	11	7,857
Meal Time	15	9,954
Meeting	19	12,511
Nap	8	5,687
Tending to Plants	8	5,501
Working	13	9,204
Total Clips:		89
Total Frames:		60,466

Table 2: Activity prediction accuracy.

	Accuracy
Video Only (3D ResNeXt [5])	86.31%
Video with HOMs	86.31%
Video with Time	89.49%
HOMs Only	91.58%
Video with HOMs and Time	**97.89%**

A time-line of the dataset with sample frames of each activity is shown in Fig. 4.

For evaluating our method, we split the dataset into training and test sets which contain 68 and 21 clips respectively.

4.2 Recognition Performance

As stated before in section 3.3, we have compared the following methods:

- Only RGB video frames (we call this method 'video only')
- RGB video frames with HOMs and time features
- RGB video frames with HOMs features
- RGB video frames with time features
- Only HOMs features

For training, we trained each method separately until convergence using the training split in our dataset. We have used stochastic gradient decent as our optimizer and learning rate of 0.01 which is annealed using a learning rate scheduler. Most of the training parameters are the same as training the 3D ResNeXt [5]. Out of the 68 training data, we have used image augmentation for training. The input for 3D ResNeXt is

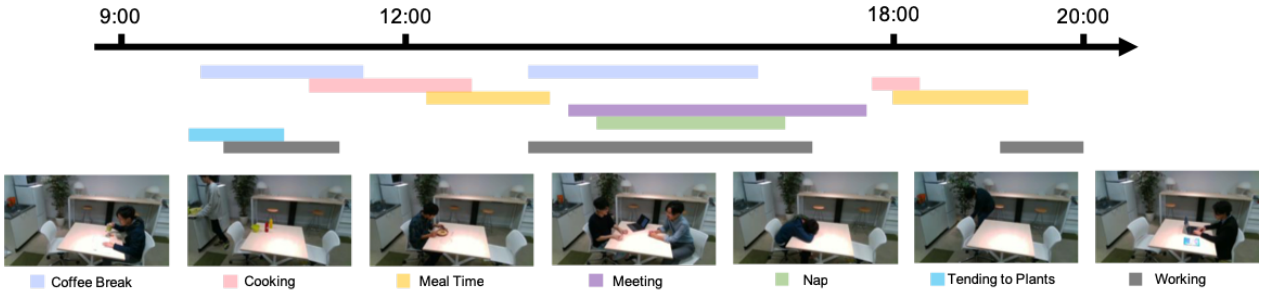


Figure 4: A time-line of when the activities in our dataset occurred with sample frames of each of the activity.

16 frames, thus we sampled 16 frames using temporal augmentation. The HOM images are created using the same 16 frames.

For evaluation of our method, we split the testing split into even smaller test dataset and calculated the classification accuracy based on evaluation metrics used in action recognition [5]. Even though it is a smaller test dataset, the frames are selected so that it encompasses the whole video clip. We have not added any image augmentation, but have selected every M frames instead of random sampling. For example, if there are total of N frames in a test clip, we would sample, 1st frame, then $N/16$th frame, $2N/16$th frame, and so on until we obtain 16 frames. The prediction results is shown in Table. 2.

We have also included confusion matrices of the comparison methods in Fig. 5 for analysis of the effects of the proposed features.

5 Discussion

By taking a look at the results in section 4.2, our proposed method has higher prediction accuracy than the baseline method of using only using 3D ResNeXt by around 11%. Even more, our proposed method surpass all of the comparison methods as well. This proves our presumption of the use of location information as well as time enhances the action recognition accuracy. Taking a look at 5e, the model failed to recognize the activity, 'tending to plants', and mistook it for 'cooking', which was the only failure it had made. The same mistakes were made with the other comparison methods, which means the sample was very hard to recognize.

However, to our surprise, even though the method that used only HOM features surpassed the baseline, the method that used video with HOM features had the same accuracy as the baseline method. Despite having visual and location information to better understand what is happening in a scene, the opposite can be seen by taking a look at Fig. 5a and Fig. 5b. For example, 'Cooking' and 'Tending to Plants' shows very close visual features due to the fact that a person doing the activity are usually around the sink or the potted plant (shown in sample images in Fig. 4), which is a similar trend that can be seen in HOMs as well. Even though in video only and HOMs only predictions those methods could classify the two as different activities, when the two methods are combined, the visual location and HOM's 2D location feature may have been strengthened, which results in high prediction fail rates in activities that relates to location information instead of understanding human-object interaction.

From Fig. 5d, we can also conclude that time alone is not as effective in increasing the accuracy. This can be explained from Fig. 4 in how much each activities overlap with each other.

6 Conclusion

We have proposed a novel representation of human and object interaction using probability maps called Human-Object Maps for the use of recognizing daily routines. These maps are location based probability maps, thus for our next step, we would like to investigate the use of these maps for robot navigation and human computer interactions. In addition to the new representation, we have used HOMs as features for enhancing conventional action recognition algorithm for predicting long term daily activities. To validate our proposed method, we have proposed a new daily activity dataset, which consists of real human data in long term activities. Since our goal for this research was to recognize action in a particular scene, in our future work, we will enlarge the dataset to include more variety of activity classes in variety of scenes.

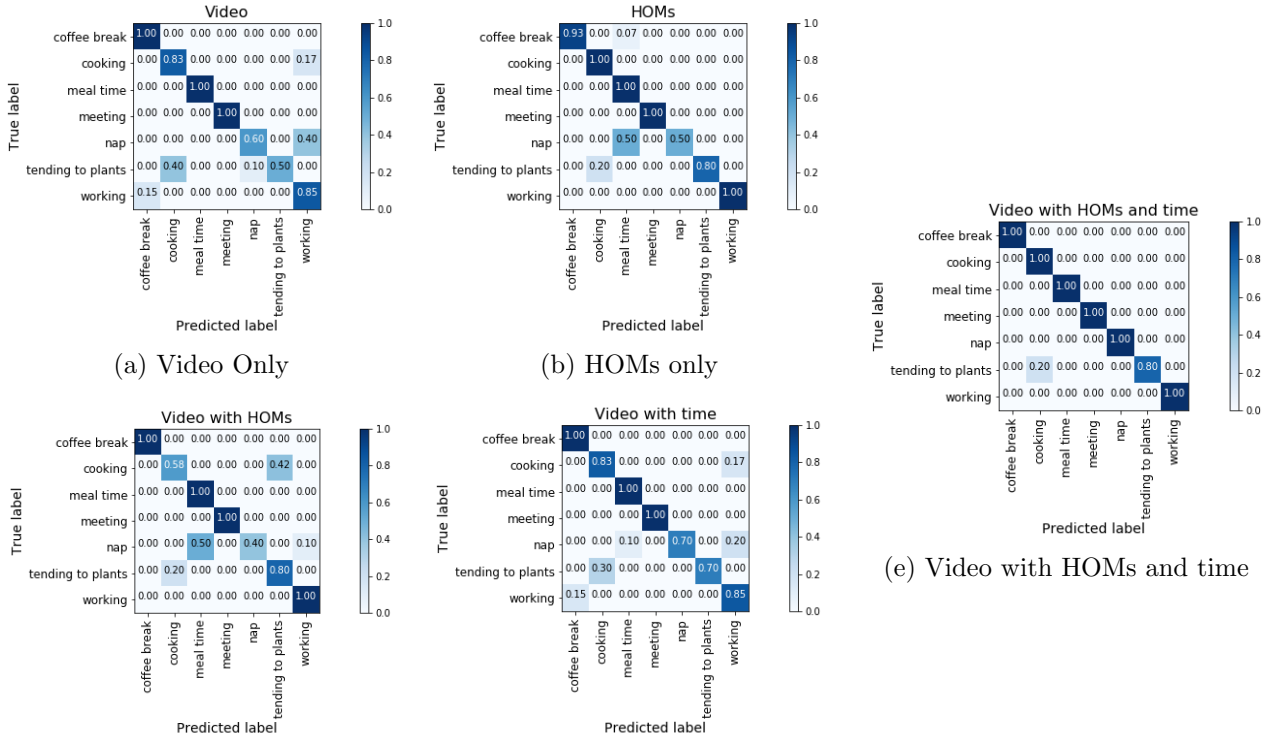
Acknowledgement

We are grateful to the researchers at Honda R&D Co., Ltd. for the extensive discussions for making this research possible.

References

[1] The Economist Group Limited: "Where the smart is," *The Economist*, Jun. 2016.

[2] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa: "Predicting



(a) Video Only

(b) HOMs only

(e) Video with HOMs and time

(c) Video with HOMs

(d) Video with time

Figure 5: Confusion matrices of the comparison methods.

Daily Activities From Egocentric Images Using Deep Learning," ISWC, 2015.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei: "Large-scale Video Classification with Convolutional Neural Networks," CVPR, 2014.

[4] M. Ma, H. Fan, and K. Kitani: "Going Deeper into First-Person Activity Recognition," CVPR, 2016.

[5] K. Hara, H. Kataoka, and Y. Satoh: "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," CVPR, 2018.

[6] F. Heilbron, V. Escorcia, B. Ghanem, and J. Niebles: "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding," CVPR, 2015.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman: "The Kinetics Human Action Video Dataset," arXiv, 2017.

[8] Y. He, S. Shirakabe, Y. Satoh, and H. Kataoka,: "Human Action Recognition without Human," CoRP, 2016.

[9] L. Liao, D. Fox, and H. Kautz: "Location-based Activity Recognition," NIPS, 2005.

[10] M. Williams, J. Burry, and A. Rao: "Understanding social behaviors in the indoor environment: a complex network approach," ACADIA, 2014.

[11] S. Bokhari and K. Kitani: "Long-Term Activity Forecasting using First-Person Vision," CVPR, 2016.

[12] M. Savva, A. Chang, P. Hanrahan, M. Fisher, and M. Niener: "SceneGrok: Inferring Action Maps in 3D Environments," SIGGRAPH, 2014.

[13] N. Rhinehart and K. Kitani: "Learning Action Maps of Large Environments via First-Person Vision," CVPR, 2016.

[14] G. Gkioxari, R. Girshick, P. Dollar, K. He: "Detecting and Recognizing Human-Object Interactions," CVPR, 2018.

[15] V. Delaitre, J Sivic, and I. Laptev: "Learning person-object interactions for action recognition in still images," NIPS, 2011.

[16] K. Kato, Y. Li, and A. Gupta: "Compositional Learning for Human Object Interaction," ECCV, 2018.

[17] K. He, G. Gkioxari, P. Dollr, R. Girshick: "Mask R-CNN," ICCV, 2017.

[18] L. Keselman, J. Woodfill, A. Grunnet-Jepsen, A. Bhowmik: "Intel RealSense Stereoscopic Depth Cameras," CVPR, 2017.