Detection of Generic Human-Object Interactions in Video Streams

Lilli Bruckschen, Sabrina Amft, Julian Tanke, Jürgen Gall, and Maren Bennewitz

Humanoid Robots Lab, University of Bonn, Germany[*]

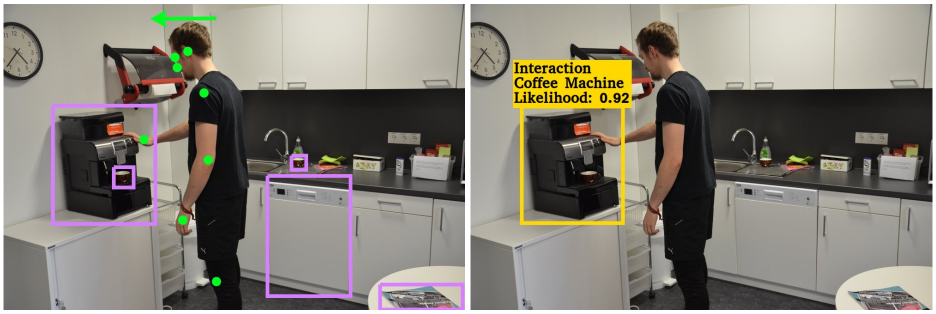
Abstract. The detection of human-object interactions is a key component in many applications, examples include activity recognition, human intention understanding or the prediction of human movements. In this paper, we propose a novel framework to detect such interactions in RGB-D video streams based on spatio-temporal and pose information. Our system first detects possible human-object interactions using position and pose data of humans and objects. To counter false positive and false negative detections, we calculate the likelihood that such an interaction really occurs by tracking it over subsequent frames. Previous work mainly focused on the detection of specific activities with interacted objects in short prerecorded video clips. In contrast to that, our framework is able to find arbitrary interactions with 510 different objects exploiting the detection capabilities of R-CNNs as well as the *Open Image* dataset and can be used on online video streams. Our experimental evaluation demonstrates the robustness of the approach on various published videos recorded in indoor environments. The system achieves precision and recall rates of 0.82 on this dataset. Furthermore, we also show that our system can be used for online human motion prediction in robotic applications.

Keywords: intention understanding, video understanding, domestic robots

1 Introduction

The ability to detect interactions of humans with objects is of great use for a variety of applications, especially for service robots. Examples include the identification of customer browsing patterns in retail scenarios [8], activity recognition based on used objects and monitoring of daily activities [15, 14], and, as we showed in our previous paper, the prediction of human movements based on subsequent object interactions [1]. We now present a system that extracts such interactions from RGB-D streams. Most work regarding interaction detection focuses on well-constrained scenarios often with the goal to identify a small set of potential activities in prerecorded videos [10, 12]. In this paper, we present a novel approach to detect and extract arbitrary human-object interactions from video streams, which is based on spatio-temporal and pose information. To achieve robustness, our framework verifies interactions found in one frame using subsequent observations. In contrast to existing detection approaches, we do not assume

[*] This work has been supported by the DFG Research Unit FOR 2535 Anticipating Human Behavior.



(a) (b)

Fig. 1: Our system detects human-object interactions based on object positions (purple) as well as pose and orientation information of the human (green) (a). To deal with uncertainty in the observations, we then compute for each found human-object interaction the likelihood that this interaction really occurs using previous observations (b). In this example, the human interacts with the coffee machine over several frames, resulting in a high likelihood for this interaction.

specific activities but allow for the detection of arbitrary human-object interactions with 510 different objects from the *Open Image* dataset [7]. Our framework focuses on video streams, but can also be applied to static images or pre-recorded videos. We define a human-object interaction as an action in which a human places at least one hand on an object while facing it, see Fig. 1 for a demonstration. Our method detects relevant objects inside each frame using regional convolutional neural networks (R-CNNs) [3] and estimates humans and their body pose using the *OpenPose* system [2]. We then detect possible interactions based on the pose of the human and spatial information about humans and objects. To deal with uncertainty in the observations, our system computes for each found interaction the likelihood that it really occurs by tracking it over subsequent frames. The output of our framework is the set of all detected human-object interactions with a sufficiently high likelihood. Fig. 1 illustrates the methodology of our approach.

As our approach is constrained by the types of recognizable objects our framework is able to utilize a vast amount of available training data [7]. This allows us to recognize any interaction with objects known by an interchangeable R-CNN. At the time of publication, our system is able to detect interactions with 510 different objects. We will publish the source code of our framework.

As we show in the experimental evaluation our system achieves recall and precision rates of 0.82 with respect to the detection of human-object-interactions. We additionally show the application of our system to predict human motions online and improve existing motion prediction systems [1].

2 Related Work

The detection of interactions between humans and objects is closely intertwined with activity recognition, as the type of the used object is typically associated with an activity.

An interesting work in this context is presented by Prest *et al.* [10]. The goal of the authors is to detect smoking and drinking activities in realistic videos. To accomplish this, Prest *et al.* trained an action classifier on example interactions and use this classifier in combination with a generic, part-based human detector [11] to spot the previously learned interactions in a prerecorded video. The system tracks objects and persons in space and time and uses the action classifier on the tracked data. In contrast to our approach the application domain of this system is limited, as it is only able to detect interactions with cigarettes and glasses. Similarly Yang *et al.* [14] proposed to use object and interaction information to assign a predefined role, in their example *kidnapper* and *hostage*, to a human. To detect human-object interactions, the authors apply depth information and R-CNNs [3] and assume that an object is in use when it is very close to a human in terms of position and depth. As interaction detection is primarily done using position information obtained from an R-CNN, detection errors can easily lead to wrong results. While our framework also uses an R-CNN, we additionally make use of pose and spatio-temporal information to increase the robustness of the detection.

Several other related systems use static images rather than videos for example the work by Yao *et al.* [15]. The authors use the assumption that objects are associated with activities with the goal to increase object detection rates in static scenes by utilizing information about pose and activities of humans. The work of Gupta *et al.* [5] follows a similar idea. The authors propose a Bayesian model that incorporates functional and spatial context for object and action recognition. Another approach that focuses on action detection in static images was presented by Gkioxari *et al.* [4]. The authors detect humans and objects with an R-CNN and estimate action-type specific densities to localize the used object. In most cases, this corresponds to the position of a hand of the human.

In our work we use pose information [15, 5], especially about the hands of the human [4], alongside R-CNNs [3] to detect possible interactions in individual frames. We then apply a verification step in the video stream to deal with false positive detections. We extend the state of the art by allowing arbitrary interactions with known objects, thereby shifting the focus from action recognition to the detection of human-object interactions, allowing the use of a large amount of freely available training data [7]. Several applications can utilize the information provided by our framework ranging from motion prediction, as we show in this work, to intention or activity recognition at a larger scale.

3 Detection of Human-Object Interactions

Our goal is to detect all human-object interactions that occur in a video stream. We define a human-object interaction as an action in which a human places at least one hand on the object while facing it, see Fig. 1 for an illustration.

A video stream is a sequence of frames $V = [f_0, ..., f_t]$ with f_0 as the first observed frame and f_t as the currently observed frame at time t. Our approach uses the current frame f_t and all previously found interactions on $[f_0, ..., f_{t-1}]$ as input and returns all human-object interactions in V.

In summary, our approach to find all human-object interactions inside V works as follows:

1. Apply an R-CNN to detect objects and the *OpenPose* system [2] to detect humans and their poses from RGB data.
2. Use position and depth data to find overlaps between object bounding boxes and human hand positions. Use pose information of the human to check whether they are facing an object that overlaps with their hand, if so, record a possible interaction.
3. Update the likelihood of interactions based on the new observations. This step is necessary to verify that a detected interaction really occurs.

The output of our system are all human-object interactions with a likelihood over a threshold min_L, which is determined using a training data set. A learning process for min_L is shown in our evaluation. An example video demonstrating our approach is shown on our website [1].

3.1 Detection of Objects and Humans and Estimation of the Human Pose

To efficiently detect objects in the current frame we use an R-CNN from Google's object detection API [6], which was trained on the Open Images dataset [7]. Note that the R-CNN is interchangeable and its object detection capabilities can be extended using transfer learning techniques [9] in case new objects need to be detected. For the detection of humans and their poses we apply the *OpenPose* framework [2]. The estimated pose directly contains information about the position of ears, eyes, nose, shoulders, hands, and legs of the human. We further trained an estimator to classify the general direction in which the human is oriented. We use as orientations with respect to the point of view of the camera: *right*, *left*, *back*, and *frontal*. The estimation is based on information about the visibility of the ears, eyes, nose and shoulders. During this step, we used pre-existing systems. In the following, novel approaches are presented.

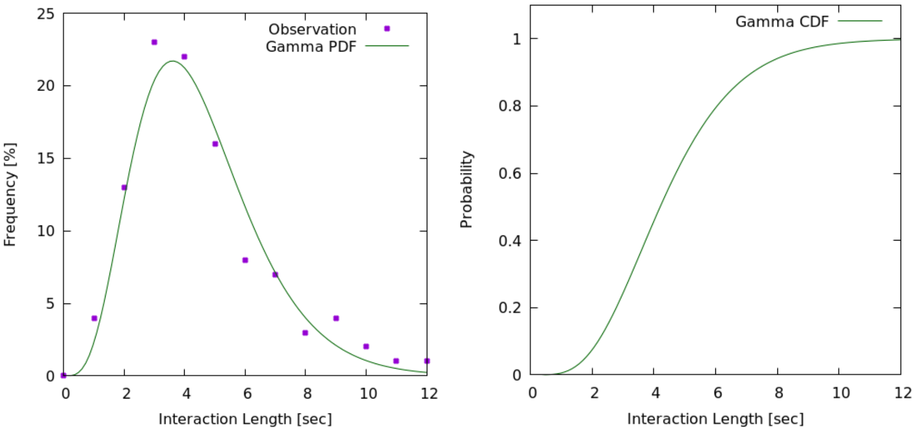
3.2 Detection of Possible Human-Object Interactions

Depending on the results of the orientation estimation we can infer which objects the human is facing based on their x coordinates in the frame and depth levels with respect to the human. In particular, our approach processes each detected object in the current frame and checks whether the following conditions are satisfied for f_t:

- The position of a human hand is inside the bounding box of the object.
- The human is facing the object.
- The depth level of the hand and the object are similar.

If an object fulfills all these conditions, a possible interaction of the human with this object is recorded for f_t.

[1] `https://www.hrl.uni-bonn.de/icsr_interaction_demo.mp4`



(a)

(b)

Fig. 2: Gamma probability density function (green) approximating the observed interaction durations (purple) in our real-world data set (a). Cumulative form of the *gamma probability density function* (b) that indicates the probability that an interaction lasts for at most x seconds. Both functions were modeled with k=5 and θ=0.9 .

3.3 Dealing With False Positive and False Negative Detections of Human-Object Interactions

A common problem of human-object interaction systems are false positive object detections [10], e.g., when image regions are wrongly classified as objects. Furthermore, we observed in early tests of our framework a drop in the recall rates due to occlusions while the human interacted with an object, e.g., while drinking from a cup. To deal with such effects, we explicitly consider uncertain observations and compute the likelihood of possible human-object interactions to estimate the probability that the interaction really occurs based on their detection in subsequent frames.

To define the likelihood function, we evaluated the typical minimal length of human interaction with an object on a training data set collected in an university setting. Most interactions were shorter than 12 seconds. Longer interactions often last for several minutes, e.g., working with a laptop. This results in a distribution with a significant amount of data points during the first 12 seconds and very scattered data points for longer durations. Using fitting techniques with common probability functions we found that the *gamma probability density function*

$$G(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} \exp(-\frac{x}{\theta}) \qquad (1)$$

with k=5 and θ=0.9 and $\Gamma(a)$ as the *gamma function* [13], is a close approximation to the data. The resulting distribution for the first 12 seconds is visualized in Fig. 2a. As can be seen its fits the data points closely.

Given this distribution we compute the cumulative distribution function $G_C(x)$ of $G(x)$, which models the probability that an interaction has a duration of x or less seconds

$$G_C(x) = \frac{1}{\Gamma(k)}\gamma(k, \frac{x}{\theta}) \tag{2}$$

with $\gamma(a, b)$ as the *incomplete gamma function* [13], see Fig. 2b for a visualization of this cumulative distribution function.

Using $G_C(x)$, we determine the likelihoods of the detected possible human-object interactions (see Sec. 3.2) using their estimated duration. In more detail, for an interaction with a given object we say that a frame is a hit if an interaction with this object was detected and a $miss$ otherwise. To get the length of an interaction over several frames we mark a hit as the start point of a new interaction if the elapsed time since the last hit for this interaction is greater than a threshold $t_{max} = 5\,s$. We determined this value from $G_C(x)$ as 50% of the interactions are within a duration of 5 seconds. Not detecting a single hit for a specific human-object interaction during this time is a strong indication that no interaction with the object took place. First, each likelihood is initialized with 0 for all objects in each frame. Then, we compute the likelihoods of all detected interactions and possibly update the likelihoods of object interactions on previous frames where the interaction was not detected. Alg. 1 lists our complete approach to compute the likelihood of an observed human-object interaction, with *timeDiff*(a, b) as the time difference between a and b.

```
Input: Possible human-object interaction $I$ on frame $f_t$,
time of previous hit for $I$ $t_{phit}$, start time of $I$ $t_{start}$.
Output: Likelihood $L$ that $I$ really occurred.
$t_{diff} = timeDiff(t, t_{phit})$
if $t_{diff} > t_{max}$ then
    //new interaction with this object
    $t_{start} = t$
end
$L = G_C(timeDiff(t, t_{start}))$
if $(t - 1) < t_{diff} < t_{max}$ then
    //false negative detections for I occurred
    set likelihood of $I$ in frames $f_{tphit+1}, \cdots, f_{t-1}$ to $L$
end
$t_{phit} = t$
return $L$
```

Algorithm 1: Likelihood computation.

4 Experimental Evaluation

We performed extensive experiments to demonstrate the robustness of our approach with respect to precision and recall. Furthermore, we show the improvement that can

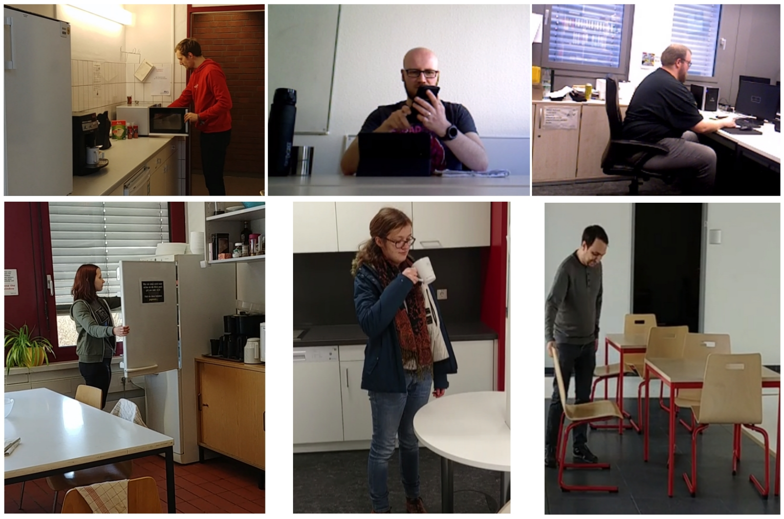


Fig. 3: Six example interactions from our evaluation dataset in different environments.

be achieved by computing the likelihood that an interaction is really happening by considering subsequent observations. In particular, we collected a dataset containing 195 human-object interactions of 10 different people with objects from the *Open Image* dataset [7], over 27 minutes of video data. All videos were recorded with 12 frames per second in indoor environments [2]. Fig. 3 shows 6 example interactions from our dataset in different environments.

We manually created the ground truth for each frame, i.e., the information which human-object interactions are taking place.

4.1 Precision and Recall

We compare the output of our approach on each frame with the ground truth to compute the recall and precision rates. We hereby perform the evaluation with respect to the likelihood value min_L from which on we assume our framework to be certain enough to return a found interaction with an object. Fig. 4 shows the evolution of the precision and recall for 100 different values of min_L equally distributed in the range from 0 to 1. The results were fitted with a function using least squares.

Using a min_L value of 0.21 our framework is able to achieve recall and precision rates of 0.82. Accordingly, in practice we use this as threshold for the likelihood as both the recall and precision are relatively high. As can be seen in Fig. 2b this corresponds to a minimal interaction length of approximately 2 seconds. Shorter minimal interaction lengths are possible but would result in a high precision loss.

[2] Videos from our dataset are availably under https://www.hrl.uni-bonn.de/icsr2019

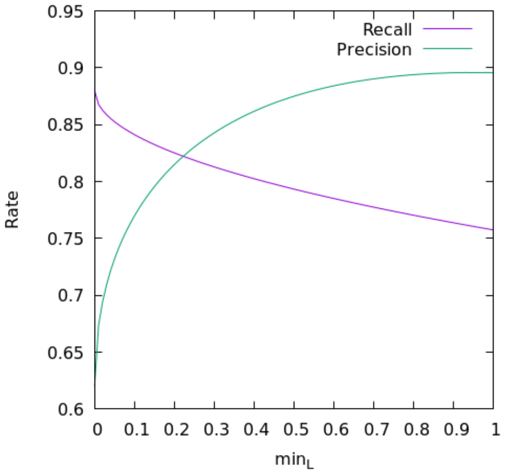


Fig. 4: Evolution of precision and recall rates of our approach with respect to different values of min_L. The precision rate strongly increases with higher min_L values while the recall rate only slowly decreases.

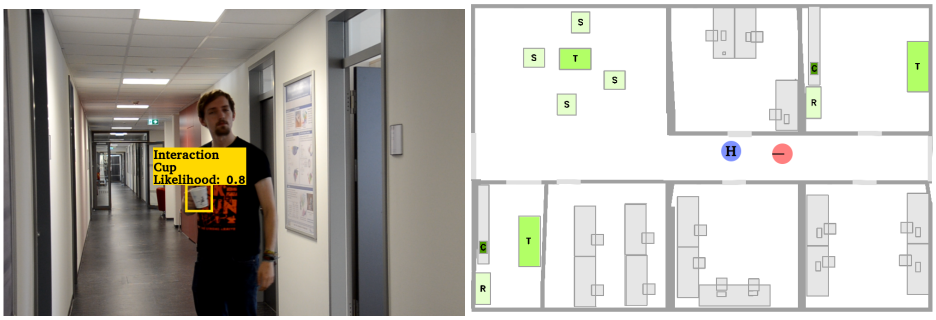
Increasing min_L to a value close to 1 results in higher precision values up to 0.88 and lower recall values of 0.75. Decreasing min_L on the other hand has the opposite effect resulting in a precision rate of 0.62 and recall rate of 0.82 for a value of min_L close to zero. This evaluation clearly highlights the usefulness of the verification step using the likelihood computation as it can significantly improve the returned precision value while only slightly reducing the recall value. In general, values for min_L between 0.2 and 0.6 seem to be a good compromise between high precision and recall rates.

False negative detections naturally happen at the start of an interaction since the corresponding likelihood is initially low as the duration of the interaction is very short at this point in time. False positive detections typically happen if objects are very close together. In this case it is very difficult to differentiate between the interacted and the passive object.

Comparison with the literature is difficult as the focus of most approaches that we are aware of is activity recognition and not human-object interaction detection. The most similar approach we found in the literature [10] lists recall values of 0.90 and precision values of 0.62. It should also be noted that this system was only able to detect interactions with 2 types of objects, while our approach is able to detect interactions with 510 different objects.

4.2 Application to Human Motion Prediction

As demonstrated, our framework is able to robustly detect human-object interactions in video streams. By applying the system online to the video stream recorded with the camera of a real robot, the robot is able to predict human motions. To do so, we first learn a distribution from collected data to represent the probability that after an



(a)

(b)

Fig. 5: Example application of our approach to predict human movement goals. The robot (red) detects a human-object interaction with a cup using our framework (a). Based on a pre-learned probability distribution about interaction transitions, the likelihood of possible next interaction objects is computed (b). The darker the green the higher the likelihood. Object names are abbreviated: table (T), sofa (S), refrigerator (R), coffee machine (C).

interaction with an object A the human will next interact with an object B. The robot can then use this knowledge to predict future movement goals of the human based on the known locations of objects in the environment when it detects human-object interactions.

Fig. 5 shows an application example. In this scenario, the robot detects a human-object interaction with a cup and computes based on an interaction model transition probabilities to other known objects. The most likely next objects in this example are sofas, tables, refrigerators, and coffee machines.

In our previous work[3], we showed that such knowledge about subsequent object interactions can improve the prediction of human motions compared to approaches which rely on learned trajectories [1].

5 Conclusion

In this paper, we present a novel approach to automatically extract human-object interactions from video streams. In comparison to existing frameworks, our system focuses on the detection of general interactions with objects rather than specific activities. Furthermore, we use spatio-temporal information to verify found interactions. We use an R-CNN to detect objects and the *OpenPose* pose estimator [2] to detect humans and their poses. Based on this information, we find human-object interactions on the current frame and compute for each interaction the likelihood that it is really happening based on subsequent observations.

As the experimental evaluation demonstrates, our approach is able to robustly detect human-object interactions with recall and precision rates of 0.82 on our test dataset.

[3] A video showing the capabilities of this approach can be found under `https://www.hrl.uni-bonn.de/icsr_application_demo.mp4`

Acknowledgments

We would like to thank Nils Dengler, Sandra Höltervennhoff, Sophie Jenke, Saskia Rabich, Jenny Mack, Marco Pinno, Mosadeq Saljoki and Dominik Wührer for their help during our experiments.

References

1. Bruckschen, L., Dengler, N., Bennewitz, M.: Human motion prediction based on object interactions. In: Proc. of the European Conference on Mobile Robots (ECMR) (2019), to appear
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2017)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2014)
4. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2018)
5. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(10) (2009)
6. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2017)
7. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Kamali, S., Malloci, M., Pont-Tuset, J., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. (2017)
8. Li, H., Ye, C., Sample, A.P.: IDSense: A human object interaction detection system based on passive uhf rfid. In: Proc. of the ACM Conf. on Human Factors in Computing Systems. ACM (2015)
9. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI) **40**(12) (2018)
10. Prest, A., Ferrari, V., Schmid, C.: Explicit modeling of human-object interactions in realistic videos. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI) **35**(4) (2013)
11. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI) **34**(3) (2012)
12. Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., Schiele, B.: Coherent multi-sentence video description with variable level of detail. In: German Conf. on Pattern Recognition. Springer (2014)
13. Weisstein: Gamma function. `http://mathworld.wolfram.com/GammaFunction.html`, accessed: 2019-02-24
14. Yang, C., Zeng, Y., Yue, Y., Siritanawan, P., Zhang, J., Wang, D.: Knowledge-based role recognition by using human-object interaction and spatio-temporal analysis. In: Proc. of IEEE Int. Conf. on Robotics and Biomimetics (ROBIO) (2017)
15. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: Proc. of Computer Vision and Pattern Recognition (CVPR) (2010)