Functional Descriptors for Object Affordances

Alessandro Pieropan Carl Henrik Ek Hedvig Kjellström

***Abstract*— In the context of robot learning from demonstration, it is very important that a robot understands what an object can be used for. By observing a human performing an activity, a robot should be able to identify the human motion, the objects involved and the outcome of the performed activity. One important aspect of this challenging problem is to detect and reason about objects in terms of affordance or alternatively, about their function in the current activity. Affordance is often modeled in terms of appearance however appearance does not necessarily map one-to-one with functional classes. In this paper we propose two alternative features that characterize objects directly in terms of how they are used. Our approach show a significant improvement compared to the traditional appearance based methods.**

I. INTRODUCTION

Reasoning about activities is an important skill for robots that operate in unstructured environments. A robot should be able to observe a human and reproduce the same performed activity [1]. To achieve that it is essential that a robot understands the function of objects in the observed activity. Features such as SIFT or HOG are often used to model affordance [2] however appearance does not necessarily map one-to-one with functional classes (Fig. 1). Instead we propose to characterize objects directly in terms of how they are used. There are some works in this spirit: [3] classifies activities by looking at how the different segments interact each other, or [4] proposes to use the human pose as a descriptor for the *sitable* affordance. In this paper the scenario we focus on is a human demonstrator teaching a robot about the affordances of objects, used in a kitchen scenario, by showing how they are used. Therefore we propose to describe objects in terms of how they are manipulated by the human and in terms of the spatio-temporal relationships existing between each pair of object used in an activity. The performances of our descriptors are compared to a baseline of appearance-based descriptors (SIFT bag-of-words) in terms of classification of affordance classes. The accuracy using our interaction descriptor is 0.92, of our spatio-temporal relationship descriptor 0.95 while the appearance baseline has an average accuracy of 0.64. Finally performing classification with the interaction and relationship descriptors jointly our model achieves an accuracy of 0.97.

II. DATA SET

Given RGB-D data from the scene we wish to cluster the information such that elements corresponding to locations on the same object are merged together within the same cluster or object hypothesis. Our approach is presented in [5]. The relevant information used to extract our functional descriptors are a set of object hypothesis tracks including object mask, 3D position of the centroid, size and bounding box.

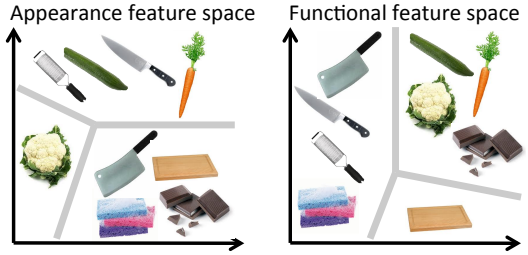


Fig. 1. Object classification based on appearance and affordance.

III. INTERACTION DESCRIPTOR

The position of human hands to extract the interaction descriptor is acquired by the Kinect device. We let the temporal signature of the relative position between the hands and the object constitute our object representation. Rather than using a continuous state-space we discretize the space into five different states (idle, close to hand, active, approach, leave) and represent each object hypothesis as a sequence of states over time (Fig. 2). We will refer to this as a string.

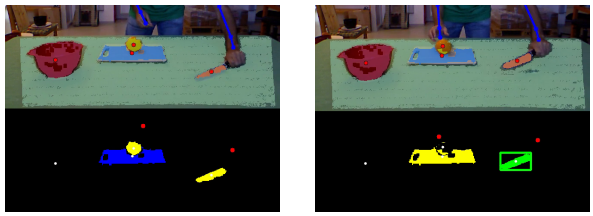


Fig. 2. Example of extraction of hand-objects interaction features. The knife is first close to a human hand (yellow color) and then it becomes active (green) when the human uses it.

IV. SPATIO-TEMPORAL RELATIONSHIP DESCRIPTOR

Given the set of object hypotheses: let $C^{\text{hyp}}_{i,t}$ be the centroid of object hypothesis i in the image at time t. The Euclidean distance between two object hypotheses (Fig. 3) at time t is:

$$d_{i,j,t} = \| C^{\text{hyp}}_{i,t} - C^{\text{hyp}}_{j,t} \| \quad (1)$$

The relationship between two object hypotheses can be described as the object-object distance profile defined by Eq.1.

V. FEATURE COMPARISON

The performances of the proposed descriptors are compared to a baseline using Support Vector Machines (SVM). Each SVM's parameters are tuned with the help of the

This research has been supported by the EU through TOMSY, IST-FP7-Collaborative Project-270436, and the Swedish Research Council (VR).

The authors are with CVAP/CAS, KTH, Stockholm, Sweden, pieropan,chek,hedvig@kth.se.

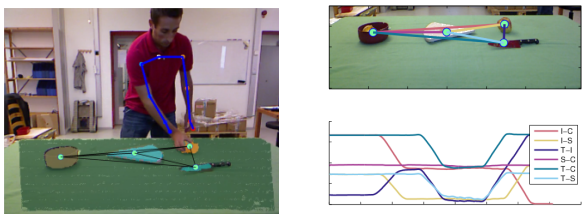
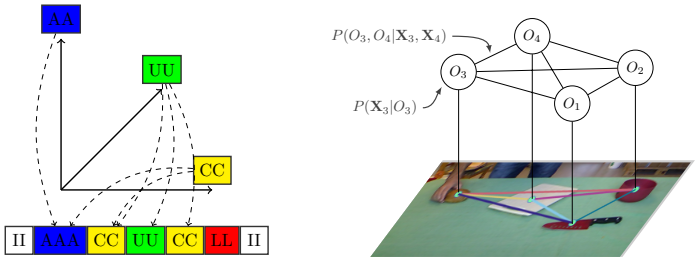


Fig. 3. Extraction of object-object spatial relationships superimposed on the segmentation video. On the right relationships between different functional classes of objects are shown in different colors.



(a) hand-object interaction (b) Spatio-temporal relationship

Fig. 4. (a) describes the feature space induced by the string kernel used to describe the features extracted (b) shows the graphical model used to infer the correct functional classes using as an input the confidence values produced by a SVM that classifies object-object spatial relationships.

validation set. In the case of the interaction feature, each object descriptor is a string of symbols (Fig. 4(a)), therefore the the similarity between two object strings x1 and x2 can be measured using the string kernel [6]. The pairwise similarities using this measure are computed between all pairs of object hypotheses in the data and it is used with a SVM to perform the classification. In the case of the spatio-temporal relationships, the pairwise relationships of objects are first classified according to the classes of the two objects involved using a SVM. The similarity between two sequences of continuous data such as the pairwise distance is measured with the recently proposed path kernel [7]. This kernel specifies an inner product between two sequences by summarizing the characteristics of all possible alignments. As such it should respect both the temporal stretch/compression and alteration in execution speed that our data exhibits. The output of the classifier are multi-class confidence values that represent the joint probability $P(O_i, O_j|X_i, X_j)$. Assuming that the relationships of each pair of objects involved in an activity are dependent, we use the output of the classifier in a graphical model to infer the functional class of the objects O_iO_j as in Fig. 4(b). The details of the proposed model are described in our recent work [8].

The classification results using the different features are shown in Table I. The average classification accuracy of the appearance descriptor is 0.64 with a high classification rate of 0.84 due to the low variation in appearance for the container class. The average classification of functional classes with our descriptors is respectively 0.92 using the manipulation descriptor and 0.95 using the spatial relationships descriptor. It is interesting to notice that the spatial descriptor outperforms the manipulation feature in 3 out of 4 classes, but the manipulation feature has a higher accuracy in recognizing the *tool* functional class. This meets our expectation as the tool class contains the set of objects that are handled more often by the human, making the manipulation descriptor very representative. In our latest experiments we use the confidence values of the classification performed with the interaction descriptor to model the unary term $P(X_i|O_i)$ shown in Fig. 4(b). This allows to perform classification using both proposed features achieving an average precision of 0.97 as shown in Table I.

VI. Conclusions

We propose two functional descriptors for objects to reason about human activities. Objects are represented in terms of how they are being handled and the spatio-temporal relationships with other objects present on a scene. The proposed descriptors have an average accuracy of 0.92 and 0.95 outperfomring a standard classifier based on appearance features. This support our idea that affordance can be capture by looking at how objects are being used during an activity. Moreover experiments have shown that the manipulation descriptor performs the best with objects that are manipulated often (e.g. tools) while the pairwise performs better in extracting the affordance of objects that are rarely used directly by a human but still involved in the undergoing activity.

References

[1] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor. Learning object affordances: From sensory motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26, 2008.

[2] J. J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1979.

[3] E. E. Aksoy, A. Abramov, F. Wörgötter, and B. Dellen. Categorizing object-action relations from semantic scene graphs. In *ICRA*, 2010.

[4] H. Grabner, J. Gall, and L. van Gool. What makes a chair a chair? In *CVPR*, 2011.

[5] A. Pieropan, C. H. Ek, and H. Kjellström. Functional object descriptors for human activity modeling. In *ICRA*, 2013.

[6] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

[7] A. Baisero, F. T. Pokorny, D. Kragic, and C. H. Ek. The path kernel. *International Conference on Pattern Recognition Applications and Methods*, 2013.

[8] A. Pieropan, C. H. Ek, and H. Kjellström. Recognizing object affordances in terms of spatio-temporal object-object relationships. In *International Conference on Humanoid Robots*, 2014.

TABLE I

Classification of object affordance using appearance features (SIFT), hand-object interaction feature (H-O), pairwise object interacion feature (O-O) and both proposed features together.

	Tools	Ingredients	Support	Containers
Sift	0,60	0,52	0,62	0,84
H-O	0.93	0.88	0.90	0.96
O-O	0.91	0.92	0.99	0.99
Joint	**0.96**	**0.94**	**0.99**	**0.99**