Attribute Based Affordance Detection from Human-Object Interaction Images

Mahmudul Hassan(✉) and Anuja Dharmaratne

School of IT, Monash University Malaysia,
Jalan Lagoon Selatan, 47500 Bandar Sunway, Malaysia
{mahmudul.hassan,anuja}@monash.edu

Abstract. The detection of functional classification of an object, which is also referred as affordance is a prevalent researched topic in the domain of robotics and computer vision. Typically, the approaches regarding fine level affordance (affordance related to core traits of an object i.e. graspability, rollability etc.) detection are often disjoint from the techniques in higher level affordance detection (i.e. drinkability or pourability of a glass). In this paper, we have proposed an attribute based technique for higher level affordance detection which integrates methods from both fine level and high level affordance detection, and takes three prominent contexts (Human, Object and the ambience) into account. It further represents each of these contexts as a cluster of attributes rather than singular entities thus making the affordance detection process more semantic, efficient, dynamic and general.

Keywords: Affordance · Attribute transfer · Modelling mutual contexts

1 Introduction

The astonishing stature that separates humans from other forms of biological entities is its inherited capability of learning. Our ancestors were been able to excavate the potentiality of fire woods, coals and other objects that afford, not only to lit fire but also cooking. In plain context the tacit knowledge of learning the usability of different objects is an integral part of the success story of the human race. In the domain of Computer vision and Artificial intelligence, it is a very important topic of research. Though it sounds very linearly simple, but the research paradigm in this field is quite multidimensional and encompasses both fine level and high level layers. For example in practical robotics the premier focus is to master the identification of fine level usability of the objects (i.e. rolling ability, graspability etc.) in contrast in computer vision the focus is on some higher tones. Here researchers are more inclined to detect higher level usage of objects (i.e. how a human performs interaction with a computer). The way these two streams of researches are approached is often disjoint. In this paper we have tried to portray the benefits of combining the techniques from these two sects of researches. In broader terms we have used objects basic visual features

F. Huang and A. Sugimoto (Eds.): PSIVT 2015 Workshops, LNCS 9555, pp. 220–232, 2016.
DOI: 10.1007/978-3-319-30285-0_18

(SIFT, textons, edges, color histograms etc.) to infer some higher level attributes such as its material, shape, size, visual parts etc. to find the objects usability. We have also tried to boost the detection process of objects usability by using different contexts such as the human demonstration (i.e. body poses) and ambient objects (i.e. the induced effects of one object to others). We believe this integration of attributes and contexts make the detection process more robust, semantic and general.

1.1 Psychological Perspective of Affordance

The theory of affordances [1] was introduced as a theory of direct perception, which can account for findings in development psychology. According to [2,3] An affordance is an intrinsic property of an object. In broader sense, affordance is the functional classification of objects. Affordance is neither subjective nor objective. It depends on the object being interacted, the human who interacts and the ambient objects. For example, a chair is meant for sitting for an adult but for a toddler it does not have the sitting affordance rather, it has the climbing affordance. On the other hand when we see a mug alone, we infer its affordance as drinkable but as soon as we see a pitcher on top of it, its affordance space accumulate the pourable affordance as well. Hence, it is evident to state that, the affordance of an object is basically the mapping of these three contexts.

1.2 Ontological Classification of Affordance Detection Techniques

Generally affordance prediction has been approached from two different prospective. Firstly, the methods that learn the affordances passively by observing the humans interacting with the objects [4–6] and on the other hand methods that use objects visual features (appearance) to learn the objects affordance [7,8]. Usually the first category of works does focus on the high level affordance detection where the later works concentrate on the finer level affordances. But these conventional views towards affordance detection are now changing and there is a new trend where the researchers are combing both human actions and objects perception for affordance detection [9–13]. These mixed approaches are primarily used for higher level affordance detection [4], [9–12], [14–16]. Apart from robust affordance detection, these blended techniques, emerging substantially as an adequate tool for solving different problems in the computer vision domain, such as: classifying object categories [9,10,17], scene understanding [18–20], segmenting sub-activities from continues high level activity [21], robot navigation [22], robot-object placement [23], anticipation of human action [24–26] etc.

There is also a variant ontological prospective in the fine level affordance modeling approaches. It is visual features [7,8] versus the physical attributes of the objects [27–29]. In the visual features based approaches, finer features like Corner points (SIFT, HOG), edges, texture (textons), colour (Histograms) etc. are extracted from the objects and directly mapped with affordances. In contrast with that, in the attribute approaches, the finer level visual features are used to predict mid-level physical attributes [27] such as size, shape, material, weight

etc. A key advantage of attribute based detection is the ability to leverage object properties which are shared by multiple affordances, leading to more effective generalization to novel examples and the ability to learn new affordances with limited training data.

1.3 Challenges in Affordance Detection

There are fundamental difficulties in both the above mentioned approaches. Regarding direct perception approaches, there are three major issues. Firstly, affordances are not actually determined-in the physical sense, by visual features, rather by the physical properties of the objects [30]. Whether an object can roll or not depends on the shape of the object; whether it can be pushed is influenced by its material properties. Secondly, the visual features are very much vulnerable from different imaging and viewing phenomenon. Thirdly, a liability of the direct perception based methods is that there is no form of knowledge transfer between the object classes. In contrast, problems with the approaches related to human demonstration (both considering the objects and without it) are that, human could perform same actions with different objects (like mopping a table has similar body pose of ironing). Moreover a single object can have multiple affordances, therefore it is required to train the system with each action-object pair and consequently the training process becomes very complex and lacks the generalization (Systems usually suffers if an unseen object or body poses are considered). Another challenge in demonstration based affordance detection is that affordance depends on the 'attributes' of a person; it will not remain same for all the humans with same object. For example if the height of the human changes, than the possible actions that can be performed with a certain object will vary. Even the attributes of the objects and the ambient environment (other objects nearby) influence the affordance of an object.

2 Overview and Contribution of Our Approach

The core competency of our model is that it takes into account mutual contexts of the attributes of the object, the human and the ambient environment. We believe the affordance detection process of an object can improve substantially by considering the mutual relations of these contexts and their attributes. For example a knife has the primary affordance of cutting. But if the knife is made of plastic, it does not afford to cut harder objects. Here, we can see the change in the attributes can change the affordance space of that object. Similarly if we see a human is performing a stirring action with a knife, than the affordance space of the knife again changes and accumulate stirring. It shows that the body poses of the human helps us to infer the affordance of an object. Again if we see a knife is near to a food can or a biscuit tin, it may afford opening them. Here the ambient object induced the change in the affordance space of the object. Especially from static images, where unlike the video no temporal references are available this process of mutual context analysis helps the system

to develop a knowledge base and detect affordance more robustly. Furthermore, since our approach of affordance detection represents these contexts as sets of different attributes rather than considering them stand alone entities. It ensures the system to be more semantic, efficient, dynamic and general. For instance, in accordance with the previous example, we do not detect/classify the object as knife rather we describe it as a rectangle, metal object with sharp edges. We describe the objects with different attributes according to [30]. Simultaneously we also describe the mutual relations of the objects with human and the other ambient objects by a number of attributes [29]. Have used this attribute based representation for unseen object class detection, and they claim that this method does possess knowledge transfer mechanism and helps to recognize unseen and untrained objects. We have find that attributes are shared between the spectrums of different affordance classes also. For instance most of the objects that afford drinking (i.e. mug, cup, bottle, flask etc.) are cylindrical in shape and may have a handle (i.e. a mug handle).

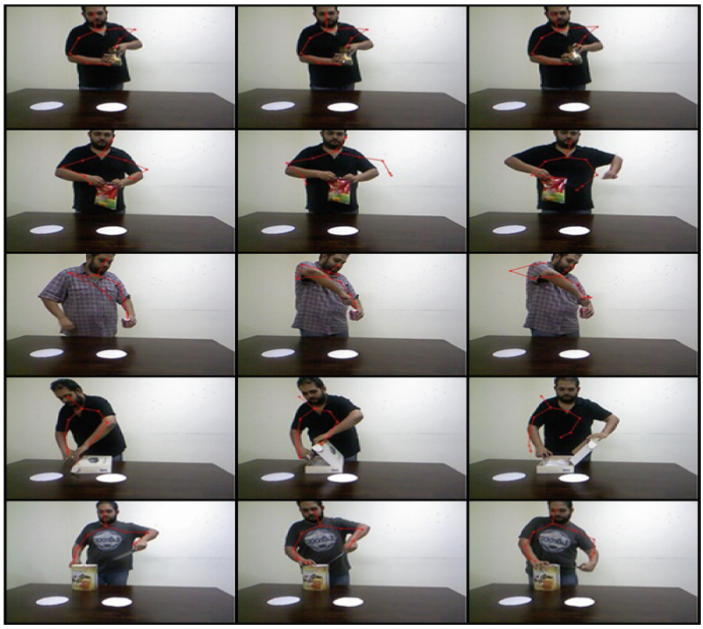

Fig. 1. Opening a jar (top row), Opening a poly(2nd row), Opening a drinks can (3rd row), Opening packet (4th row) and opening a tin with knife (bottom row).

In this paper, we have portrayed the importance the attributes in detecting human object interactions robustly. For instance, in Fig. 1 we consider opening object actions. We have multiple opening scenarios, like: opening a can, opening a packet of potato chips, opening a flask, opening a box etc. We analyzed and inferred that, we have different opening body poses for different object classes

due to the difference in the attributes of the objects. At the same time the attributes related to the human and the ambient objects are also important. Our work is inspired by the works of [27] and [31]. The main focus of this research is to combine the different notions of affordance modeling in order to achieve a robust affordance model. We are using the visual features to predict mid-level physical attributes of the objects and as well as the human and the environment (the other nearby objects). After that we use the physical attributes as the features to learn (both as parameters and the structure) our high level affordance detection model.

Our novel attribute based affordance model, encompasses two types of features related to the human, object and the environment in order to model objects affordances, namely: visual features and physical attributes [27]. The visual features are the basic image features extracted from images.

The Visual Features that We have Considered are:

- **For the Objects:** SURF(speeded up robust features) features, HOG (Histograms of oriented gradients) features, Edges, Textons, Region properties of bounding boxes, Image histograms, Euclidian distances between multiple objects.
- **For the Subjects:** Human body joint coordinates from kinect, the angles between the shoulder-arm-wrist (for both left and right hand).

After extracting the visual features, we have created multiple classifiers to classify physical attributes related to both objects, human and ambient objects.

The Physical Attributes that We have Considered are:

- **For the objects:** Material, Aspect ratio, Height, Objects shape, Color, Orientation.
- **For the Human:** Body poses, angle of the arms.
- **For Human-Object:** The distance of the object(s) from each body joints.
- **For Object-Object:** Euclidean distance between multiple objects, the spatial location of objects relative to other objects, relative aspect ratio of multiple objects.

The flow of our system is as follow: First, given images with human interacting with different objects, we select the bounding boxes of the object(s). After that we extract the base features from the selected bounding boxes (objects). We also extract the body joint coordinates of the human and the angles of the arms. Then we use these base features to train mid-level attribute classifiers. Subsequently we use these mid-level attributes as the features of our overall affordance model. In the test scenario, given the bounding boxes (the user provide the bounding boxes), the system can detect affordance of the selected objects more semantically and robustly.

3 Attribute Based Affordance Model

Our affordance model can be formalized by the following statements:

- The affordance space as (λ) where (λ) is a m dimensional vector.
- Objects visual features are (θ) where (θ) is a t dimensional vector.
- Objects physical attributes are (α) where (α) is a p dimensional vector.
- Body pose features are (β) where (β) is a q dimensional vector.
- Humans physical attributes are (γ) where (γ) is a r dimensional vector.
- Ambient environment attributes are (ε) where (ε) is a s dimensional vector.

Then, we can formalize the model as:

$$(\lambda) = f(\alpha, \beta, \gamma, \varepsilon, \theta) \tag{1}$$

So, if we want to represent the relations of these components in a joint distribution form:

$$p(\lambda, \alpha, \beta, \gamma, \varepsilon, \theta) = p(\lambda \mid \alpha, \beta, \gamma, \varepsilon, \theta)p(\alpha \mid \beta, \gamma, \varepsilon, \theta)p(\beta \mid \gamma, \varepsilon, \theta)p(\gamma \mid \varepsilon, \theta)p(\varepsilon \mid \theta)p(\theta) \tag{2}$$

$$p(\lambda, \alpha, \beta, \gamma, \varepsilon, \theta) = p(\lambda \mid \alpha, \gamma, \varepsilon)p(\alpha \mid \theta)p(\gamma \mid \beta)p(\varepsilon \mid \beta, \theta) \tag{3}$$

So, for finding the affordance, we can marginalize λ, and we get by the variable elimination method:

$$p(\lambda \mid \alpha, \beta, \gamma, \varepsilon, \theta) = \sum_{\alpha} \sum_{\beta} \sum_{\gamma} \sum_{\varepsilon} \sum_{\theta} p(\lambda, \alpha, \beta, \gamma, \varepsilon, \theta) \tag{4}$$

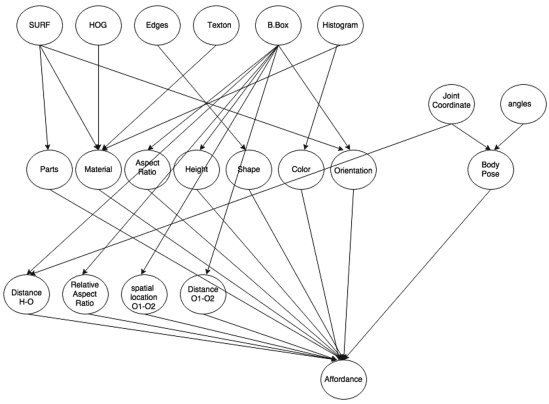

Fig. 2. The Bayesian network representation of the proposed model.

Currently we have implemented our attribute based affordance model with Bayesian network (Fig. 2). We have compared two different scoring methods

for learning the structural model of the Bayesian network, (1) Bayesian Information Criterion and (2) Greedy hill climbing (optimization). For inference we have used junction tree algorithm. Apart from the Bayesian network, we have also implemented our model with a multi-dimensional SVM and K-Nearest neighbor algorithm. In the case of the KNN, we have tested the model with Euclidean, Cityblock and Minkowski distances. We have used the N-fold validation for cross validation of the model.

3.1 Attribute Classifiers

As we have stated earlier the mid-level attributes are classified from the base features. We have implemented separate classifiers for each of the mid-level physical attributes.

Parts Classification. We have introduced a novel physical attribute called Parts. It is basically distinct image patches of objects which are common in all the objects in a single affordance class. Different object classes can have a single affordance, we argue that though these object classes are dissimilar in visual aspects but they do share some common parts. For example the objects which have affordance of drinking or pouring usually have visual patches of a handle. These parts have proved to be a robust cue in our affordance detection model. For the part class detection, first we manually selected distinct parts of different objects (5 parts per affordance class) class that has a common affordance, and then these parts (patches) are cropped out from the object images (we have used 750 patches for each part class). In Fig. 3, different selected parts of the sitting affordance class is shown. These cropped patches are finally used as features for our part class detection classifier. We have trained our parts detection classifier with Bag of features algorithm, where we have trained the classifier with vocabulary sizes from 1000 to 4000 with 500 interval and final set the vocabulary size to 1500 (1500 clusters) since it has given us the highest accuracy. Patch size was set to [64 128 192 256] for the optimal efficiency. Finally a multiclass SVM is used as classification algorithm. The grid points (SURF points) were selected densely for the bag of features algorithm. For the part classification, we have achieved 71 % accuracy.

Material Classification. For the material detection of different objects (what the object is made of), we have extracted SURF points, HOG features, Textons [32] and image histograms from object images. These features are then subsequently given as inputs into a K-nearest neighbor classifier to detect materials. We have considered material type of: paper, metal, plastic, poly, food, glass and cloth. We have tested and compared our classifier with [32] and [33] where Textons and Fractals are used, and found that, our classifier is more suitable in detecting materials of real life objects. Real life objects have a lot of labeling and undesired interest points. Though [32] performs better with basic surface texture images of different materials but loses accuracy for real life objects. For the material classification images of each object class (1200 images per object class).

Fig. 3. The common parts in diverse objects that afford sitting.

Aspect Ratio and Height. Aspect ratio is a popular measurement that gives us a cue about the size of an object with some sort of scale invariance perspective. We have calculated the aspect ratio as the width over height of the selected bounding boxes. On the other hand height is simply the vertical height of the selected bounding box. We have simply used the measurements of the object bounding boxes as features for our aspect ratio and the height classifier, where a multi class SVM is used for training and classification.

Shape Classification. Shape is a very prominent feature of the object. Most of the time, the objects which share same affordances have their shape in common. We have classified shapes as Square, Cylindrical, round and 3D-boxy. We have first extracted the edges of the objects via Prewitt edge detector filter. Then we performed some morphological operations on the edges and the Hough matrix is calculated. Finally a curve fitting algorithm is used to find the similarities in shapes. We have compared our algorithm with [34], and observed that our efficiency is lower than it, but due to the complexity, we remained with our algorithm as the difference of the efficiency is not substantial. Currently, our shape detection classifiers accuracy is 78 %.

Color and Orientation Classification. For color classification, we have used simple histograms of the object images as features. The histograms of all Red channel, Green channel and Blue channel are used. The KNN algorithm is used to implement the classifier.

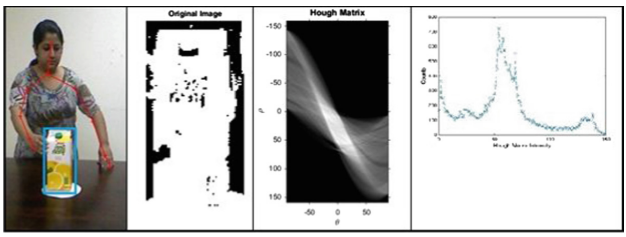

Fig. 4. The process of shape detection.

For the orientation, we have initially used rotating patches of object images with 900 variations and trained a SVM based object classifier to classify the Horizontal and Vertical version of the objects. But the classifier has not performed optimally due to different traits of object poses. Moreover, the object detection itself is a substantial challenge to handle and increases the complexities of the overall system to a significant. For time being, as the main focus of the current work is to depict the effects of attributes in the overall affordance detection we have manually input the orientation values of the object in our final affordance classifier.

Body Pose Classifier. For the objective to implement a robust body pose classifier, we had to first identify and segment human body from a cluttered scene. Then we had to acquire the body joint locations. We have first tried to use simple Part based method [35], where week classifiers are trained with HOG features to detect and track the body parts, but the results were not optimal. Later we have used the Microsoft Kinect sensor to capture RGBD images and got the articulated human skeleton by Kinect SDK. The Kinect Skeleton viewer function, that is a part of the support package of Kinect SDK, provides coordinates 20 body joints of detected human body robustly. We have considered only the 10 joints of the upper body (Shoulder center, Head, Left shoulder, Right shoulder, Left elbow, Right elbow, Left wrist, Right wrist, Left hand, Right hand). We have used these coordinates as the base features for human action pose detection. Our novel action pose detection classifier is inspired by the concept of [34]. We have represented the body poses not by the mere coordinates of the body joints but by the distance of each body joints from the head. This method helped us to offset view point variance and translation variance to some extent. For the classification we have used the K-NN algorithm.

For the human action pose classifier, we have also used the inner angles of the elbows as base features. The vector dot products were used to determine the angle (Fig. 5).

Fig. 5. Detected skeletons in different actions.

Human-Object Distance Classifier. For Human-Object distance attribute classifier, we have used the Euclidian distance between the object centroid and the human body joints (Skeleton joints, acquired by Kinect) as base features. A multidimensional SVM is used for the classification.

Relative aspect Ratio and Relative Spatial Location Classifier. Relative aspect ratio and relative spatial locations are the attributes which are only used in the case of multiple objects. Relative aspect ratio implies the comparison of the aspect ratio of one object to other. We have find that the relative aspect ratio gives us a useful insight of objects affordance in a multiple object setting. For instance for a pouring action, most of the time the larger object is Pour from object and the smaller object is pour to object. For the relative spatial location, we have decomposed each image frames into nine cells as: Center, Above, Bottom, Left, Right, Upper left, Upper right, Bottom left, Bottom right. We index the locations of each object by these cells and use them as base features.

4 Training and Inference

For training the affordance classifier, we have used 9632 images of different actions being performed. There are 22 action classes performed with 43 objects. 4 subjects (person) were used to perform the actions. The action classes are: (1) Spraying in the body (2) Chopping (3) Cutting (4) Drinking with both hands (5) Drinking with single hands (6) Eating snacks (7) Eating fruit (8) Ironing (9) Mopping (10) Opening poly (11) Opening box (12)Opening can (13) Opening jar (14) Opening packet (15) Opening tin (16) Pouring with both hands (17) Pouring with single hands (18) Spraying in the air (19) Stacking with both hands (20) Stacking with a single hand (21) Waving (22) Answering mobile phone.

For the training of the attributes (material, shape, color and parts) classifiers, we have used features extracted from objects images from different datasets such as 'Caltech 256' and SHORT-100 and also downloaded images from the web. For testing our model we have used Human-Object-Interaction images from known object classes (The affordance classes which are trained) and also novel object classes. In the test dataset, there are also instances where the objects are partially occluded and the human body poses are unknown.

5 Model Evaluation

We have tested our model with a test dataset of 3 subjects performing 22 actions with 18 objects. Total instances of the test dataset are 528. We have initially implemented our model with SVM, KNN and Bayesian networks to find the most suitable algorithm for our model (pilot testing). Due to the best empirical results, a Bayesian Network based method is used for constructing the final affordance model. For comparing these three algorithms a prototype test dataset was used which is different from the actual testing dataset.

We have compared our model with two baseline models. For baseline (a) we tested the models which used only human body pose as features for Human-Object-Interaction detection and for base line (b), the models which used the mutual contexts of human body poses and detected object classes for affordance detection.

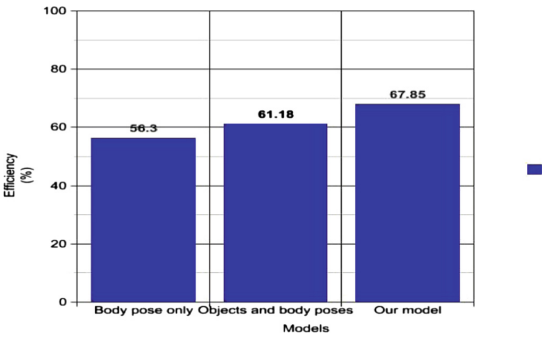

Fig. 6. The comparison of our model with the base lines.

Figure 6 shows the comparative results of our attribute based affordance model with the base lines. The overall accuracy of our model is 67.85 %. This accuracy is acquired by testing the model with both known and unseen object classes. It shows the accuracy and generalization improves a substantial amount with our model. The overall accuracy of the base line algorithms are 61.18 % (Objects and body poses) and 56.3 % for body pose only.

6 Conclusion

In contrast with the current affordance detection models in the computer vision and robotics domain, we have implemented our model by considering mutual contexts of Human-Object and ambient environment. Moreover we have represented each context with a cluster of attributes. Due to the inclusion of multiple contexts and knowledge sharing capability within the attributes our model proved to perform more efficiently, semantically and has generalization quality.

References

1. Gieseking, J.: The People, Place, and Space Reader. Hilldale, New Jersey (2014)
2. Helbig, H.: Action observation can prime visual object recognition. Exp Brain Res. **200**(3–4), 251–258 (2009)
3. Kjellstrm, H.: Visual object-action recognition: Inferring object affordances from human demonstration. Comput. Vis. Image Underst. **115**(1), 81–90 (2011)
4. Manuela, V., Rybski, P., von, F.: FOCUS: A generalized method for object discovery for robots that observe and interact with humans. In: Proceedings of the 2006 Conference on Human-Robot Interaction, IEEE Press, Salt Lake City (2006)

5. Zhu, Y., Fathi, A., Fei-Fei, L.: Reasoning about object affordances in a knowledge base representation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 408–424. Springer, Heidelberg (2014)
6. Grabner, H., Gall, J., Gool, V.: What makes a chair a chair? In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colaradao (2011)
7. Castellini, C., Tommasi, T., Noceti, N., Odone, F., Caputo, B.: Using object affordances to improve object recognition. IEEE Trans. Auton. Mental Dev. **3**, 207–215 (2011)
8. Moldovan, B., Moreno, P., van Otterlo, M., Santos-Victor, J., De Raedt, L.: Learning relational affordance models for robots in multi-object manipulation tasks. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 4373–4378, Minnesota (2012)
9. Roudposhti, K.K., Dias, J.: Probabilistic human interaction understanding. Pattern Recognit. Lett. **34**, 820–830 (2013)
10. Desai, C., Ramanan, D., Fowlkes, C., Kesselman, C.: Discriminative models for static human-object interactions. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2010)
11. Bangpeng, Y., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco (2010)
12. Bangpeng, Y.H., Liu, M., Philipose, M., Pettersson, H., Sun, M.: Subsequences. J. Vis. Commun. Image Representation **25**, 719–726 (2014)
13. Bangpeng, Y., Xiaoye, J., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1331–1338, Barcelona (2011)
14. Packer, B., Saenko, K., Koller, D.: A combined pose, object, and feature model for action understanding. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1378–1385. Rhode Island (2012)
15. Ikizler-Cinbis, N., Sclaroff, S.: Object, scene and actions: combining multiple features for human action recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 494–507. Springer, Heidelberg (2010)
16. Peursum, P., Venkatesh, S., West, G.A.W., Bui, H.H.: Object labelling from human action recognition. In: Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, pp. 399–406, Dallas (2003)
17. Jakkula, V., Diane, J.C.: Mining sensor data in smart environment for temporal activity prediction. In: Poster session at the ACM SIGKDD, San Jose, CA (2007)
18. Jiang, Y., Saxena, A.: Infinite latent conditional random fields. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 262–266. IEEE (2013)
19. Jiang, Y., Saxena, A.: Modeling high-dimensional humans for activity anticipation using Gaussian process latent CRFS. In: Robotics: Science and Systems, San Francisco (2014)
20. Koppula, H., Gupta, R.: Saxena: ILearning human activities and object affordances from rgb-d videos. Int. J. Rob. Res. **32**, 951–970 (2013)
21. Sun, J., Moore, J., Bobick, A., Rehg, J.: Learning visual object categories for robot affordance prediction. Int. J. Robot. Res. **29**, 174–197 (2009)
22. Jiang, Y., Koppula, H., Saxena, A., Kesselman, C.: Hallucinated humans as the hidden context for labeling 3D scenes. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 2993–3000, Portland (2013)

23. Koppula, H., Ashutosh, S.: Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13). Atlanta (2013)
24. Koppula, H.S., Saxena, A.: Anticipating human activities for reactive robotic response. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), p. 2071, Tokyo (2013)
25. Jiang, Y., Saxena, A.: Modeling high-dimensional humans for activity anticipation using gaussian process latent CRFs. In: Proceedings of Robotics: Science and Systems, San Francisco (2014)
26. Hermans, T., Rehg, J.M., Bobick, A.: Affordance prediction via learned object attributes. In: IEEE International Conference on Robotics and Automation (ICRA): Workshop on Semantic Perception, Mapping, and Exploration, pp. 181–184. IEEE Press, New York (2011)
27. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 951–958. Florida (2009)
28. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1785, Florida (2009)
29. Sun, J.: Learning visual object categories for robot affordance prediction. Int. J. Robot. Res. **29**, 174–197 (2010)
30. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: IEEE Conference on Computer Vision and Pattern Recognition, 17–22, Minnesota (2007)
31. Leung, T., Malik, J.: IRepresenting and recognizing the visual appearance of materials using three-dimensional textons. Int. J. Comput. Vis. **43**, 29–44 (2001)
32. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. Int. J. Comput. Vis. **62**, 61–81 (2005)
33. Salve, S.G., Jondhale, K.C.: Shape matching and object recognition using shape contexts. In: 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), pp. 471–474 (2010)
34. Lu, X., Chia-Chih, C., Aggarwal, J.K.: Human detection using depth information by kinect. In: 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Colarado (2011)