Chapter 5
Modeling Paired Objects and Their Interaction

Yu Sun and Yun Lin

5.1 Introduction

Object categorization and human action recognition are two important capabilities for an intelligent robot. Traditionally, they are treated separately. Recently, more researchers started to model the object features, object affordance, and human action at the same time. Most of the works build a relation model between single object features and human action or object affordance and uses the models to improve object recognition accuracies [16, 21, 12].

In our daily life, it is natural that we not only pay our attentions to the objects we hold and manipulate, but also the interactive relationship between the objects. We also select our motions according to the intended interaction we want, which is mostly defined by both objects. For example, when a person holds a pen, there could be many different kinds of motions. However, if the pen is associated to a piece of paper, the human motion with the pen is significantly confined. Most likely, a writing motion will occur. Likewise, if we want to detect the type of object in a human hand, and we have detected a human writing motion and a piece of paper, we have more confidence to believe that the object is a pen than without detecting the writing motion or the paper. There are many similar examples such as a book and a schoolbag, and a teapot and a cup. The interactive motions performed by the humans have strong relationship with both objects. Therefore, the motion information can enhance our belief of the recognition results of the objects. If we can detect a stirring motion and recognize a cup, we can enhance our belief that the object in the human's hand is a spoon. Figure 5.1 shows several objects on a table that have inter-object relationship: a CD and a CD case, a pen and a piece of paper, a spoon and a cup, and a cup and a teapot.

Yu Sun · Yun Lin
Computer Science and Engineering, University of South Florida, Tampa, FL 33620, U.S.A.
e-mail: yusun@cse.usf.edu, yunlin@mail.usf.edu

© Springer-Verlag Berlin Heidelberg 2014

Y. Sun et al. (eds.), *New Development in Robot Vision*,
Cognitive Systems Monographs 23, DOI: 10.1007/978-3-662-43859-6_5

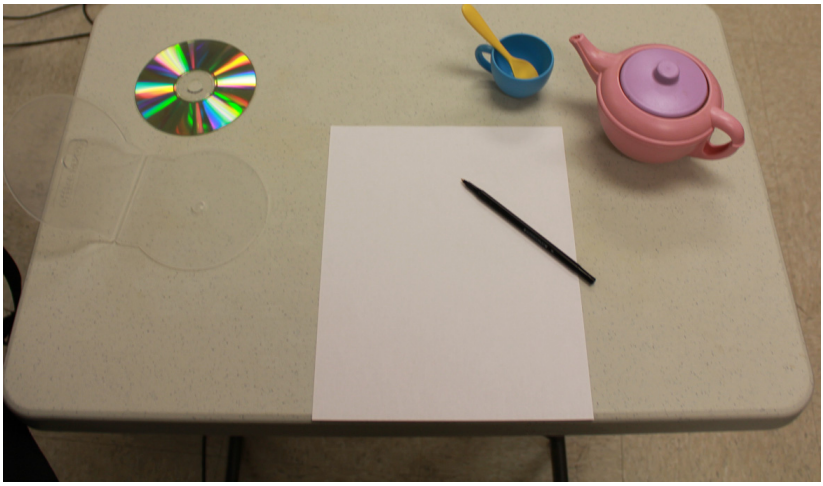


Fig. 5.1 Several objects on a table have inter-object relationships: pen-paper, teapot-cup, cup-spoon, CD-CD case

The connection between the visual recognition and motor action has been studied in neuroscience and cognitive science recently. The concept of objects' affordance has been around since 1977 [14]. Only lately studies on objects' affordance [28, 29, 24] indicated that the mirror neurons in human brains congregated visual and motor responses. In the studies, the researchers found that mirror neurons in the F5 sector of the macaque ventral premotor cortex fired both during observation of interacting with an object and during action execution, but did not discharge in response to simply observing an object [9, 13]. More close to the human-object-object interaction affordance idea, Yoon et al. [32] studied the affordances associated to pairs of objects positioned for action and found an interesting so-called "paired object affordance effect." The effect was that the response time by right-handed participants was faster if the two objects were used together when the active object (supposed to be manipulated) was to the right of the other object. Borghi et al. [3] further studied the functional relationship between paired objects and compared it with the spatial relationship and found that both the position and functional context were important and related to the motion; however, the motor action response was faster and more accurate with the functional context than the spatial context. The study results in neuroscience and cognitive science indicate that there are strong connections between the observation and the motion, and functional relationships between objects are directly associated with the motor actions.

Based on the new findings in neuroscience and cognitive science, we propose to link a pair of objects with their interaction motion directly and we call the interaction motion instead of the functionalities of the object as the inter-object affordance. In this chapter, we attempt to capitalize the strong relationship between paired objects and interactive motion by building an object relation model and associating it to

human action model in the human-object-object way to characterize inter-object affordance.

In robotics and related fields, object affordance has only been explored recently in limited works that mainly model the object affordance with the interaction between single object and a human user, and then use the mutual relation to improve the recognition of each other. For example, Gupta and Davis [16] recently achieved inspiring success in using single object action to improve the recognition rate of both the object and human motion. Kjellstrom et. al. [21] used conditional random field (CRF) and factorial conditional random field (FCRF) to model the relationship between object type and human action, in which the 3D hand pose was estimated to represent human action including open, hammer, and pour actions. Most recently, Gall, et. al. [12] have recovered the human action from a set of depth images and then represented object's function and affordance with the human action. In their work, objects were classified according to the involved human action in an unsupervised way base on high-level features.

Another recent approach in literature is to derive the objects' affordance from their low level features or 3D shapes. Stark et. al. [30] obtained the object affordance cues from human hand and object interaction in the training images, and then they detected an object and determine the objects functions according to the objects affordance cue features. Grabner et. al. [15] proposed a novel way to determine object affordance using computer graphical simulation. The system 'imagines' or simulates an actor performing actions on the objects to compute the objects affordances from the object's 3D shape.

In robotics community, there are several existing works on obtaining and using object-action relation. In [1], objects were categorized solely according to object interaction sequences (motion features), but the geometry appearance features of the objects was not considered. First, the objects were segmented out from the background in a number of video sequences, then the space interaction relationship between objects were represented with an undirected semantic graph. Their work was able to represent the object temporal and spatial interactions in an event with a sequence of such graphs.

In summary, most of the existing works focus on object-action interaction, or object geometry-related affordance features. This chapter based on our previous publications [27, 31], introduces our new works on modeling the affordance relationship between objects for object recognition and presents a way to model the inter-object affordance, and then use the inter-object affordance relationship to improve object recognition.

In this chapter, we describe a design of a graphical model that composes of two objects and the human motions that relate both objects. The graphical model contains the inter-object affordance that can be learned to represent the interaction relationship between paired objects, such as teapot-cup, and pen-paper. A Bayesian Network is structured to integrate the paired objects, their interaction, and the consequence of the object interaction. After the description of the Bayesian Network graphical model, we introduce an approach to recognize the paired objects by analyzing and classifying the interactive motions with the statistical knowledge learned

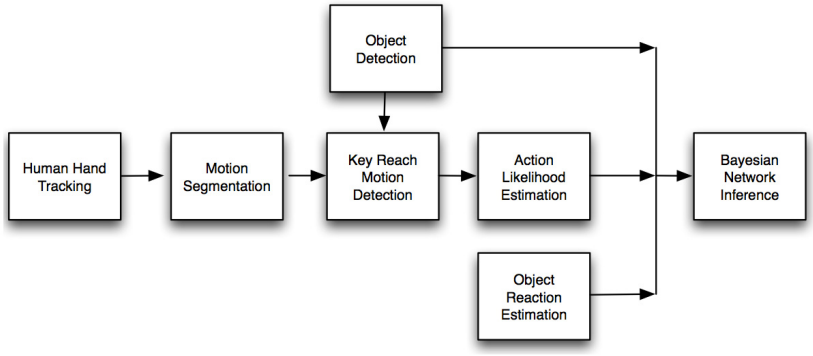


Fig. 5.2 The workflow of building the human-object-object interaction model starts with object detection, human hand tracking, and object reaction estimation. In the end the likelihoods are used to build a Bayesian inference network.

from training data. In addition, at the end of the chapter, we extend this approach to leverage the object recognition accuracy from videos with the interactive motion recognition and demonstrate the benefit of the approach with results in several experiments, which show that the detection accuracy of the interactive objects was significantly improved with the introduced approach.

5.2 Human-Object-Object-Interaction Modeling

The workflow to build the human-object-object interaction model is illustrated in Figure 5.2. First, the initial likelihood of the objects' manipulation and reaction is computed. The object initial likelihoods were estimated with a sliding window object detector, which is based on the Histogram of Oriented Gradients (HoG). The initial likelihood of human action is estimated based on the feature of human hand motion trajectory. The human hand was tracked in the whole process, and the hand motion was segmented according to the velocity changing. With motion segmentation and possible object locations, the interactive object pairs were detected in the step of key reach motion detection. The start time of the manipulation was estimated based on the object pair locations and hand motion trajectory. Then, the initial belief of the manipulation was computed.

Object interaction usually leads to a state change of the associated objects. For example, if a CD is put into a CD case, the color of the CD case probably will change. The likelihood of object reaction was estimated by comparing with the training datasets. Finally, the belief in each node was updated with the inference algorithm for Bayesian Networks.

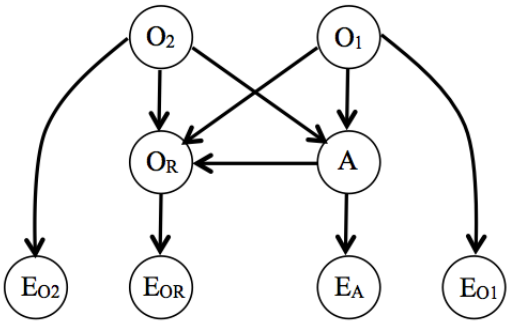


Fig. 5.3 The Bayesian network model used to represent objects, actions and object interactions. O_1 and O_2 represent the two interacting objects, A denotes hand manipulation motion, and O_R is the object reaction.

5.2.1 Bayesian Network Model for HOO Interaction

$$\begin{aligned} P(O_1, O_2, A, O_R|e) \propto\ & P(O_1|e_{O_1})P(O_2|e_{O_2}) \\ & P(A|O_1, O_2)P(A|e_A) \\ & P(O_R|O_1, O_2, A)P(O_R|e_{O_R}) \end{aligned} \tag{5.1}$$

Bayesian network is chosen to model the HOO interaction because it is a powerful inference tool for decision making in the observation of several or many interrelated factors. As illustrated in Figure 5.3, the Bayesian network introduced here has eight nodes. The two interactive objects are represented as node O_1 and node O_2. Node A denotes hand manipulation action, also represents the inter-object affordance. The node O_R represents the object reaction that reflects the change of object state after the interaction. The rest notes are the evidences $e = \{e_{O_1}, e_{O_2}, e_A, e_{O_R}\}$, and they represent the evidence for O_1, O_2, A, and O_R respectively. The nodes are connected according to their conditional dependencies. Since node A is determined by the two interacting objects (O_1 and O_2), they are the parents of node A. Similarly, since the object reaction is the consequence of the two objects and the manipulation, it is the child of those three nodes. The belief for each node can be updated with the messages from the corresponding evidence node. According to the Bayesian rule and conditional independence relations, the joint probability distribution of the paired objects, inter-action, and reaction can be represented with Equation 5.1.

The Bayesian network model can be scaled up by increasing the number of variables for object and action in each node without changing the graphical model

structure. Alternatively, we can combine multiple Bayesian networks to form a large-scale graphical model if there are inter-connections between different pairs of objects.

5.2.2 Object Detection

To estimate the initial likelihood of the objects, a detector similar to [7] was designed. The detector works in the sliding window manner, and uses a variant of the HoG feature from [10] to represent the object local features. At each pixel, the color channel with the largest gradient magnitude was used to represent the gradient orientation and magnitude. In each detecting window, the image was divided into 8x8 pixel cells and, for each cell, the pixel level feature was aggregated to a feature map.

Objects were modeled as object type and object location. We computed the object likelihoods:

$P(O_1 = \{obj_1, l^{O_1}\}|e_{O_1})$ and
$P(O_2 = \{obj_2, l^{O_2}\}|e_{O_2})$

for each sliding window with the SVM estimation, in which l^{O_1} is the location of start object and l^{O_2} is the location of the end object . Figure 5.4 shows a sample of the detection results using training image images from the Image-Net [8] and Google Image Search. All of the training images were labeled. For each object, 50 positive and 70 negative examples were used to train an SVM (Support Vector Machine) classifier. The window size and aspect ratio were learned from the training data set. The LibSVM library [4] was used to obtain the probability of the classification for each window.

Fig. 5.4 Example result of object detection with SVM classifier using HoG features. Dots indicate detected object centers.

5.2.3 Motion Analysis

The object detector in the previous section can only give us the possible object locations with their types. Since the inter-object affordance is represented by the object interaction, that affordance should be modeled with motion features. To represent the inter-object action – the affordance of the pair, it is necessary to detect and analyze the hand motion that is associated with one of both of the objects. Here the trajectories are segmented and the motion segments are used to represent and recognize the motion types. Generally, there are two kinds of object interactive motion – putting an object into a container and manipulating one of the objects relative to the other [18, 19]. In this chapter, these two kinds of motion are treated the same, although they are considered different in cognition science.

5.2.3.1 Human Hand Tracking in 2D

It is difficult to track an arbitrary hand in a daily-living environment with various background solely based on the hand's shape as a hand can have many different shapes for different gestures. To simplify the discussion, this chapter describe an approach using the human skin color as tracking features since it is much more stable and has been used successfully in previous works [2]. In addition, the skin color model in [5] and the TLD object tracker [20] are combined to build a stable hand tracker. In this approach, the hands in the initial several frames are located using optical flow and the skin color. Then for each additional frame, the hand location is updated according to the color information around the previous hand location and the shape features from TLD tracker. Figure 5.5(a) shows one example of the tracking result and the Figure 5.5(b) shows the tracked trajectory for a whole inter-action motion – putting a CD into its case.

5.2.3.2 Motion Segmentation

From the tracked hand motion trajectory, motion features should be extracted to represent the motion. Here, the obtained trajectories can then be segmented into several pieces according to the velocity and represented with the motion features in the segments. According to [26], there are two kinds of human limb motions: ballistic motion and mass spring motion. In those two kinds of motions, the velocity provides natural indications of the motion segments. The local minimal points in their velocity curves are used to segment the trajectories, and then these small pieces can be either merged or segmented further into possible ballistic and mass spring segments. Similar to the method in [26], the segments are classified into ballistic and mass sprint types according to their velocity features. The features used in this chapter include the maximum velocity, average velocity, number of local minimum point, standard deviation, and motion distance etc. Figure 5.5(c) shows the motion segments in velocity for one motion that is putting a pencil into a pencil case. Similar motion analysis approaches exist in neuroscience and cognitive science to classify and represent motion segments with action chains [11, 17].

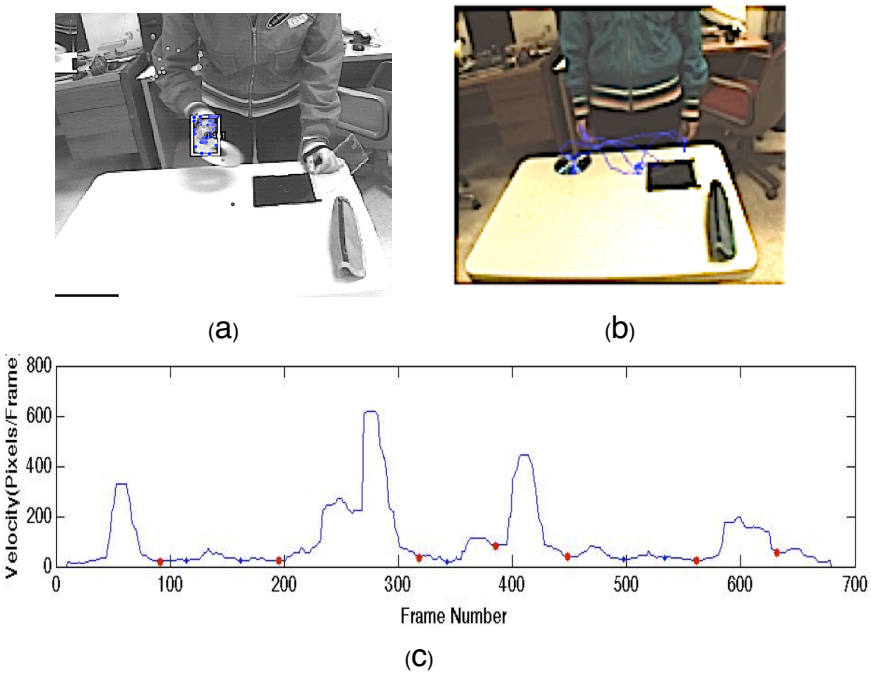
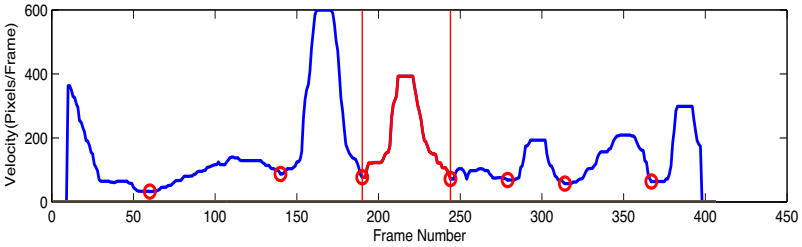


Fig. 5.5 Hand tracking and motion segmentation: (a) right-hand motion tracking; (b) right-hand motion trajectory; (c) motion segmentation with velocity – horizontal axis is time (frame number), and vertical axis represents velocity (pixels per frame). Red circles are detected motion segment boundaries.

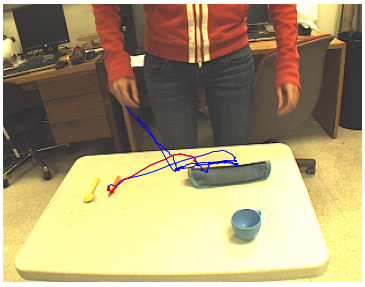
5.2.3.3 Key Reach Motion Detection

In each object interaction process, a human hand carries one object to the location of another object. For example, in the stirring water example, a human hand carries a spoon and moves it to the cup. This reach motion is called the key reach motion. There could be several reach motions in one action. For example, in a process of putting a book into a schoolbag, there are three reach motions. A person first opens the schoolbag, the first reach motion; reach to the book, the second reach motion; and then take the book to the schoolbag to put into it, the third reach motion. However, only the taking the book to the schoolbag is defined as the key reach motion for this interaction as only this reach motion involves both objects. Therefore we name the book as the start object and the schoolbag as the end object as object1 and object2 respectively in the graphical model.

The ballistic segments are then further classified into reach motion and non-reach motion according to motion features including the velocity during acceleration and deceleration, time duration, average velocity, and stand deviation of the velocity. However, it is difficult to segment out the key reach motion only based on the hand



(a)



(b)

Fig. 5.6 Key reach motion detection: (a) red velocity segment represents key reach motion in velocity graph. The red circles are detected motion segment boundaries; (b) The red curve shows key reach motion in image.

motion and to detect if a hand is carrying object or not if the object is small. Instead, we rely on the motion of the object since it is easy to detect the object state around the start and end location of the reach motion. The key reach motion starts from one location (l^a_{r1}), and ends at another location (l^a_{r2}). The distance between the location of start object (l^{O_1}) and the start of the key reach motion location l^a_{r1} is modeled with a normal distribution, $N(|l^a_{r1}l^{O_1}|, \mu_r^{O_1}, \sigma_r^{O_1})$. Likewise, the distance between the location of the end object (l^{O_2}) and l^a_{r2} is modeled with $N(|l^a_{r2}l^{O_2}|, \mu^{O_2}, \sigma_a^{O_2})$. The start and end locations for each reach motion are obtained in the tracking. Then, the start object, end object, and the key reach motion are detected at the same time, according to the two distributions values. Here $\mu_r^{O_1}$, $\sigma_r^{O_1}$, $\mu_a^{O_2}$, and $\sigma_a^{O_2}$ are learned from the training data set. In the key reach motion, human hand carries object1 from location l^{O_1} to location l^{O_2} , so the belief of the key reach motion can be further enhanced by checking if the detected start object (object1) is removed or not. This can be carried out by comparing the likelihood value of object1 at location l^{O_1} before and after the key reach motion. Figure 5.6 shows the key reach motion segment detected (marked as red) from the entire motion that put a pencil into a pencil case.

5.2.3.4 Manipulation Motion Estimation

A manipulation action can be modeled with the features in the human hand trajectory. The features are the start time (t_s^a), the end time (t_e^a), the two reach locations (l_{r1}^a, l_{r2}^a), and the manipulation type (T^a). According to Equation 1, we model the conditional probability $P(A|O_1O_2)$, and the initial likelihood of A, $P(A|e_A)$. $P(A|O_1O_2)$ can be computed with Equation refeq:moo1. If we define l_s^a as the hand location for the start time t_e^a, we can model $P(t_s^a, t_e^a|O_1O_2)$ with $N(|l_s^a l^O|, \mu_r^O, \sigma_r^O)$, and O is either O_1 or O_2. μ_r^O is the mean of the grasping distance for the object O, while σ_r^O is the variance, which can be learned from the training data. $P(l_{r1}^a|O_1)$ and $P(l_{r2}^a|O_2)$ are modeled as normal distributions $N(|l_{r1}^a l^{O_1}|, \mu_r^{O_1}, \sigma_r^{O_1})$ and $N(|l_{r2}^a l^{O_2}|, \mu_a^{O_2}, \sigma_a^{O_2})$, which have been discussed in Section 5.2.3.3. $P(T^A|obj_1, obj_2)$ is computed according to the occurrence of manipulation type and object type in the training data.

$$\begin{aligned} P(A|O_1O_2) = {} & P(t_s^a, t_e^a|O_1O_2)P(l_{r1}^a|O_1) \\ & P(l_{r2}^a|O_2)P(T^a|obj1, obj2) \end{aligned} \tag{5.2}$$

We estimate the likelihood $P(A|e_A)$ with the features from the hand motion trajectory. Based on the segmentation results in Section 5.2.3.2, the ballistic and mass spring segments are replaced with labels. The manipulation motions are classified according to the numbers of ballistic and mass spring segments, the translation rate of the two segments, and time duration etc. Linear SVM is trained as the classifier and gives the likelihood of the manipulation.

5.2.4 Object Reaction

The object reaction node is modeled with two parameters: reaction type (T^R) and reaction location (l^R). It is difficult to fully model the object reaction. Therefore, we only consider the state change of the object2 after the interaction. Similar to [4], we use the color histogram at the object2 to represent the object reaction. We estimate $P(O_R|e_{O_R})$ by comparing the histogram of the object2 with the histogram of the training instances from the training data set. Then we model the prior $P(O_R|O_1, O_2, A)$ according to Equation (5.3). $P(l^R|O_2)$ is model with $N(|l^R l^{O_2}|, \mu^R, \sigma^R)$, and parameters μ^R and σ^R are learned from the training data. $P(T^R|O_1, O_2, A)$ is learned from the training data set by counting the occurrence of T^R, O_1, O_2 and A.

$$P(O_R|O_1, O_2, A) = P(l^R|O_2)P(T^R|O_1, O_2, A) \tag{5.3}$$

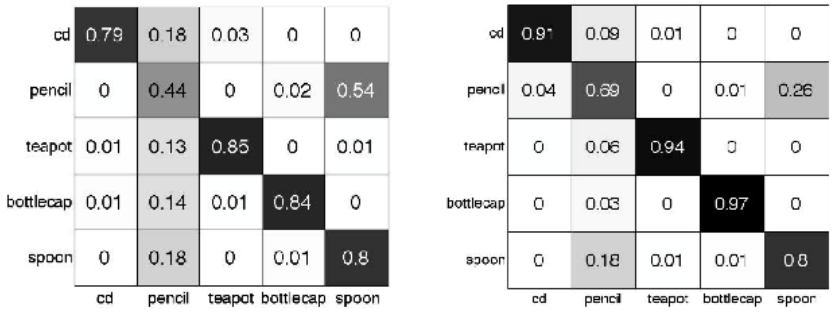
5.2.5 Bayesian Network Inference

After getting the key reach motion and the interaction object pair locations, we estimate the parameters for A and O_R according to Sections 5.2.3.3 and 5.2.3.4. We perform the inference with Pearls algorithm [25] once all of the initial likelihoods

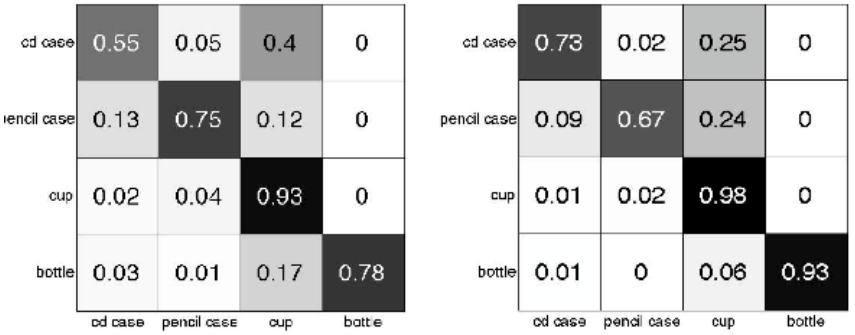
for O_1, O_2, A, and O_R are estimated. The Bayesian Network, the object classifier and the manipulation classifier are trained with fully-labeled data.

5.3 Experiments and Results

The following experiment and evaluation results demonstrate how this approach is used and its performance. A dataset was collected from six subjects who performed five types of interactions of five pairs of objects. The interaction object pairs included teapot-cup, pencil-pencil case, bottle cap-bottle, CD-CD case and spoon-cup. The actions for these object pairs were pouring water from a teapot to a cup, putting a pencil into a pencil case, screwing on a bottle cap, putting a CD into the CD case and stirring a spoon in a cup. All of these objects and actions were chosen because they are very common in everyday life, and they are representative for different inter-object affordance relationships. The data from four subjects were used



(a)



(b)

Fig. 5.7 Results comparison: (a) Object1 likelihood confusion matrix. The left one shows the result using HoG detector. The right shows the result using the described approach; (b) Object2 likelihood confusion matrix. The left one shows the result using HoG detector. The right shows the result using our framework.

for training, and the data from the rest two subjects were used for testing. Each subject performed every action for two or three trials.

The object classifier, the action classifier and the Bayesian Network were trained in supervised manner. As stated before, the training images for the object classifier were collected from the ImageNet [8] and Google Image Search. The training data for the action classifier and the Bayesian Network were collected from manually labeled video sequences taken in our experiments. Fifty videos sequences that performed by four subjects were used for training. In each training video sequence, object locations, reach locations and action type and the start frame of the manipulation were manually labeled.

The test data set are video sequences that contain the action sequences performed by the other two subjects. Figure 5.7(a) shows the object classification confusion matrixes for object1 for the testing data, which is the object at the beginning of the key reach motion. Figure 5.7(b) presents the likelihood confusion matrixes for object2 that is the object at the end of the key reach motion. In each of the confusion matrices, the *i*th row represents the likelihood value when the *i*th type of object presents. For object1, as we can see from the confusion matrices, it was difficult to distinguish a pencil from a spoon only based on the appearance, which is consistent with the fact that they have the similar shape and both of them are small. With our approach, by including the context of human-object-object interaction, our Bayesian network was able to distinguish and recognize the spoon and the pencil more much accurately. The average recognition success rate of our approach for object1 was improved from 72.6% to 86.0% and improved from 75.3% to 82.8% for object2.

Among the five actions studied, if based only on motion features, it was difficult to distinguish putting a CD into a CD case, putting a pencil into a pencil case, pouring water into a cup, and stirring water in a cup because they had the similar motion patterns. With the human-object-object interaction framework, they could be distinguished. Figure 5.8(a) shows the likelihood confusion matrix that was estimated

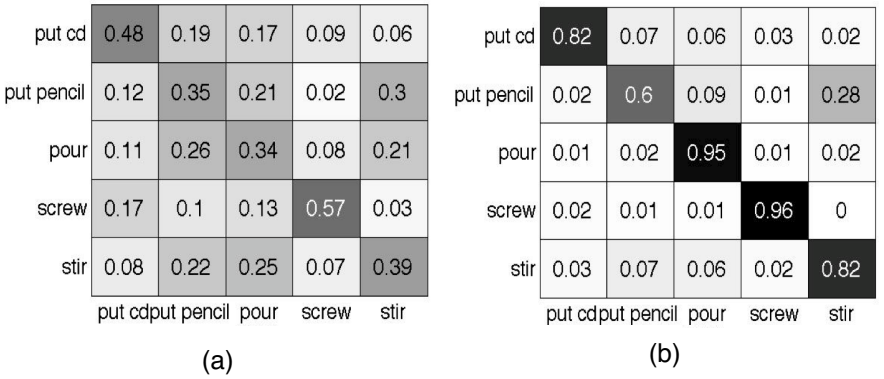


Fig. 5.8 Action likelihood confusion matrix: (a) result using only motion features; (b) result using framework. The *i*th row shows likelihood value when *i*th action is categorized.

with only hand motion features. Figure 5.8(b) shows the action confusion matrix using human-object-object interaction framework. We can see that the overall average recognition rate across all objects improved from 42.6% to 83.0%.

5.4 Conclusions

This chapter described a recent investigation on modeling the human-object-object interaction with Bayesian network. The object categorization and action recognition are linked using human-object-object-interaction affordance framework. The knowledge of object affordance is learned from labeled video sequences, and represented with a Bayesian Network. The elements of the Bayesian Network include objects, human action and object reaction. The experiments with six subjects and about 70 video sequences have shown that with human-object-object-interaction affordance knowledge, the object classification rate, and especially the action recognition rate were significantly improved.

The learned affordance knowledge represented in the Bayesian network can also help us to learn affordance motion more precisely and apply the learned motion to guide and control robot motions in a learning from demonstration framework such as in [22], since the interaction affordance knowledge can suggest proper actions that the robot should perform. The interaction motion can also be used to compute a feasible and stable manipulation-task oriented grasp planning [23] with the help of the object categorization. The motion analysis presented in this chapter is only one of many approaches. A functional motion analysis could be applied (similar to [6]) to capture more dynamic features and represent the motion in a lower dimensional space.

References

1. Aksoy, E., Abramov, A., Worgotter, F., Dellen, B.: Categorizing object-action relations from semantic scene graphs. In: IEEE Intl. Conference on Robotics and Automation, pp. 398–405 (2010)
2. Argyros, A.A., Lourakis, M.I.A.: Real-time tracking of multiple skin-colored objects with a possibly moving camera. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 368–379. Springer, Heidelberg (2004)
3. Borghi, A., Flumini, A., Natraj, N., Wheaton, L.: One hand, two objects: emergence of affordance in contexts. Brain and Cognition 80(1), 64–73 (2012)
4. Chang, C., Lin, C.: Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 1–27 (2011)
5. Conaire, C., O'Connor, N.E., Smeaton, A.F.: Detector adaptation by maximising agreement between independent data sources. In: IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum, pp. 1–6 (2007)
6. Dai, W., Sun, Y., Qian, X.: Functional analysis of grasping motion. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1–7 (2013)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)

8. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Int. Conf. on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
9. Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., Rizzolatti, G.: Understanding motor events: A neurophysiological study. Exp. Brain Res. 91, 176–180 (1992)
10. Felzenszwalb, P.F., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
11. Fogassi, L., et al.: Parietal lobe: From action organization to intention understanding. Science 29(308), 662–667 (2005)
12. Gall, J., Fossati, A., Gool, L.: Functional categorization of objects using real-time markerless motion capture. In: Conference on Computer Vision and Pattern Recognition, pp. 1969–1976 (2011)
13. Gallese, V., Fogassi, L., Fadiga, L., Rizzolatti, G.: Action representation and the inferior parietal lobule. In: Prinz, W., Hommel, B. (eds.) Attention and Performance XIX. Common mechanisms in perception and action. Oxford University Press, Oxford (2002)
14. Gibson, J.: The theory of affordances. In: Shaw, R., Bransford, J. (eds.) Perceiving, Acting and Knowing. Erlbaum, Hillsdale (1977)
15. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: Conference on Computer Vision and Pattern Recognition, pp. 1529–1536 (2011)
16. Gupta, A., Davis, L.: Objects in action: An approach for combining action understanding and object perception. In: Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
17. Hamilton, A., Grafton, S.: The motor hierarchy: from kinematics to goals and intentions. In: Haggard, P., Rosetti, Y., Kawato, M. (eds.) Attention and Performance, ch. 22 (2007)
18. Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J., et al.: Grasping the intentions of others with one's own mirror neuron system. PLoS Biol. 3(3) (2005)
19. Jax, S.A., Buxbaum, L.J.: Response interference between functional and structural actions linked to the same familiar object. Cognition 115(2), 350–355 (2010)
20. Kala, Z., Matas, J., Mikolajczyk, K.: P-n learning: Bootstrapping binary classifiers by structural constraints. In: Conference on Computer Vision and Pattern Recognition, pp. 49–56 (2010)
21. Kjellstrom, H., Romero, J., Kragic, D.: Visual object-action recognition: Inferring object affordances from human demonstration. Computer Vision and Image Understanding 115, 81–90 (2010)
22. Lin, Y., Shaogang, R., Clevenger, M., Sun, Y.: Learning grasping force from demonstration. In: IEEE Intl. Conference on Robotics and Automation, pp. 1526–1531 (2012)
23. Lin, Y., Sun, Y.: Task-oriented grasp planning based on disturbance distribution. In: International Symposium on Robotics Research (ISRR), pp. 1–16 (2013)
24. Oztop, E., Kawato, M., Arbib, M.: Mirror neurons and imitation: a computationally guided review. Epub Neural Networks 19, 254–271 (2006)
25. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Network and Plausible Inference. Morgan Kaufmann (1988)
26. Prasad, V.S.N., Kellokumpu, V., Davis, L.S.: Ballistic hand movements. In: Perales, F.J., Fisher, R.B. (eds.) AMDO 2006. LNCS, vol. 4069, pp. 153–164. Springer, Heidelberg (2006)

27. Ren, S., Sun, Y.: Human-object-object-interaction affordance. In: IEEE Workshop in Robot Vision, pp. 1–6 (2013)
28. Rizzolatti, G., Craighero, L.: The mirror neuron system. Ann. Rev. Neurosci. 27, 169–192 (2004)
29. Rizzolatti, G., Craighero, L.: Mirror neuron: A neurological approach to empathy. In: Changeux, J.P., Damasio, A.R., Singer, W., Christen, Y. (eds.) Neurobiology of Human Values, Springer, Heidelberg (2005)
30. Stark, M., Lies, P., Zillich, M., Wyatt, J.C., Schiele, B.: Functional object class detection based on learned affordance cues. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 435–444. Springer, Heidelberg (2008)
31. Sun, Y., Ren, S., Lin, Y.: Object-object interaction affordance learning. Robotics and Autonomous Systems (in Press)
32. Yoon, E., Humphreys, W., Riddoch, M.: The paired-object affordance effect. J. Exp. Psychol. Human 36, 812–824 (2010)