Discovering Object Classes from Activities

Abhilash Srikantha[1,2] and Juergen Gall[1]

[1] University of Bonn
[2] MPI for Intelligent Systems, Tuebingen
abhilash.srikantha@tue.mpg.de, gall@informatik.uni-bonn.de

Abstract. In order to avoid an expensive manual labelling process or to learn object classes autonomously without human intervention, object discovery techniques have been proposed that extract visually similar objects from weakly labelled videos. However, the problem of discovering small or medium sized objects is largely unexplored. We observe that videos with activities involving human-object interactions can serve as weakly labelled data for such cases. Since neither object appearance nor motion is distinct enough to discover objects in such videos, we propose a framework that samples from a space of algorithms and their parameters to extract sequences of object proposals. Furthermore, we model similarity of objects based on appearance and functionality, which is derived from human and object motion. We show that functionality is an important cue for discovering objects from activities and demonstrate the generality of the model on three challenging RGB-D and RGB datasets.

Keywords: Object Discovery, Human-Object Interaction, RGBD Videos.

1 Introduction

Approaches for object detection require a fair amount of annotated images in order to perform well [10]. Contemporary solutions such as crowdsourcing will be suboptimal in the long run due to high costs involved. As a result, there has been a recent shift of focus towards utilizing readily available weakly labelled data [2,6,26,36,40], particularly videos [27,33,34]. The fundamental assumption in all these approaches is that the object of interest is dominant and can be easily segmented. In other words, motion or appearance of the object are assumed to be distinct from the background. This is a valid constraint for large active objects such as moving vehicles or animals and is further aided by object- or action-centric nature of labelled videos on the Internet.

Moving away from commonly used categories such as airplanes, boats, cars, cats, horses etc., we propose to work on small and medium sized object categories such as pens and mugs that are used in daily routine. Weakly supervised learning in such areas is largely unexplored inspite of their obvious impact on applications in robotics, assisted living etc. One reason is the scarcity of data because such objects do not form popular subjects for generating and sharing videoclips. However, videos labelled with the context of human activity, like drinking or

D. Fleet et al. (Eds.): ECCV 2014, Part VI, LNCS 8694, pp. 415–430, 2014.

writing, are available in plenty. These videos, however, violate the fundamental assumption since the dominant subjects here are humans, their body parts and their immediate environment; thus forcing present day methods to failure as verified in our experiments. Another problem of existing methods for such videos is the assumption that similarity of objects is completely defined by their appearance features inspite of their frequent occlusion and low resolutions. Also, appearance and pose of objects change within a single video due to the human-object interaction involved. For these reasons, appearance-only approaches are limited for mining such objects as verified in our experiments.

We therefore propose an approach that addresses the problem of weakly supervised learning for medium or small sized objects from action videos where humans interact with them. Since existing methods fail for this task, we introduce a novel method consisting of two parts as illustrated in Figure 1. The first part addresses the problem that objects cannot be segmented by searching for dominant motion segments. Instead, we track randomly selected superpixels to generate many tubes per video as object candidates as illustrated in Figure 2. To this end, we do not rely on a single algorithm with a single parameter setting due to the variety in objects but sample from a space of algorithms and their parameters. We condition the sampling on human pose in order to make it more efficient. The second part addresses the problem that similarity of generated tubes is not well described by appearance features alone. We therefore propose a similarity measure that not only includes appearance and size but also encodes functionality of the object derived from relative human-object motion during the activity.

To demonstrate the generalization capabilities of our approach, we evaluate on three challenging datasets, namely two RGB-D datasets [16, 25] and one RGB dataset [35]. The datasets have been recorded with three different types of sensors: a time-of-flight camera [16], a structured light camera [25], and a color camera [35]. The quality of 3d or 2d pose information also differs greatly since it is automatically extracted with different methods. On all three datasets, we show that our approach is suitable to discover objects from videos of activities and investigate the importance of functionality in the current setup.

2 Related Work

Unsupervised object discovery in images [26, 39] or videos [37] aims at finding similar objects in a set of unlabelled visual data. In many cases, weak label information is available and can be used. For instance, images with an object class label can be collected thereby reducing the problem to identifying instances that co-occur in images and localizing them either through bounding boxes or segmentation masks [2, 6, 36, 40].

There are a few works that exploit weakly labelled videos for learning [27, 30, 33, 34]. The element these approaches have in common is that they strongly rely on motion in videos and often assume that deformation in objects is either rigid or articulated i.e. can be approximated by rigid parts. For instance, part-based

models of animals are learned from videos and applied for detection in [34]. [30] uses a structure-from-motion approach to discover objects that share similar trajectories. Videos that are not necessarily related to an object class are used to learn features for object detection that are robust to temporal deformations in [27]. In [33] videos with object labels are utilized to generate training data for object detection. While the approach focuses on objects that can be relatively easily discovered and segmented in videos, our approach deals with medium and small sized objects that do not move at all or move only when they are used by a human. This makes it very difficult for them to be discovered in videos via conventional motion or appearance based methods. One can, however, exploit human motion as additional cue.

The idea of using human motion for scene understanding has recently gained attention [8, 15, 16, 18, 20, 21, 23, 25, 31, 32, 38] due to progress in human pose estimation and availability of commercial SDKs for depth data. In [31, 38], human trajectories in office environments or street scenes are extracted to segment image regions based on observed human behaviors. While the benefit of combining object detection and action recognition have been investigated in several works e.g. [13, 19, 29], the works [16, 23, 25, 32] focus on affordance cues that can be used for higher level video understanding e.g. action recognition. [23] learns relations between objects and their functionality to improve object detection or activity recognition. In [16], human motion is used to cluster objects of similar functionality in an unsupervised fashion. Further, descriptors learned for object functionality from hand-object interactions are applied to human activity recognition in [32]. The joint learning of activities and object affordances is addressed in [25]. Extracting object instances in egocentric videos using appearance-only cues and weak action-object labels using a framework that is made robust by incorporating motion information is dealt in [11].

Human models have also been used to hallucinate their interactions with given scenes. A detector for surfaces where humans can sit on is proposed in [18]. In this work, the sitting action is represented by a single human pose and its geometric relation to objects like chairs is learned. The approach is generalized in [21] where more relations between human poses and objects are used to label 3D scenes. In [20], a similar idea has been employed for static 2D images where the geometry of the scene is extracted. An exactly opposite approach is followed by [15] where human motion in a video is observed to extract scene geometry. Human motion has also been used for scene segmentation in [8].

3 Learning Object Models from Activities

An overview of the pipeline for discovering instances of a class in a set of RGB-D or RGB videos is illustrated in Figure 1. The input is a set of videos that is labelled with activities involving human-object interactions e.g. label *eating cereal* indicates the presence of a bowl.

To begin the pipeline, we assume that human pose either in 2d or 3d has already been extracted. This can easily be obtained from RGB-D videos using

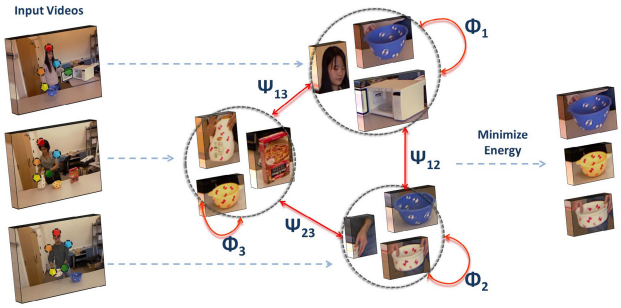


Fig. 1. Processing pipeline: Input is a set of action videos with human pose. Multiple sequences of object proposals (tubes) are generated from each video. By defining a model that encodes the similarity between tubes in terms of appearance and object functionality, instances of the common object class are discovered.

freely available SDKs, but is also straightforward for RGB videos due to the enormous progress in 2d pose estimation over the past years [35]. There are no other additional restrictions on the nature of input videos. In other words, videos may contain multiple activities, multiple persons and/or multiple objects e.g. *microwaving food, cleaning microwave* are different activities involving various objects, but they commonly feature a microwave.

In the next step, several object proposals are generated by selecting spatio-temporal regions called tubes from each video. This is modelled as a sampling process and is explained in Section 3.1. While most of these tubes will not contain the object of interest, the aim is to extract at least one tube that sufficiently overlaps with the object.

After having selected a set of tubes, we jointly select one tube per video that best describes the object. This is achieved by minimizing an energy functional built upon potentials that describe either the presence of an object in a tube or the similarity between tubes as explained in Section 3.2. As for these potentials, we employ similarity in appearance and functionality.

3.1 Generating Tubes

A straightforward way to generate tubes is to extract motion segments from videos as in [4, 33]. However, such methods do not generate meaningful tubes in the current scenario because motion is predominantly caused by entities like body parts. Instead, we extract frame based superixels in these videos and track them over time. We observed that the quality of tubes is sensitive to the method chosen and its parameters and that there is no single universal setting. We therefore consider a pool of trackers and randomly sample from it to extract tubes T_v from a video v. The probability that any tube of the video is selected therefore depends on the tracking algorithm τ and a superpixel S:

$$p(T_v) = \sum_S \sum_\tau p(T_v|\tau, S)p(\tau)p(S). \quad (1)$$

In our experiments, we use two trackers with uniform probability i.e. $p(\tau) = 0.5$. The first method uses the median optical flow [5] within the region of the superpixel to propagate it to the next or previous frame. The second method uses mean shift [7] where the RGB(D) histogram of the superpixel is used as template. While the method using dense optical flow works well for medium sized rigid objects, it easily gets distracted by fast or background motion for small objects.

Since long-term tracking is unreliable in either case, we limit the length of each tube to 300 frames or the shortest length of a video.

The superpixels S are generated using [12] which is modified to incorporate depth as feature. Similar to tracking, there is no single configuration optimal for all objects. While depth is helpful for many objects, it becomes unreliable for very small objects or reflective surfaces. We therefore compute superpixels in three different settings $\sigma \in \{RGB, D, RGBD\}$. The sampling of the superpixel also depends on the frame f in the video and a spatial prior $p(l|f)$ which depends on the frame:

$$p(S) = \sum_\sigma \sum_f \sum_l p(S|f, l, \sigma)p(l|f)p(f)p(\sigma). \quad (2)$$

For RGB videos, $p(\sigma = RGB) = 1$; otherwise $p(\sigma)$ is uniform. $p(f)$ is a prior on frames where the interaction is happening in the video. In our experiments, we use a uniform distribution i.e. we assume that the activity occurs anywhere in the video. For the spatial prior $p(l|f)$, we make use of the pose information since we are considering activities with human-object interactions. To this end, we compute the location variance of all joints within a temporal neighborhood of 15 frames and select the joint with highest variance. For RGB-D videos, we model $p(l|f)$ as a uniform distribution within the sphere centered at the joint location j at frame f and radius 400mm. Since RGB videos do not provide 3d information, we use the location of the parent joint j_p to compute the radius of the circle $\|\gamma(j - j_p)\|$ and its center $j + \gamma(j - j_p)$. In our experiments, we use $\gamma = 0.2$.

Sampling from (1) is straightforward and the tube generation process is illustrated in Figure 2. In our experiments, we sample 30 tubes T_v per video. It is important to note that we are only generating candidates at this point, the evaluation of the tubes is performed in the next step.

3.2 Joint Object Hypothesis Generation

Given a set of candidate tubes $\mathcal{T}_v$ in each video v, the goal is to select the tubes that contain the object class and are tight around the object. Similar to [9,33], this can be formulated as an energy minimization problem defined jointly over all videos N. Let $l_v \in \{1, ..., |\mathcal{T}_v|\}$ be a label that selects one tube out of a video, then the energy of all selected tubes $L = (l_1, \ldots, l_N)$ is defined as

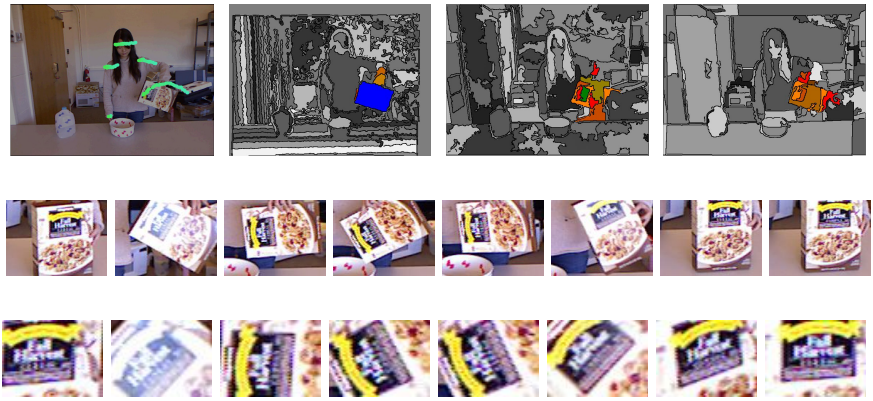


Fig. 2. Illustrating the tube generation process. The top row from left to right: The first image shows joint trajectories. The most active joint is used to compute the spatial prior for selecting superpixels. The three images next to it show three superpixel representations computed using depth (D), color (RGB) and both (RGBD). Colored superpixels are within the specified distance of the most active joint. Second and third rows visualize tubes T_v sampled from the blue and green superpixel S respectively.

$$E\left(L\right) = \sum_{v} \Phi\left(l_v\right) + \sum_{v,w} \Psi\left(l_v, l_w\right). \tag{3}$$

The unary potentials Φ measure the likelihood of a single tube being a tight fit around an object. The binary potentials Ψ measure the homogeneity in object appearance and functionality of a pair of tubes. The energy is minimized by Tree-Reweighted Message Passing [24]. While the method does not find always the global optimum, it produces satisfying results as we show in our experiments. We now describe the various potentials involved.

3.3 Unary Potentials Φ

Unary terms measure the quality of tube l_v in video v. We identify four aspects that distinguish tubes tightly bound to objects that are manipulated from the rest.

Appearance Saliency has routinely been used for object discovery since the appearance of objects is often distinct from the background. We define saliency of the k^{th} frame of a tube by the average χ^2 distance between the RGB-D or RGB distributions of the region inside each frame of the tube, I_k, and its surrounding region, S_k, which is of the same size.

$$\Phi^{app}(l_v) = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{1}{2}\sum_{i}\frac{(I_{k,i} - S_{k,i})^2}{I_{k,i} + S_{k,i}}\right) \tag{4}$$

The unary penalizes tubes that contain the object but are not tight, or tubes that cover only a part of the object. In both cases, the appearance inside and outside the tube is more similar than for a tight tube.

Pose-object Relation is useful to identify the object that is being manipulated. To this end, we measure the distance between the locally active end effector j_k and the center of the tube c_k for each frame of the tube. Depending on the data i.e. RGB or RGB-D videos, the distance is measured in 2d or 3d. Since the body does not need to be very close to the object over the entire length of the video e.g. for a microwave the contact might be very short, we perform $\alpha = 0.3$ trimmed mean filtering

$$\Phi^{Pose}(l_v) = \frac{1}{K} \sum_{k=\alpha \cdot K}^{(1-\alpha)\cdot K} \|c_{D(k)} - j_{D(k)}\| \tag{5}$$

where D is the sorted list of distances. The parameter α also makes the potential more robust to pose estimation errors.

Body part avoidance is necessary since they are dominant parts of input videos and satisfy the previous terms perfectly, hands in particular. To discourage trivial solutions such as these, we define a potential that penalizes the selection of body parts

$$\begin{aligned} \Phi^{body}(l_v) &= \max\left\{\bar{p}_{skin}(I), \bar{p}_{upper}(I), \bar{p}_{lower}(I)\right\}, \\ \text{with} \quad \bar{p}_x(I) &= \frac{1}{K} \sum_k p_x(I_k) \end{aligned} \tag{6}$$

where I_k is the color histogram of the tube at frame k. The probabilities for upper and lower body are modelled by 5-component Gaussian Mixture Models, which are learned from the video directly using the estimated pose. For skin, we use a generic model [22].

Size prior of an object is a cue that can be computed relative to human size independently of the dataset. Such priors are useful in scenarios where tubes are very small such that the other potentials become unreliable. To this end, we impose a Gaussian prior on the size of an object

$$\Phi^{size}(l_v) = \exp\left(\frac{(w_{l_v} - 2w_h)^2 + (h_{l_v} - 2h_h)^2}{2\sigma_h^2}\right) \tag{7}$$

where (w_h, h_h) and (w_{l_v}, h_{l_v}) are average width and height of the hand and tube respectively and σ_h is $0.75 \times (w_h + h_h)$.

Unary potential is formed by linearly combining the four terms as

$$\begin{aligned} \Phi(l_v) = \lambda_1 \Phi^{app}(l_v) + \lambda_2 \Phi^{pose}(l_v) \\ + \lambda_3 \Phi^{body}(l_v) + \lambda_4 \Phi^{size}(l_v) \end{aligned} \tag{8}$$

where the weighting parameters λ_i are learned from a held out validation set as explained in Section 4.

3.4 Binary Potentials Ψ

The binary term measures similarity between two tubes l_v and l_w. We use two terms in this regard. While the first term measures similarity in appearance, the second measures similarity in human motion involved during the interaction.

Shape. As in [33], we use PHoG [3] to measure the similarity between two tubes. We describe the appearance of a tube by uniformly sampling 50 frames along its temporal extent and spatially binning each frame's gradients at different resolutions. Since objects can be transformed during object manipulation, we additionally align the sequences using dynamic time warping, where we use joint locations of the head, shoulders and hands as features. Since the alignment of two very different action sequences is meaningless, we apply the warping only if the average alignment error is below a certain threshold. The $\Psi^{shape}(l_v, l_w)$ is then defined as the median χ^2 distance between PHoG features from the corresponding frames k of l_v and l_w given as

$$\Psi^{shape}(l_v, l_w) = \underset{k}{\text{median}} \left\{ \frac{1}{2} \sum_i \frac{\left(P_{\omega_v(k),i} - P_{\omega_w(k),i}\right)^2}{P_{\omega_v(k),i} + P_{\omega_w(k),i}} \right\} \tag{9}$$

where ω_u is the dynamic time warping function for tube l_u and $P_{\omega_u(k),i}$ is i^{th} bin of the PHoG feature extracted from k^{th} frame of tube l_u after warping.

Functionality. Assuming that functionality the of an object correlates with its trajectory with respect to human motion, we measure the relative distance between the center of the tube and the human. After having tubes aligned as for the shape term, we sample 50 uniformly distributed corresponding frames of both tubes. To this end, we compute the distance between the center $c_{u(k)}$ of the tube l_u at frame k and the head position $h_{u(k)}$ and normalize it by the distance between the head and the locally active end effector $j_{u(k)}$:

$$d_{u(k)} = \frac{\|h_{u(k)} - c_{u(k)}\|}{\|h_{u(k)} - j_{u(k)}\|} \tag{10}$$

The normalization is important for 2d poses, but it also compensates in 3d for different body sizes. The potential $\Psi^{func}(l_v, l_w)$ is then the median of these differences after applying the dynamic time warping functions ω_u:

$$\Psi^{func}(l_v, l_w) = \underset{k}{\text{median}} \left\{ |d_{\omega_v(k)} - d_{\omega_w(k)}| \right\} \tag{11}$$

Binary potential is formed by linearly combining the two terms as

$$\Psi\left(l_v, l_w\right) = \lambda_5 \Psi^{shape}\left(l_v, l_w\right) + \lambda_6 \Psi^{func}\left(l_v, l_w\right) \tag{12}$$

where the weighting parameters λ_i are learned together with the weights of the unary potential (8) from a validation set.

4 Experiments

To evaluate the proposed approach and demonstrate its generalization capabilities for different types of input data, we perform experiments on two RGB-D and one RGB-dataset. We show that motion segmentation, e.g. as used in [33], fails drastically for discovering objects from videos with activities and evaluate the impact of various potentials in detail. We further compare our approach to an unsupervised approach [14] and a weakly supervised approach [33] that formulates an energy functional similar to (3). For any given tube, the unary potential is composed of the objectness measure [1], shape similarity calculated as PHoG consistency and appearance similarity calculated via SIFT Bag-of-Words. The binary potential quantifies similarity between a pair of tubes by evaluating PHoG based shape and SIFT-BoW based appearance congruity.

4.1 Datasets

We use three action datasets of varying modalities: ETHZ-activity [14], CAD-120 [25] and MPII-Cooking [35]. The ETHZ-activity is an RGB-D dataset captured by a color and a ToF camera with a resolution of 640×480 and 170×144, respectively. It contains 143 sequences of 12 high level activities performed by 6 different actors. Human pose extracted via a model based method consists of 13 3d joint locations from the upper body. Interactions are mostly restricted to a single object but with varying appearances. The 12 object classes vary from medium-size e.g. *teapot* and *mug* to small-size e.g. *marker* and *phone*. A typical frame illustrating the relative size of the objects is shown in Figure 3.

The CAD-120 is an RGB-D dataset recorded with a color camera and structured light for depth having VGA-resolutions for both modalities. It contains 120 sequences of 10 different high level activities performed by 4 different actors. Human pose consisting of 15 3d joint locations from the whole body is extracted using OpenNI SDK. The pose is noisy which is more pronounced for hands and legs. The activities involve interactions with various objects e.g. *making cereal* indicates the presence of instances of the object classes *box*, *milk* and *bowl*.

The MPII-Cooking is a high resolution (1624×1224) RGB dataset. It contains 65 sequences of 2 high-level activities performed by 12 different actors. The

Fig. 3. Sample images of human-object interaction from ETHZ-activity dataset, CAD-120 dataset and MPII cooking dataset in that order. Object of interest is bounded in red and pose overlayed in orange.

human pose is extracted by a part-based detection approach and consists of 8 2d joint locations for the upper body without head. For the binary potential $\Psi^{func}(l_v, l_w)$ (11), we take therefore the mean of both shoulders instead of the head as reference joint. Apart from involving multiple objects, the objects are often occluded or covered by food during the activity e.g. *plate* during the process of preparing a salad.

For evaluation, we labelled the objects in the three datasets by drawing tight bounding boxes around the objects for every 10^{th} frame and interpolating intermediate bounding boxes. The annotations and evaluation scripts will be made publicly available.

4.2 Inference

The output of the system is a collection of tubes that best describe an object class common in all input videos. Discovered instances of object classes are shown in Figure 5. In order to evaluate the quality of these tubes, we study frame- and class-wise PASCAL IoU measures. A frame-IoU measure is defined as a ratio of areas of intersection over union of the ground truth and inferred bounding boxes. A tube-IoU is defined as the average of all frame-IoUs. Similarly, a class-IoU is defined as the average of all inferred tube-IoUs.

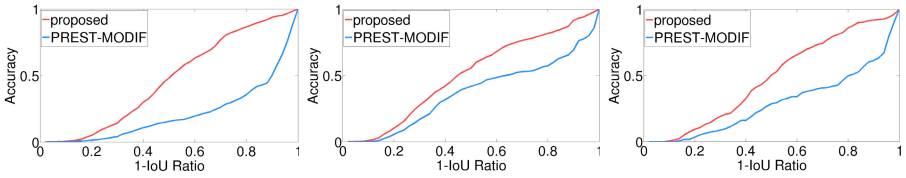
To learn the scalar weights λ of the energy model (3), (8), (12) and [33], we use ground-truth object annotations of one randomly chosen object class as validation in each dataset: *puncher* (ETHZ), *milkbox* (CAD) and *whisker* (MPII). In order to set these parameters, we perform a grid-search in $\{0.05, 0.25, 0.50, 0.75, 1.00\}$ and take the configuration that maximizes class-IoU for the validation class. We therefore exclude validation classes from all performance evaluations that follow.

4.3 Comparison

Firstly, we compare the proposed tube generation process with the object proposal technique [28] considering every 10^{th} frame in the ETHZ-action dataset. While the recall of the proposal technique was (0.19, 0.58, 0.67) for (10^2, 10^3, 10^4) proposals per frame respectively, our approach as described in Section 3.1

Table 1. Average class-IoU of the proposed model (APP+SIZ+FUN) for the three datasets. All three types of potentials that model object appearance (APP), size prior (SIZ) and object functionality (FUN) are important for the final performance. Our proposed approach outperforms the method [33], which relies on motion segments and object appearance.

	prest-exact [33]	prest-modif	proposed	APP	APP+SIZ	FUN	APP+FUN	FUN+SIZ
ETHZ-Action	0.063	0.249	**0.447**	0.192	0.305	0.292	0.312	0.390
CAD-120	0.039	0.246	**0.410**	0.168	0.191	0.147	0.202	0.350
MPII-Cooking	0.023	0.221	**0.342**	0.079	0.149	0.229	0.235	0.288



(a) ETHZ-action (b) CAD-120 (c) MPII-Cooking

Fig. 4. Accuracy measured as fraction of bounding boxes with an IoU ratio greater equal than a given threshold. The x-axis plots 1-IoU i.e. the higher the value on the x-axis the more tolerant is the success threshold and the higher the accuracy. The accuracy is averaged over all classes.

achieves a recall of 0.65 for only 30 tubes per video. This verifies that the proposed tube generation module is well suited to the current scenario.

Further, we compare the proposed approach with a method for learning from weakly labelled videos [33] on all three datasets. The average class-IoU is presented in Table 1. The performance of the proposed approach supercedes that of [33] significantly. The reason for such poor performance of [33] is that the extracted motion segments do not correspond to objects in most cases and are therefore not suitable for the task at hand. We therefore modify the method by using the tube sampling approach introduced in Section 3.1 and the energy functional proposed in [33] to select tubes that most likely contain instances of the object class. We denote the modified approach as prest-modif in Table 1. In contrast to [33], prest-modif achieves improved results but is still inferior when compared to the energy functional used in the proposed approach.

To evaluate the quality of inferred tubes, we define class-accuracy as the fraction of bounding boxes with an IoU ratio greater equal than a given threshold. Figure 4 shows class-accuracy averaged over all classes for decreasing IoU ratios. For [33], the IoU ratio for nearly all bounding boxes is close to zero. We therefore plot the accuracy only for prest-modif. As can be seen, the average class-accuracy of the proposed method for different thresholds consistently outperforms that of prest-modif in all three datasets. The biggest difference in performance is for the ETHZ dataset at 1-IoU=0.8 where the performance of the proposed approach and prest-modif are 0.86 and 0.36 respectively. At IoU=0.5, the accuracies of the methods are (0.48, 0.16) for ETHZ, (0.56, 0.42) for CAD and (0.53, 0.29) for the MPII dataset respectively.

4.4 Impact of Potentials

In order to characterize the contribution of designed potentials, we group them into three categories: APP consisting of potentials that are intrinsic to object appearance $\{\Phi^{app}, \Psi^{shape}\}$, SIZ denotes the size prior $\{\Phi^{size}\}$ and FUN consisting of potentials derived from human-object interaction $\{\Phi^{pose}, \Phi^{body}, \Psi^{func}\}$. Performances of different group combinations are presented in Table 1.

Table 2. Percentage change in average class-IoU performance when any given potential is discarded from the model

	Φ^{app}	Φ^{pose}	Φ^{body}	Φ^{size}	Ψ^{shape}	Ψ^{func}
ETHZ-Action	0.35	1.88	-25.49	-13.50	-4.62	-8.86
CAD-120	-48.66	-15.73	-18.89	-20.80	-40.15	-9.19
MPII-Cooking	-15.85	0.06	-31.09	-10.70	0.058	-60.95

The first observation is that the group APP performs worse when compared to prest-modif for all datasets. This fall in performance is expected because APP uses only 2 potentials while the energy functional of prest-modif uses 6 terms to model the appearance of an object. The performance improves when the size prior is added (APP+SIZ). The functionality terms (FUN) outperform prest-modif and APP on the ETHZ and MPII datasets emphasizing the fact that human interaction is a valuable cue to discover objects, but not sufficient. Using the functionality and the appearance terms (FUN+APP), the performance is higher than using only one of them. Finally, the pair of (FUN+SIZ) performs best amongst all subset combinations, but only attains 80% of the accuracy attained by the full model. This indicates that object appearance, functionality and size prior are all important for maximal performance.

In addition, we present percentage change in class-IoU performance when each potential is discarded from the model in Table 2. It can be seen that performance drops upon eliminating any potential almost in all cases. For the CAD dataset, removing any potential has a negative effect. Appearance based features have minimal impact on the ETHZ dataset as they are not reliable for small objects and Ψ^{shape} has negligible impact on the MPII dataset owing to drastic variations in object appearances during interaction. The terms Φ^{body}, Φ^{size} and Ψ^{func} are required by all datasets as indicated by loss in performance when they are discarded.

Further, we study the robustness of pose-related potentials with respect to strong pose estimation noise on the CAD dataset. To this end, we add normally distributed noise with variance $100cm^2$, $200cm^2$ and $400cm^2$ to each 3d joint position. The average class-IoU then drops to 0.365, 0.342 and 0.323 respectively from the baseline of 0.410 (see Table 1). The performance, however, is still higher than without using these potentials (see APP+SIZ in Table 1).

4.5 Evaluating Object Models

As a final comparison, we study the quality of the inferred tubes for object detection. We split each dataset such that no actor occurs in both training and testing data. For training, we considered data from 5 out of 6 actors in ETHZ-action, 3 out of 4 actors in CAD-120 and 9 out of 12 actors in MPII-cooking datasets. The rest of the data was used for testing.

For object detection, we use a Hough forest [17] with 5 trees each trained with 50,000 positive and 50,000 negative patches (drawn uniformly from the background) and a maximal depth of 25. We do not make use of depth for

Table 3. Average precision (%) for different datasets comparing object models built from ground truth data (GTr.) and data inferred by the proposed method (Infer).

Class	GTr.	Infer	Class	GTr.	Infer	Class	GTr.	Infer	Class	GTr.	Infer
ETHZ-Action											
brush	45.1	33.6	calcul.	100.0	100.0	camera	83.5	73.0	remote	49.4	34.4
mug	38.0	39.5	headph.	69.8	69.8	marker	39.7	33.3	teapot	63.2	50.9
videog.	78.3	82.0	roller	99.6	69.0	phone	0.05	0.06	**Avg.**	60.6	53.2
CAD-120											
book	11.2	08.0	medbox.	58.3	40.4	bowl	24.5	25.0	mwave.	71.4	71.0
box	24.4	19.1	plate	16.2	14.1	cup	14.8	09.4	remote	14.1	17.6
			cloth	20.1	15.1	**Avg.**	29.4	24.4			
MPII-Cooking											
bowl	69.2	11.1	spiceh.	100.0	100.0	bread	25.5	06.2	squeez.	61.5	61.5
plate	43.4	43.4	tin	33.0	23.9	grater	02.2	01.2	**Avg.**	47.8	35.3

this experiment. For comparison, we use manually annotated bounding boxes of training images, i.e. every 10^{th} frame of training sequences. This is denoted as 'GTr.' in Table 3. The 'Infer' training data is based on an equal number of frames from the automatically extracted tubes inferred by the proposed model.

The results show that optimal performance is achieved for categories like *calculator*, *mug* in ETHZ, *bowl*, *microwave* in CAD-120 and *spiceholder*, *squeezer* in MPII. A loss in performance is observed for many categories due to weaker supervision which is explained by the fact that the bounding boxes of extracted tubes are noisier than manually annotated training data. Nevertheless, performances of the object detectors trained on weakly supervised videos achieve 87.7% (ETHZ), 83.0% (CAD) and 74.4% (MPII) of that from full supervision.

We also compare with [14] which is an unsupervised approach that segments and clusters videos based on pose features. [14] generates 20 clusters for the ETHZ-action dataset without labels and only 3–21 object samples per cluster while our approach generates more than 300 samples per class. Although the resulting clusters cannot be directly compared with our approach, we manually labelled the clusters and trained object detectors for all 12 classes. The resulting average precision on ETHZ is 24.85% in comparison to 53.23% of our approach.

5 Conclusion

We have addressed the problem of discovering medium and small sized objects from videos with activities. Our experiments have shown that current approaches for learning from weakly labelled videos that rely on motion segmentation fail for this task. We have also shown that using object appearance alone is insufficient in such scenarios and that encoding functionality greatly improves performance. Interestingly, the results also revealed the complementary nature of appearance and functionality related potentials for object discovery. The generalization capabilities of our approach were demonstrated on three datasets that span a variety of different activities, modalities (RGB vs. RGB-D), and pose representations (2d vs. 3d). Finally, our weakly supervised approach outperformed an unsupervised approach and achieves between 74% and 88% of the performance of a fully supervised approach for object detection.

Fig. 5. Discovered instances of the object classes: *Marker, Mug, Camera, Roller, Milkbox, Bowl, Cloth, Microwave, Plate, Tin, Bread, Squeezer* and failure cases *Teapot, Brush*. The first image in each row shows relative object size by illustrating a typical action scene with overlayed pose and a bounding box around the object of interest. Since the objects are relatively small, images are best viewed by zooming in.

Acknowledgements. Authors acknowledge financial support from the DFG Emmy Noether program (GA 1927/1-1).

References

1. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR, pp. 73–80 (2010)
2. Blaschko, M.B., Vedaldi, A., Zisserman, A.: Simultaneous object detection and ranking with weak supervision. In: NIPS, pp. 235–243 (2010)
3. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: ACM Int. Conf. on Image and Video Retrieval, pp. 401–408 (2007)
4. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 282–295. Springer, Heidelberg (2010)
5. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. PAMI 33(3), 500–513 (2011)
6. Chum, O., Zisserman, A.: An exemplar model for learning object classes. In: CVPR, pp. 1–8 (2007)
7. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. PAMI 24(5), 603–619 (2002)
8. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 284–298. Springer, Heidelberg (2012)
9. Deselaers, T., Alexe, B., Ferrari, V.: Localizing objects while learning their appearance. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 452–466. Springer, Heidelberg (2010)
10. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV 88, 303–338 (2010)
11. Fathi, A., Ren, X., Rehg, J.: Learning to recognize objects in egocentric activities. In: CVPR, pp. 3281–3288 (2011)
12. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. IJCV 59(2), 167–181 (2004)
13. Filipovych, R., Ribeiro, E.: Recognizing primitive interactions by exploring actor-object states. In: CVPR (2008)
14. Human Body Analysis. In: Fossati, A., Gall, J., Grabner, H., Ren, X., Konolige, K. (eds.) Consumer Depth Cameras for Computer Vision. Springer (2013)
15. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single view geometry. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 732–745. Springer, Heidelberg (2012)
16. Gall, J., Fossati, A., van Gool, L.: Functional categorization of objects using real-time markerless motion capture. In: CVPR, pp. 1969–1976 (2011)
17. Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V.: Hough forests for object detection, tracking, and action recognition. PAMI 33(11), 2188–2202 (2011)
18. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: CVPR, pp. 1529–1536 (2011)
19. Gupta, A., Davis, L.: Objects in action: An approach for combining action understanding and object perception. In: CVPR, pp. 1–8 (2007)

20. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3D scene geometry to human workspace. In: CVPR, pp. 1961–1968 (2011)
21. Jiang, Y., Koppula, H., Saxena, A.: Hallucinated humans as the hidden context for labeling 3D scenes. In: CVPR, pp. 2993–3000 (2013)
22. Jones, M., Rehg, J.: Statistical color models with application to skin detection. IJCV 46(1), 81–96 (2002)
23. Kjellström, H., Romero, J., Kragic, D.: Visual object-action recognition: Inferring object affordances from human demonstration. CVIU 115, 81–90 (2010)
24. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. PAMI 28(10), 1568–1583 (2006)
25. Koppula, H., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. IJRR 32(8), 951–970 (2013)
26. Lee, Y.J., Grauman, K.: Learning the easy things first: Self-paced visual category discovery. In: CVPR, pp. 1721–1728 (2011)
27. Leistner, C., Godec, M., Schulter, S., Saffari, A., Werlberger, M., Bischof, H.: Improving classifiers with unlabeled weakly-related videos. In: CVPR, pp. 2753–2760 (2011)
28. Manen, S., Guillaumin, M., Van Gool, L.: Prime object proposals with randomized prim's algorithm. In: ICCV, pp. 2536–2543 (2013)
29. Moore, D., Essa, I., Hayes, M.: Exploiting human actions and object context for recognition tasks. In: ICCV, pp. 80–86 (1999)
30. Ommer, B., Mader, T., Buhmann, J.: Seeing the Objects Behind the Dots: Recognition in Videos from a Moving Camera. IJCV 83, 57–71 (2009)
31. Peursum, P., West, G., Venkatesh, S.: Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In: ICCV, pp. 82–89 (2005)
32. Pieropan, A., Ek, C.H., Kjellstrom, H.: Functional object descriptors for human activity modeling. In: ICRA, pp. 1282–1289 (2013)
33. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: CVPR, pp. 3282–3289 (2012)
34. Ramanan, D., Forsyth, D.A., Barnard, K.: Building models of animals from video. PAMI 28(8), 1319–1334 (2006)
35. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. In: CVPR, pp. 1194–1201 (2012)
36. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: CVPR, pp. 1939–1946 (2013)
37. Schulter, S., Leistner, C., Roth, P.M., Bischof, H.: Unsupervised object discovery and segmentation in videos. In: BMVC, pp. 391–404 (2013)
38. Turek, M.W., Hoogs, A., Collins, R.: Unsupervised learning of functional categories in video scenes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 664–677. Springer, Heidelberg (2010)
39. Tuytelaars, T., Lampert, C.H., Blaschko, M.B., Buntine, W.: Unsupervised object discovery: A comparison. IJCV 88, 284–302 (2010)
40. Winn, J.M., Jojic, N.: Locus: Learning object classes with unsupervised segmentation. In: ICCV, pp. 756–763 (2005)