Human-Object-Object-Interaction Affordance

Shaogang Ren
University of South Florida
Tampa, FL USA 33620
shaogangren@mail.usf.edu

Yu Sun
University of South Florida
Tampa, FL USA 33620
yusun@cse.usf.edu

Abstract

This paper presents a novel human-object-object (HOO) interaction affordance learning approach that models the interaction motions between paired objects in a human-object-object way and use the motion models to improve the object recognition reliability. The innate interaction-affordance knowledge of the paired objects is modeled from a set of labeled training data that contains relative motions of the paired objects, humans actions, and object labels. The learned knowledge of the pair relationship is represented with a Bayesian Network and the trained network is used to improve recognition reliability of the objects.

1. Introduction

Traditionally, object categorization and human action recognition 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 [1, 2, 3].

It is natural in our daily life that we not only pay our attentions to the objects we hold and manipulate, but also the the interactive relationship between objects. We also select our motions according to what kind of the interaction will happen and that 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 motions with the pen is significantly limited. Most likely, a writing motion will occur. Likewise, if we detect a human writing motion and a piece of paper, the chance that the object in the human hand is much higher than without writing motion or the paper. In addition to the functionalities of the object, the interaction motion is more confined and associated to a pair of objects and we call it the inter-object affordance. 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 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. In this paper, we attempt to capitalize the strong relationship between paired objects and interactive motion by building an object relation model and associating it human action model in the human-object-object way to characterize inter-object affordance.

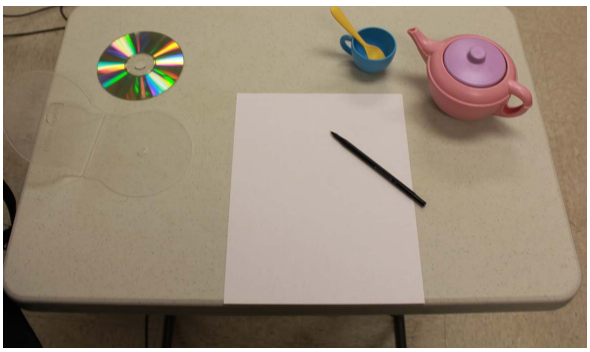


Figure 1. Several objects on a table have inter-object relationships.

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 [1] recently achieved inspiring success in using single object action to improve the recognition rate of both the object and human motion. Kjellstrm et. al. [2] 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. [3] have recovered the human action from a set of depth images and then represented object's function and affordance with the human action. In

978-1-4673-5648-0/13/$31.00 ©2013 IEEE

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. [4] 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. [5] proposed a novel way to determine object affordance using computer graphical simulation. The system imagines or simulate 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 [6], 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. To the authors' knowledge, there is no existing investigation on modeling the affordance relationship between objects for object recognition. This paper presents a way to model the inter-object affordance, and then use the inter-object affordance relationship to improve object recognition.

Different from existing work, we design a graphical model that composes of two objects and the human motions that relate the both object. 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, the interact action, and the consequence of the object interaction.

From the Bayesian Network graphical model, we developed an approach to recognize the paired objects by analyzing and classifying the interactive motions with the statistical knowledge learned from training data. In addition, we extend this approach to leverage the object recognition accuracy from videos with the interactive motion recognition. Our results in several experiments show that the detection accuracy of the interactive objects is significantly improved with our proposed approach.

2. Human-Object-Object-Interaction Modeling

We start by obtaining the initial likelihoods of the objects, human manipulation, and object reaction. Among them, the objects' initial likelihoods are estimated by a sliding window object detector that is based on the Histogram of Oriented Gradients (HoG). We then estimate the initial likelihood of human manipulation action from the features in the trajectories of human hand motions. In our approach, we assume the human hand can be tracked at all time. The hand motion can be segmented according to the trajectory's velocity characters. The start time of the manipulation is estimated based on the object pair locations and hand motion trajectory.

With the motion segmentation and possible object locations in an image, the interactive object pair can established. Then the initial believe of manipulation is changing. For example, if a CD is put into the CD case, the color of the CD case probably will change. The likelihood of object reaction is estimated by comparing with the training datasets. Finally the belief in each node is updated with the inference algorithm for Bayesian Networks.

2.1. Bayesian Network Model for HOO Interaction

$$\begin{aligned} P(O_1, O_2, A, O_R|e) \quad \propto \quad & P(O_1|e_{O_1})P(O_2|e_{O_2}) \\ & P(A|O_1, O_2)P(A|e_A) \\ & P(O_R|O_1, O_2, A)P(O_R|e_{O_R}) \end{aligned} \tag{1}$$

We choose Bayesian network because it is a powerful inference tool for decision making in the observation of several or many interrelated factors. As illustrated in Figure 2, our Bayesian network 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 1.

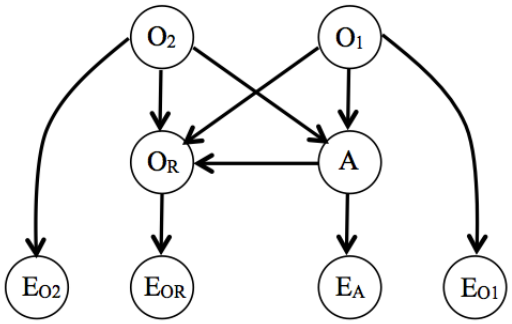


Figure 2. The Bayesian network model used to represent objects, actions and object interactions.

To estimate the initial likelihood of the objects, we implemented an approach according to Ref. [7]. The object is detected with a sliding window and compare the local features represented in HoG features [8]. The window size and aspect ratio are learned from training data set. Our training images are from the Image-Net [9] and the Google Image Search and they are labeled. We use around 50 positive and 70 negative examples to train a bi-class Support Vector Machine (SVM) classifier for each object. The LibSVM library [10] is used to obtain the probability of the classification for each window.

Each of the paired objects can be modeled with the object type (obj) and its current location (l). Therefore their initial likelihood is represented as $P(O_1 = \{obj_1, l^{O_1}\}|e_{O_1})$ and $P(O_2 = \{obj_2, l^{O_2}\}|e_{O_2})$. They are computed for each sliding window with the SVM estimation.

2.2. 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, we need to detect and analyze the hand motion that are associated with one of both of the objects. The hand motion should be tracked, and the motion trajectory should be analyzed. Here we break the trajectories to segments and use the motion segments to represent and recognize the motion types.

2.2.1 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. In this work, we use the human skin color since it is much more stable and has been used successfully in previous works [11]. In addition, we combine the skin color model in Ref. [12] and the TLD object tracker [13] to build our own hand tracker. In our 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. Here, since only the right hand was used in our experiments, the right hand motion is tracked. It is the same approach if we want to track the left hand as we don't distinguish them. Figure 3(a) shows one example of the tracking result and the Figure 3(b) shows the tracked trajectory for a whole inter-action motion – putting a CD into its case.

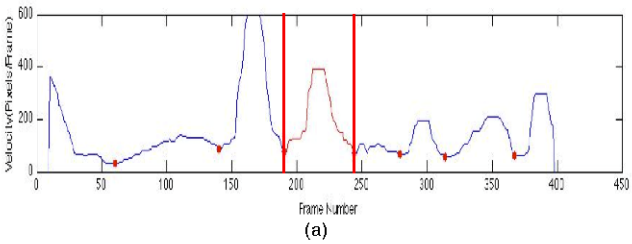
(a)

(b)

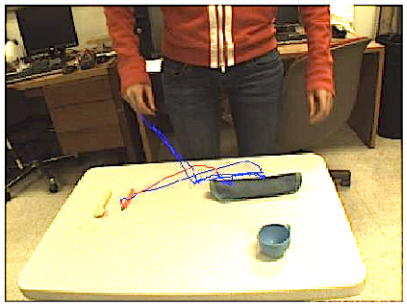
Figure 3. One example of the tracking in a 2D image: (a) the hand is tracked with a window; (b) the hand motion trajectory for a motion that puts a CD into a its case.

2.2.2 Motion Segmentation

From the tracked hand motion trajectory, motion features should be extracted to represent the motion. In this approach, we segment the obtained trajectories into several segments according to the velocity and represent the motion with motion features in the segments. According to Ref. [14], there are two kinds of human limb motions: ballistic motion and mass spring motion. In those two kind of motions, the velocity provides natural indications of the motion segments. We segment the trajectories with the local minimal points in their velocity curves, and then these



(a)



(b)

Figure 4. The segmented motion shown in red and between vertical lines is the motion putting a pencil into a pencil case.

small pieces can be either merged or segmented further into possible ballistic and mass spring segments. Similar to the method in Ref.[14], the segments are classified into ballistic and mass sprint types according to their velocity features. The features used in this paper include the maximum velocity, average velocity, number of local minimum point, standard deviation, and motion distance etc. Figure 4 shows the motion segments in velocity for one motion that is putting a pencil into a pencil case.

2.2.3 Key Reach Motion Detection

We noticed that the interaction motion usually include a reaching motion in which a human hand carries one object to the location of another object or they both reach toward each other. For example, in the stirring water example, a human hand carries a spoon and moves it to the cup. We call this reach motion as 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, we only call the taking the book to the schoolbag as the key reach motion for this interaction as this only reach motion involves both objects. Therefore we name the book as the start object and the schoolbag as the end object as object 1 and object 2 respectively in our 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 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, $N(|l^a_{r1}l^{O_1}|, \mu^{O_1}_r, \sigma^{O_1}_r)$. Likewise, the distance between the location of the end object (l^{O_2}) and l^a_{r2} is modeled with $N(|l^a_{r2}l^{O_2}|, \mu^{O_2}, \sigma^{O_2}_a)$. 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 $\mu^{O_1}_r$, $\sigma^{O_1}_r$, $\mu^{O_2}_a$, and $\sigma^{O_2}_a$ are learned from the training data set. In the key reach motion, human hand carries object 1 from location l^{O_1} to location l^{O_2} , so the believe of the key reach motion can be further enhanced by checking if the detected start object (object 1) is removed or not. This can be carried out by comparing the likelihood value of object 1 at location l^{O_1} before and after the key reach motion.

2.2.4 Manipulation Motion Estimation

A manipulation action can be modeled to the features in the human hand trajectory. The features are the start time (t^a_s), the end time (t^a_e), the two reach locations (l^a_{r1}, l^a_{r2}), 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 2. If we define l^a_s as the hand location for the start time t^a_e, we can model $P(t^a_s, t^a_e|O_1O_2)$ with $N(|l^a_s l^O|, \mu^O_r, \sigma^O_r)$, and O is either O_1 or O_2. μ^O_r is the mean of the grasping distance for the object O, while σ^O_r is the variance, which can be learned from the training data. $P(l^a_{r1}|O_1)$ and $P(l^a_{r2}|O_2)$ are modeled as normal distributions $N(|l^a_{r1}l^{O_1}|, \mu^{O_1}_r, \sigma^{O_1}_r)$ and $N(|l^a_{r2}l^{O_2}|, \mu^{O_2}_a, \sigma^{O_2}_a)$, which have been discussed in Section 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.

$$P(A|O_1O_2) = P(t^a_s, t^a_e|O_1O_2)P(l^a_{r1}|O_1) \quad (2)$$
$$P(l^a_{r2}|O_2)P(T^a|obj1, obj2)$$

We estimate the likelihood $P(A|e_A)$ with the features from the hand motion trajectory. Based on the segmentation results in Section 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.

We want to detect the key reach motion and the interacting object pair at the same time once we obtain the detected objects and hand motion trajectory.

2.3. 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 object 2 after the interaction. Similar to [4], we use the color histogram at the object 2 to represent the object reaction. We estimate $P(O_R|e_{O_R})$ by comparing the histogram of the object 2 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 (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.

$$P(O_R|O_1, O_2, A) = P(l^R|O_2)P(T^R|O_1, O_2, A) \quad (3)$$

2.4. 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 (2.3.3) and (2.3.4). We perform the inference with Pearls algorithm [15] 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.

3. Experiments and Results

We have evaluated our framework with a dataset collected from six subjects who performed five types of interactions of five pairs of objects. The interaction object pairs include teapot-cup, pencil-pencil case, bottle cap-bottle, CD-CD case and spoon-cup. The actions for these object pairs are 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 are 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 for training, and the data from the rest two subjects are used for testing. Each subject performs each action for two or three trials.

The object classifier, the action classifier and the Bayesian Network were trained in supervised manner. As

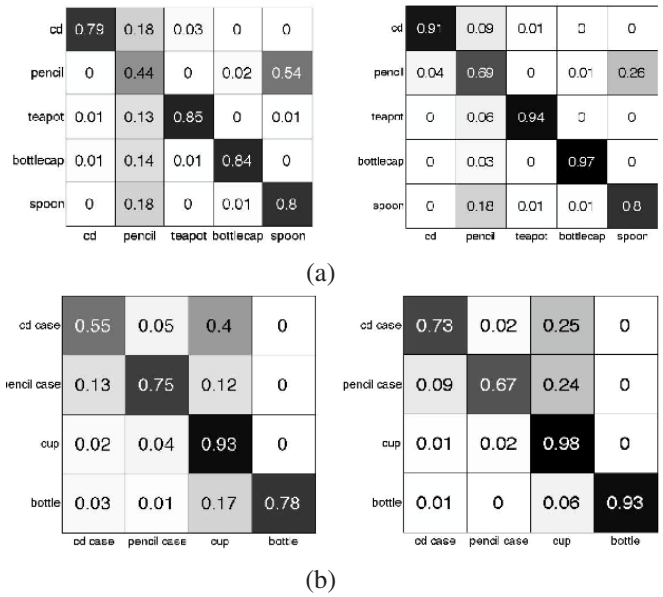


Figure 5. Results comparison: (a) Object 1 likelihood confusion matrix. The left one shows the result using HoG detector. The right shows the result using the our approach; (b) Object 2 likelihood confusion matrix. The left one shows the result using HoG detector. The right shows the result using our framework.

stated before, the training images for the object classifier are collected from the ImageNet [9] and Google Image Search. The training data for the action classifier and the Bayesian Network are collected from manually labeled video sequences taken in our experiments. About 50 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(a) shows the object classification confusion matrixes for object 1 for the testing data, which is the object at at the beginning of the key reach motion. Figure 5(b) presents the likelihood confusion matrixes for object 2 that is the object at the end of the key reach motion. In each of the confusion matrices, the ith row represents the likelihood value when the ith type of object presents. For object 1, as we can see from the confusion matrices, it is 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 can distinguish and recognize the spoon and and the pencil more much accurately. The average recognition success rate of our approach for object 1 is improved from 72.6% to 86.0% and improved from 75.3% to 82.8% for object 2.

4. Conclusions and Future Work

In this paper we investigated the object categorization and action recognition 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. Our experiments with six subjects and about 70 video sequences have shown that with human-object-object-interaction affordance knowledge, the object classification rate is significantly improved.

In the future, we plan to include more objects into our framework and investigate more complicated relations between objects. We also plan to use the learned affordance knowledge to help us to learn affordance motion more precisely and apply the learned motion in guiding and controlling robot motions in our learning from demonstration framework [16], since the interaction affordance knowledge can suggest proper actions that the robot should take to perform interactive tasks with paired objects. .

References

[1] Gupta, A., and L. Davis. (2007). Objects in Action: An Approach for Combining Action Understanding and Object Perception. Conference on Computer Vision and Pattern Recognition.

[2] Kjellstrm, H., J. Romero, and D. Kragic (2010). Visual Object-Action Recognition: Inferring Object Affordances From Human Demonstration. CVIU.

[3] Gall, J., A. Fossati and L. Gool (2011). Functional Categorization of Objects using Real-time Markerless Motion Capture. Conference on Computer Vision and Pattern Recognition.

[4] Stark, M., P. Lies, M. Zillich, J. Wyatt, and B. Schiele (2008). Functional Object Class Detection Based on Learned Affordance Cues. ICVS.

[5] Grabner, H., J. Gall and L. Gool (2011). What Makes a Chair a Chair? Conference on Computer Vision and Pattern Recognition.

[6] Aksoy, E., A. Abramov, F. Worgotter, and B. Dellen (2010). Categorizing Object-Action Relations from Semantic Scene Graphs. IEEE Intl. Conference on Robotics and Automation, 398-405.

[7] Dalal, N., and B. Triggs (2005). Histograms of Oriented Gradients for Human Detection. Conference on Computer Vision and Pattern Recognition.

[8] Felzenszwalb, P. F., D. McAllester, and D. Ramanan (2008). A Discriminatively Trained, Multiscale, Deformable Part Model. Conference on Computer Vision and Pattern Recognition.

[9] Deng, J., W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei (2009). ImageNet: A Large-Scale Hierarchical Image Database. Int. Conf. on Computer Vision and Pattern Recognition.

[10] Chang, C., and C. Lin (2011). LIBSVM : A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology.

[11] Argyros, A., and M. Lourakis (2004). Real-Time Tracking of Multiple Skin-Colored Objects with a Possibly Moving Camera. ECCV.

[12] Conaire, C ., N. E. O'Connor, and A. F. Smeaton (2007). Detector Adaptation by Maximising Agreement Between Independent Data Sources. IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum.

[13] Kala, Z., J. Matas and and K. Mikolajczyk (2010). P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints. Conference on Computer Vision and Pattern Recognition, 2010.

[14] Prasad, V ., Kellokompu, V., and Davis, L. (2006). Ballistic Hand Movements. F.J. Perales and R.B. Fisher (Eds.): AMDO 2006, LNCS 4069, pp. 153164.

[15] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Network and Plausible Inference. Morgan Kaufmann.

[16] Lin, Y., S. Ren, M. Clevenger, and Y. Sun (2012). Learning Grasping Force from Demonstration. IEEE Intl. Conference on Robotics and Automation:1-6.