Recognizing Object Affordances in Terms of Spatio-Temporal Object-Object Relationships

Alessandro Pieropan Carl Henrik Ek Hedvig Kjellström

***Abstract*— In this paper we describe a probabilistic framework that models the interaction between multiple objects in a scene. We present a spatio-temporal feature encoding pairwise interactions between each object in the scene. By the use of a kernel representation we embed object interactions in a vector space which allows us to define a metric comparing interactions of different temporal extent. Using this metric we define a probabilistic model which allows us to represent and extract the affordances of individual objects based on the structure of their interaction. In this paper we focus on the presented pairwise relationships but the model can naturally be extended to incorporate additional cues related to a single object or multiple objects. We compare our approach with traditional kernel approaches and show a significant improvement.**

I. INTRODUCTION

Reasoning about activities is an important skill for robots that operate in unstructured environments. 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 [1]. By imitation, the robot should reduce the search space of possible actions only to those that involve the same objects and effect on the environment. This is denoted as robot learning from demonstration or imitation learning [2]. One important aspect of this challenging problem is to detect and reason about objects in terms of affordances [3] or alternatively, about their function in the current activity. In the following, we will use the terms affordance and functionality interchangeably, meaning the aspect of an object relating to its function in the present activity.

In this paper we represent objects directly in terms of their functionality. We argue that the spatio-temporal relationships between objects involved in an activity can encode this information. In Fig. 1 an example of describing this notion is shown. The scenario we focus on in this paper is a human demonstrator teaching a robot about the affordances of objects by showing how they are used. To that end we will, in line with previous work [4], assume that the human is responsible for all the movement in the scene. Furthermore, we assume that the relationships of each pair of objects involved in an activity are dependent, and use a graphical model to model correlation between all object-object interactions, in order to improve the recognition of functional classes of all objects in the scene and mitigate misleading information. This is in the spirit of recent studies in cognition, which show that human perception of objects depends both on previous experience and spatio-temporal contextual knowledge [5], [6]. Humans are used to see objects in context of its environment, and thus, e.g., expect to see cars on a street but not boats. This mechanism helps to identify unknown objects and their purpose given what a human already knows of the environment [7]. In our graph model, object affordances that have been well identified in the image can help resolving other, less apparent ones. In the example of a kitchen scenario, seeing a zucchini and a cutting board in close interaction, the model can infer that the object in the humans hand is most likely a tool (knife) based on the context of the other two.

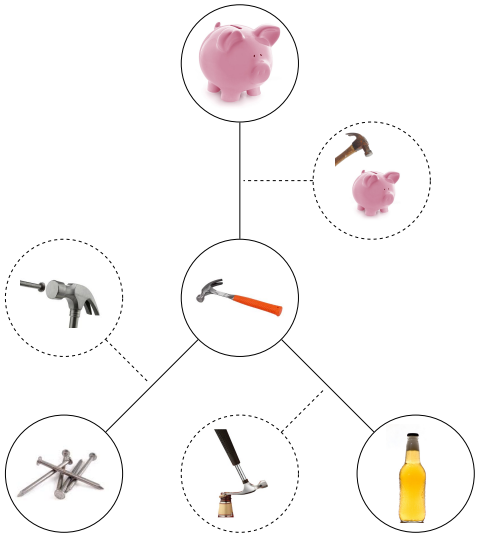


Fig. 1. Object functionality is to a high degree defined by the interaction with other objects. In this toy example, the functionality of a hammer depends highly on the context in which it is used. Together with a nail, the hammer affords hammering (the activity it is designed for). However, together with a beer bottle, the hammer also affords opening. Furthermore, together with a piggy bank, the hammer affords breaking. These three affordances are conceptually different, and tied to the other object that interacts with the hammer. We thus propose to represent object affordances in terms of object-object relationships.

The remainder of the paper is organized as follows; before we describe our method in Section III and used features in Section IV we will describe the related work in Section II. In Section V we will show the experimental evaluation of our approach before we conclude the paper in Section VI.

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.

II. Related Work

When reasoning about the environment in terms of actions and objects the concept of affordances [3] provide an attractive basis. To that end affordances have been an important focus for robotics research especially in scenarios taking a learning by demonstration approach. Enabling a robot to reason about objects in terms of affordances means that it somehow needs to learn how to encode affordances in terms of its sensory data. This is a very challenging task and have been the focus of much work. Object functionality can be learned by visually observing a human making use of object affordances. In [8] a Bayesian framework for recognizing objects based on contextual information from other objects, human actions and the scene is presented. In [9], human actions are used to infer object class.

Reversely, recognition of human actions can be guided by information on the objects involved. This is the approach in [10] which represents kitchen activities solely in terms of the sequence of objects in contact with the hand during the activity. Recently, [11] use information on which objects are active, i.e., in contact with the human hands, to guide activity recognition. Another line of work focus on grounding affordance to visual features. The traditional approach consists in modeling objects in terms of object appearance, extracting visual features from images or videos and training classifiers to categorize them in appearance classes [8]–[17].

However, appearance based classes do not necessarily correspond well to functional classes. A number of works are in the spirit of the method presented here. [18] proposes an active framework that uses contextual interaction between image regions to learn appearance and contextual models of regions. In [19], functional categories of image regions, such as roads and walkways, are segmented and clustered based on the behaviors of moving objects in the vicinity. [6], [7], [20] exploit the contextual associations to recognize objects, however this classification is only performed in terms of appearance without any reasoning about function or affordances. We here go one step further by directly observing object relations rather than constraining individual object observations with contextual information about the relation with other objects. Another approach is to extract affordances from a large set of 2D object views labeled with grasping points [21]–[23] . While [24] extracts grasping positions by looking at the human interacting with objects. However, grasping is a limited affordance feature and usually part of more complex affordance activities.

In this paper we wish to categorize objects in terms of their functional class from video. In order to do so we need to first be able to extract what parts of each frame in the video corresponds to objects. There has been a significant research effort within the vision community on perceiving and extracting objects from visual data [12]. A common approach consists in having a pre-defined set of objects and scan the input data to detect them. Even though this method works well in the context of object detection, it suffers in the context of activity reasoning for two main reasons. As previously argued objects of the same affordance class might have a very high variation in appearance. Further, the same object might have different functionalities depending on the context in which it is used (Fig. 1). As an example, a bottle opener is usually used to open a beer. However, if the proper tool is missing, a fork or any other tool can be used. A tool that affords opening of the bottle is required and that does not necessarily correspond well to the appearance of the object.

In our previous work [25] we propose to describe objects directly in terms of their functionality. More specifically, object hypotheses are represented in terms of interaction with human hands. This is more in line with other works where objects are described in terms of their functional parts [26], or mapping the shape with the function [27]. Moreover, [28] simulates humans performing activities using objects and cluster them into functional classes. Similar to our work, in [29]–[31] no appearance model is used rather humans are observed when interaction with a scene and their behavior is used to detect affordances in a similar manner. In [4], [32] a global representation of activities in terms of spatial relations between objects is presented. The recognition of activities is determined on a set of pre-defined relationships between segmented coherent regions. We will here propose to go one step further as our approach does not require a pre-defined notion of what relationships are important, rather we learn them directly from data. Moreover we learn pairwise object relationships and object functional classes rather than activities.

We will now proceed to describe the specific approach to functional object modelling from video we propose in this paper.

III. Methodology

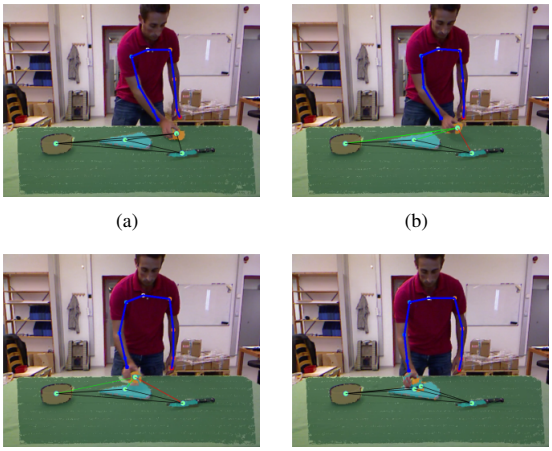
This section describes the modeling framework proposed in this paper. We take a probabilistic approach which models objects and their interactions over time in a scene. We are interested to infer an objects functionality from within this model. The approach we propose to address this problem is a two stage process. In the first step we extract object hypotheses from RGB-D measurements from the scene. Once we have acquired such hypotheses we will in a second stage model how their interaction changes over time.

A. Generating Object Hypotheses

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. In specific we will use the approach presented in [25] but will for the sake of clarity describe the method in general terms.

The basis of the analysis is a set of object hypotheses in the form of tracks of image segments over time. The actual segmentation and tracking of object hypotheses is not a focus of this paper, but is described here for completeness.

The objects extraction framework works on the assumption that at the beginning of human activities objects involved are located on horizontal support areas such as tables.



(a) (b) (c) (d)

Fig. 2. Extraction of spatial relationships superimposed on the segmentation video. (a) Initial frame of the sequence. (b) The orange is moved to the cutting board, the relative position to it is decreasing while the distance from the knife is increasing. (c) The orange is close to the cutting board. (d) The object is stationary.

Therefore, as suggested by [33], the method finds horizontal support areas at time t_0 of any video sequence. Object hypotheses are then detected using a connected component algorithm [34] that finds segments with color and/or disparity deviating from the planes. HSV histograms are then extracted from each segment. Histograms are used to extract the foreground mask of objects at each time step $t > 0$. A pyramid KLT tracker [35] is used to detect and track object masks over time.

Using this approach, a set of object hypothesis tracks is extracted. It includes object mask, 3D position, size and bounding box.

B. Object Modelling

Given the object hypotheses generated in the previous step we will now describe a probabilistic model encoding the relationship between each object, as illustrated in Fig. 2. As we are interested in determining the functionality of an object we make the assumption that each object can be fully represented in terms of its functional class. In specific, given object hypothesis i we will refer to measurements from the RGB-D data as $\mathbf{X}_i$. We will then assume that these measurements can be explained by the discrete variable O_i corresponding to the objects functionality which leads to the likelihood $p(\mathbf{X}_i|O_i)$. This term can therefore be used to describe the relationship between appearance and functionality of the object. Our main thesis is that the functionality of an object can be reliably determined based on how an object interacts with other objects in the scene. To that end we will use a fully connected undirected graph where each object is connected to every other object. We model the interaction between object i and j using the term $p(O_i, O_j|\mathbf{X}_i, \mathbf{X}_j)$. Fig. 3 shows the graphical model representation of the proposed method. The model corresponds to a fully connected Conditional Random Field (CRF) where the maximum clique

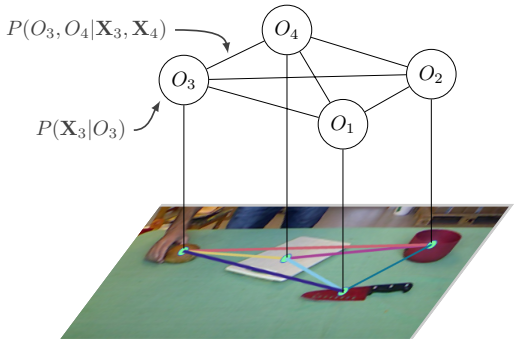


Fig. 3. Illustration of the joint functional object classification. Each node in the graphical model corresponds to an object's functional class. Each edge represents the conditional probability of the functional class given the observed objects.

size is the whole graph. In Section IV we will describe the specific form the terms $p(\mathbf{X}_i|O_i)$ and $p(O_i, O_j|\mathbf{X}_i, \mathbf{X}_j)$ takes in our experimental setting.

We are interested in extracting the functional class of the objects by observing the scene. In order to do so we wish to find the MAP solution by maximizing the posterior distribution $p(\mathscr{O}|\mathscr{X})$, where $\mathscr{X} = \{\mathbf{X}_i\}_{i=1}^N$ and $\mathscr{O} = \{O_i\}_{i=1}^N$. The Hammersly-Clifford theorem [36] states that this distribution can be expressed in terms of a product of non-negative potential functions over the maximum cliques in the graph. In this paper we will follow the approach from [37] and describe this in terms of its Gibbs distribution,

$$p(\mathscr{O}|\mathscr{X}) = \frac{1}{Z(\mathscr{X})} \exp(-\sum_{c \in C_G} \psi_c(\mathscr{O}_c|\mathscr{X}), \qquad (1)$$

where C_G is the set of maximum cliques in the graph and ψ_c is the potential function over clique c. Given that our CRF is fully connected the set of maximum cliques only contains one element, the whole graph. The normalization term $\frac{1}{Z(\mathscr{X})}$ is known as the partitioning function and is intractable to compute in the general case. However, as we are only interested in a point estimate, the MAP solution, this normalization can be ignored. To simplify computations it is therefore common practice to work in log-space with the energy rather than the on posterior directly. For a pairwise CRF the potential function ψ_C can be decomposed into two separate terms one which depends on a single node and one on a pair. This lead to the following energy function,

$$E(\mathscr{O}) = \sum_{i=1}^{N} \left(\phi_(O_i|\mathbf{X}_i) + \sum_{j \in N(i)} \phi_p(O_i, O_j|\mathbf{X}_i, \mathbf{X}_j) \right) \qquad (2)$$

where $\phi_u(O_i|\mathbf{X}_i)$ and $\phi_p(O_i, O_j|\mathbf{X}_i, \mathbf{X}_j)$ is the decomposition of the potential function into the unary and the pairwise potential of the energy and $N(i)$ is the set of nodes that shares an edge with node i. We now wish to find the MAP solution $\hat{\mathscr{O}} = \text{argmax}_{\mathscr{O}} E(\mathscr{O})$. This is, for the general case, an NP-hard problem. To proceed we will therefore approximate the true posterior with a combination of tree structured distributions in the spirit of [38]. In specific, we will average the solution of all possible trees generated from the graph, see Fig. 4.

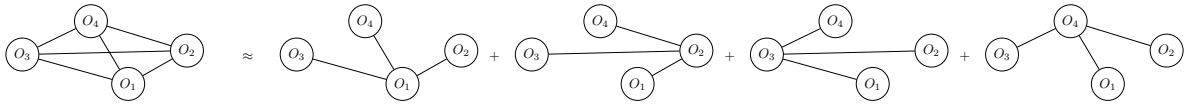


Fig. 4. Illustration of the proposed approximation. All possible trees are constructed from the fully connected graphs. The posterior is computed for each tree and the result is averaged.

The number of trees in the graph will grow rapidly with number of objects raising questions about the scalability of the suggested approach. However, fully connected CRFs have been applied to images [37] where the number of nodes are orders of magnitudes higher than in any perceivable robotic scenario. Further, we use a fully connected graph because we do not want to make any assumptions about the interaction of the objects however by pruning edges between objects that are too far away from each other for any interaction to take place the number of trees can be reduces significantly.

We will now proceed to describe how the terms in the energy in Eq. (2) is formulated.

IV. Representation

The Gibbs energy of the CRF model consists of two separate terms; the unary potentials $\phi_u(O_i|\mathbf{X}_i)$ and the pairwise potentials $\phi_p(O_i,O_j|\mathbf{X}_i,\mathbf{X}_j)$. The unary potential can be used to relate the appearance of the object to its functional class while the pairwise potential encodes the relationship between two objects. Our argument in this paper is that an objects functionality is best described by an objects interaction with others rather than its appearance. To that end we use a constant unary potential which implies that given just a single object hypothesis the functional classes are equally likely. This means that only the pairwise relation between objects are responsible for determining an objects functional class. We will now describe how we can encode interaction using the pairwise potentials.

We argue that the functional characteristics of an object should be represented in terms of the "context" to the other objects in the scene. Given a set of object hypotheses, we do so by representing each object in terms of the spatio-temporal relationships with each other object in the scene. In specific, a video of a manipulation activity is segmented into object hypotheses, as described in [25]. Let $C_i(t)$ be the centroid of object hypothesis i in the image at time t. The relationship between each object will be represented as a the 3D Euclidean distance between the centroids over time,

$$d_{i,j}(t) = \| C_i(t) - C_j(t) \|. \tag{3}$$

The signal extracted is noisy as a consequence of errors introduced in the segmentation, e.g., from object occlusion. To model major fluctuations in distance that correspond to possible change in state of the corresponding object, the signal $d_{i,j,t}$ is smoothed with a moving average low-pass filter to remove high-frequency variation.

Having extracted the tempo-spatial signature described above we need to relate this to the functional class of the objects. From Fig. 5 we can see that the feature does discriminate between the classes. However, as the signals are of different length it is not obvious how to compare two signals. The most obvious approach would be to compare a distribution of local structures in the signals through a bag-of-words descriptor [39]. However, looking at the signals in Fig. 5 the order of the changes are important why we believe such an approach would fail. Another approach would be to subsample the sequences to the same length, however this will remove the information about absolute duration which might also be important. Rather than creating an explicit feature space for the signals we will take a kernel approach and through an inner-product specify an implicit representation for the sequences. We will apply the recently proposed Path Kernel [40] to perform the embedding. The path kernel is motivated by Dynamic time warping [41] and defines an inner product between sequences as a mixture of all possible alignments of two signals. Given the resulting feature space we apply a Support Vector Machine (SVM) [42] trained on the object-object distance profiles. New observations of the distance over time between two objects O_i and O_j can then be classified using the trained SVM. The output of the classifier are multi-class confidence values [43] that represent the joint probability $P(O_i,O_j|\mathbf{X}_i,\mathbf{X}_j)$. The pairwise potential function is then the negative log of this conditional probability.

V. Experiments

Both the classification of pairwise relationships described in Sec. IV, and the inference of the object states from those relationships as described in Sec. III, are evaluated on a dataset described below.

A. Dataset Description

We use a dataset of indoor human activities presented in our previous work [25] for our spatial relationship analysis. Fig. 6 shows examples of pairwise relationships in the dataset.

There are 40 RGB-D videos in the dataset. Each video is labeled with object hypotheses in the form of coherent regions tracked over time, as described in [25]. In addition, the data set also includes the centroids and bounding boxes of these regions. All object hypotheses in the dataset are manually labeled with one of four distinct functional classes: **Tools** are used to perform activities upon other objects, **Ingredients** are the target on which the tools are used. **Support Areas** are objects that support the ongoing activity. **Containers** are used to contain other objects. The dataset includes 4 different tools (knives), 9 ingredients, 3 suppor area (cutting boards) and 4 containers. The objects are

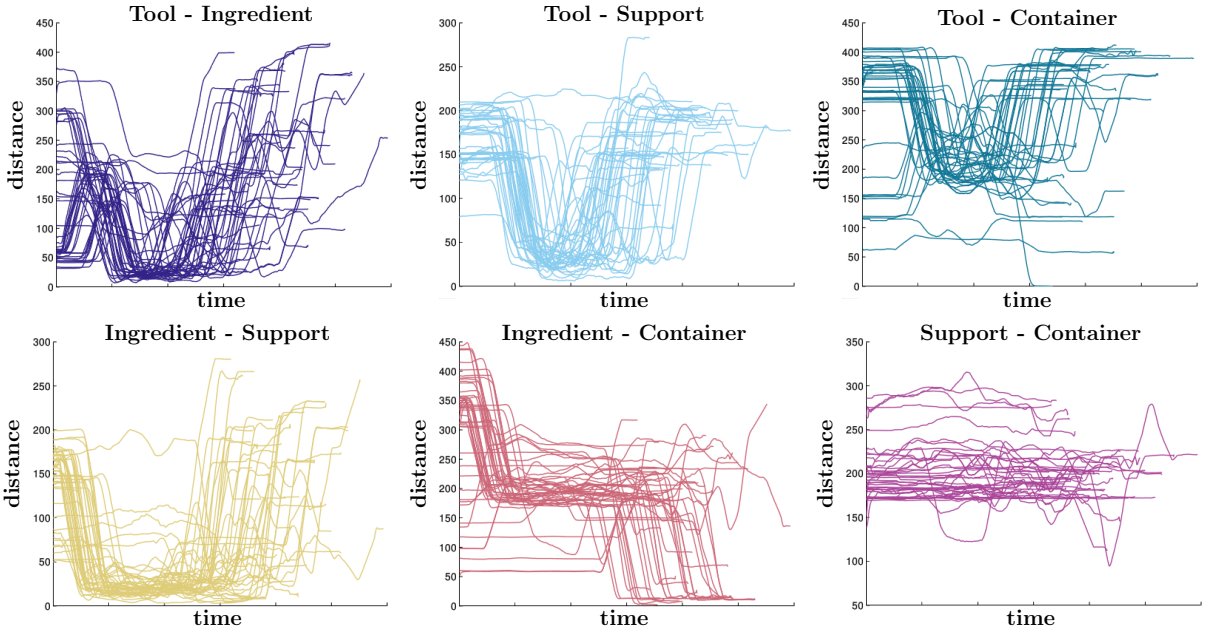


Fig. 5. Overview of 6 classes of pairwise features extracted from the dataset described in Sec. V-A. The features are divided by the pairwise functional class interaction. This figure is best viewed in color. The diagrams show the variation of the distance (y axis) between two objects over time (x axis).

located in different places relative to each other in different video sequences (Fig. 6).

B. Classification of pairwise relationships

From each video, the pairwise object distances are extracted as described in Sec. IV, with the time-span $T = 70$. The curves are then subsampled with a factor of 8 for computational reasons, and are resampled so that all curves have the same length. This enables comparisons of the proposed path kernel with regular linear and rbf kernels.

The dataset of distance curves is randomly divided into a training set of curves from 20 videos, a validating set of curves from 5 videos and a test set of curves from 15 videos. Each video is present in only one set. Curves are labeled into the following classes: **Tool-Tool**,**Tool-Ingredient**, **Tool-Support Area**, **Tool-Container**, **Ingredient-Ingredient**, **Ingredient-Support Area**, **Ingredient-Container**, **Support Area-Support Area**, **Support Area-Container** and **Container-Container**, using the labels of the two objects connect by the curve. We then train an SVM with the training set as described in Sec. IV, tuning parameters with the help of the validation set. No hard classification decisions are made, and the classification results are on the form of probability distributions $P(O_i, O_j)$ for each combination of objects i and j occurring in the same video.

Fig. 7(c) shows the confusion matrix for classification of pairwise object relationships using the SVM with path kernel. The overall result, 85%, is lowered by the results for same-object relations. The reason for the low accuracy on these four relation classes are a lack of training data: there were very few examples of **Tool-Tool**, **Ingredient-Ingredient**, **Support Area-Support Area**, and **Container-Container** examples in the dataset.

Comparing to the two baselines of SVM using a linear kernel (Fig. 7(a)) and rbf kernel (Fig. 7(b)), it is evident that the path kernel is more suitable to represent similarities between object distance curves. All experiments are performed using the best parameter setting found using a standard cross validation technique.

C. Functional object classification

The classification results in Fig. 7(c) indicate that that there is uncertainty in the classification of the pairwise features. However, the model presented in Sec. III-B can be expected to improve the functional classification of individual objects, since the estimates over pairwise features support each other. This is experimentally verified below.

The joint MAP estimate $\arg\max P(O_1, O_2, ..., O_n)$ is computed for each test video from the set of observed pairwise features in the same video, found as described in the previous section. This gives an estimate of the functional class of each object in every video. By comparing to the ground truth labels, the average accuracy of the classification is found to be 0.96 (Table I). This result shows that the spatio-temporal configuration of objects is a good feature to classify affordance of objects, and that the principled combination of individual object-object interaction classifications makes the model robust to failures in some of the object-object interaction classifications.

VI. Conclusions

We have proposed a functional descriptor for objects to reason about human activities using object context. Objects are represented in terms of their spatio-temporal relationships with other objects. The pairwise relations are combined using

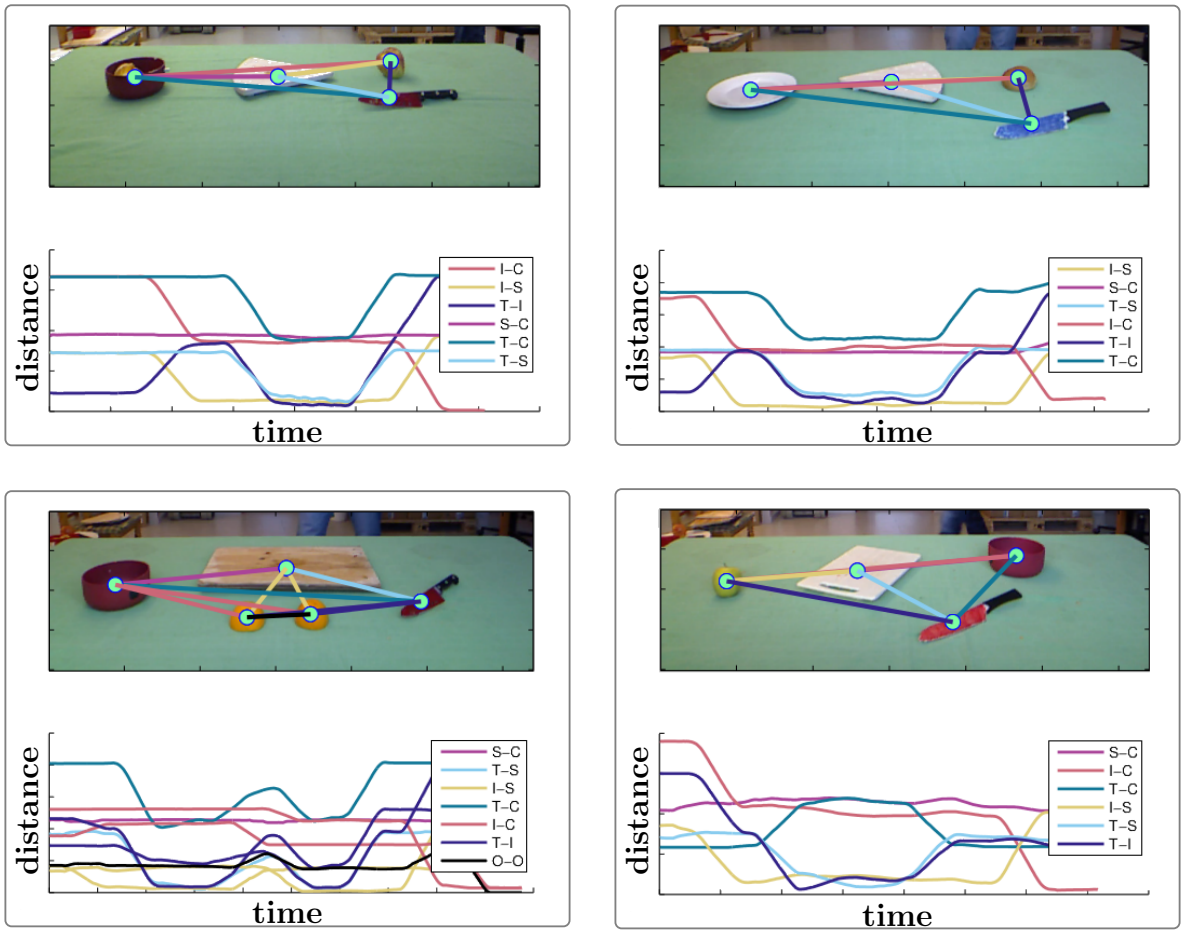
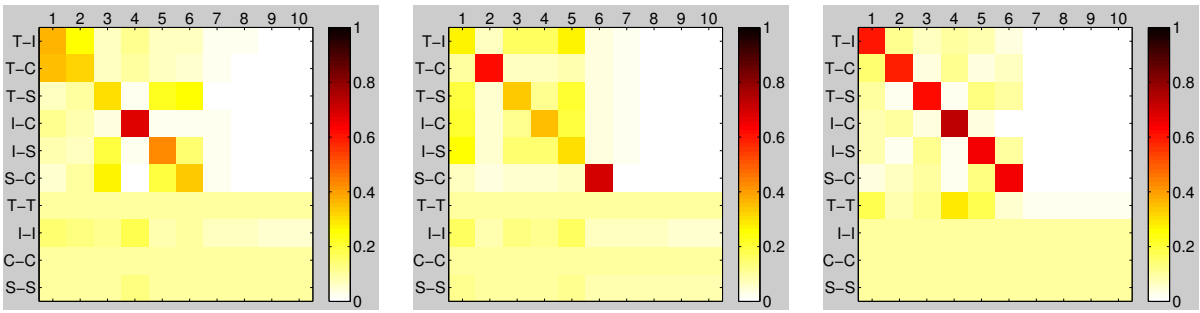


Fig. 6. Example of spatial relationship of object hypotheses in a video. Each colored curve represent the spatio-temporal relationship of two objects. The features are labeled according to the functional class of the objects involved. Ingredient-Container(**I-C**), Ingredient-Support Area(**I-S**), Tool-Ingredient(**T-I**), Support Area-Container(**S-C**), Tool-Container(**T-C**), Tool-Support Area(**I-C**). Object-Object(**O-O**) shows a relationship between two objects of the same functional class. This figure is best viewed in color.



(a) linear kernel (b) rbf kernel (c) path kernel

Fig. 7. Classification of pairwise object relationships. (a) Classification using SVM with linear kernel. (b) Classification using SVM with rbf kernel. (c) Classification using SVM with path kernel.

a graphical model that formulates the joint probability of functional classes of all objects in the scene. In a real world scenario the proposed method results in an average accuracy of 96% when jointly optimized compared to the significantly lower accuracy of 85% for the classification of the individual pairwise relations. This support our idea that contextual knowledge improves recognition of objects, in the spirit of recent studies in cognitive psychology [44]. The

TABLE I

CLASSIFICATION OF FUNCTIONAL CLASSES PERFORMED USING OUR GRAPHICAL MODEL. AVERAGE ACCURACY IS 0.96.

	Tools	Ingredients	Support	Containers
T	**0.91**	0.04	0.03	0.02
I	0.02	**0.92**	0.04	0.02
S	0.00	0.01	**0.99**	0.00
C	0.00	0.00	0.01	**0.99**

feature proposed not only performs better than our previous work on the same dataset [25] but has the advantage to work on continuous data without the need of a predefined vocabulary or manually tuned mechanism to define the state space.

The proposed method is flexible, robust and can be extended in a number of different ways. More nodes can be added, representing other types of contextual knowledge that humans exploit to recognize, human interaction with objects [25] can be modeled as a separate node connected to all objects. Object classifiers, e.g., [16] can be integrated by to provide a likelihood of each object. In the spirit of studies in cognitive psychology, contextual knowledge of the environment can be modeled to constraint expected objects, pairwise relationships and probable activities. With an accuracy of 96% we will in future work record a new and more challenging dataset with a broader set of activities to investigate the boundaries of our present method.

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] S. Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3:233–242, 1999.

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

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

[5] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

[6] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007.

[7] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, 2003.

[8] D. J. Moore, I. A. Essa, and M. H. Hayes. Exploiting human actions and object context for recognition tasks. In *ICCV*, 1999.

[9] P. Peursum, G. West, and S. Venkatesh. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In *ICCV*, 2005.

[10] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A scalable approach to activity recognition based on object use. In *ICCV*, 2007.

[11] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[12] L. Fei-Fei, R. Fergus, and A. Torralba. *Recognizing and Learning Object Categories: Short course at ICCV*, 2009.

[13] R. Filipovych and E. Ribeiro. Recognizing primitive interactions by exploring actor-object states. In *CVPR*, 2008.

[14] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE PAMI*, 31(10):1775–1789, 2009.

[15] H. Kjellström, J. Romero, and D. Kragic. Visual object-action recognition: Inferring object affordances from human demonstration. *CVIU*, 115(1):81–90, 2011.

[16] M. Madry, C. H. Ek, R. Detry, K. Hang, and D. Kragic. Improving generalization for 3d object categorization with global structure histograms. In *IROS*, 2012.

[17] M. Veloso, F. von Hundelshausen, and P. E. Rybski. Learning visual object definitions by observing human activities. In *IEEE-RAS International Conference on Humanoid Robots*, 2005.

[18] B. Siddiquie and A. Gupta. Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In *CVPR*, 2010.

[19] M. W. Turek, A. Hoggs, and R. Collins. Unsupervised learning of functional categories in video scenes. In *ECCV*, 2010.

[20] M. J. Choi, A. Torralba, and A. S. Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012.

[21] J. Bohg and D. Kragic. Grasping familiar objects using shape context. In *International Conference on Advanced Robotics*, 2009.

[22] M. Madry, D. Song, and D. Kragic. From object categories to grasp transfer using probabilistic reasoning. In *ICRA*, 2012.

[23] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic grasping of novel objects using vision. *International Journal of Robotics Research*, 27(2):157–173, 2008.

[24] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele. Functional object class detection based on learned affordance cues. In *International Conference on Computer Vision Systems*, 2008.

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

[26] E. Rivlin, S. J. Dickinson, and A. Rosenfeld. Recognition by functional parts. *CVIU*, 62(2):164–176, 1995.

[27] L. Stark and K. Bowyer. *Generic Object Recognition using Form and Function*. World Scientific Series in Machine Perception and Artificial Intelligence – Vol. 10, 1996.

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

[29] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3D scene geometry to human workspace. In *CVPR*, 2011.

[30] B. Moldovan, P. Moreno, M. van Otterlo, J. Santos-Victor, and L. De Raedt. Learning relational affordance models for robots in multi-object manipulation tasks. In *ICRA*, 2012.

[31] J. Gall, A. Fossati, and L. van Gool. Functional categorization of objects using real-time markerless motion capture. In *CVPR*, 2011.

[32] G. Luo, N. Bergström, C. H. Ek, and D. Kragic. Representing actions with kernels. In *IROS*, 2011.

[33] L. Nalpantidis, G. Sirakoulis, and A. Gasteratos. Non-probabilistic cellular automata-enhanced stereo vision simultaneous localization and mapping. *Measurement Science and Technology*, 22(11), 2011.

[34] L. Shapiro and G. Stockman. Connected components labeling. In Prentice Hall, editor, *Computer Vision*, chapter 3, pages 69–73. Prentice Hall, 2002.

[35] J.-Y. Bouguet. Pyramidal implementation of the Lucas Kanade Feature Tracker, description of the algorithm, 2000.

[36] John M Hammersley and Peter Clifford. Markov fields on finite graphs and lattices. Technical report, 1971.

[37] Philipp Krähenbül and Vladlen Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Advances in Neural Information Processing Systems*. NIPS, 2011.

[38] Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. In *International Conference on Airtificial Inteligence and Statistical Learning*. Citeseer, 2003.

[39] Harris, Zellig Sabbettai. Distributional structure. *Word*, 10:146–162, 1954.

[40] Andrea Baisero, Florian T Pokorny, Danica Kragic, and Carl Henrik Ek. The Path Kernel. In *International Conference on Pattern Recognition Applications and Methods*, February 2013.

[41] H Sakoe and S Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[42] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[43] A. Pronobis and B. Caputo. Confidence-based cue integration for visual place recognition. In *IROS*, 2007.

[44] N. Gronau, M. Neta, and M. Bar. Integrated contextual representation for objects' identities and their locations. *Journal of Cognitive Neuroscience*, 20(3):371–88, 2008.