# Weak Supervision for Detecting Object Classes from Activities

Abhilash Srikantha[a,b,*], Juergen Gall[a]

[a] *University of Bonn, Germany*
[b] *Max Planck Institute for Intelligent Systems, Tuebingen, Germany*

## Abstract

Weakly supervised learning for object detection has been gaining significant attention in the recent past. Visually similar objects are extracted automatically from weakly labelled videos hence bypassing the tedious process of manually annotating training data. However, the problem as applied to small or medium sized objects is still largely unexplored. Our observation is that weakly labelled information can be derived from videos involving human-object interactions. Since the object is characterized neither by its appearance nor its motion in such videos, we propose a robust framework that taps valuable human context and models similarity of objects based on appearance and functionality. Furthermore, the framework is designed such that it maximizes the utility of the data by detecting possibly multiple instances of an object from each video. We show that object models trained in this fashion perform between 86% and 92% of their fully supervised counterparts on three challenging RGB and RGB-D datasets.

*Keywords:* Weakly Supervised, Object Detection, Human-Object Interaction, RGB-D Videos

## 1. Introduction

Data driven approaches have been shown to perform well [1, 2, 3] for the tasks of object detection, face recognition and image classification. Trained in a completely supervised approach, their high performance stems majorly from two sources. Firstly, an increased model complexity allows for designing classifiers with increased capability of handling data-in-the-wild [4], and secondly, vast amount of rich training data is made available to retain the generalization capabilities of the designed classifier. Having access to labelled data has therefore become a pre-requisite for designing robust solutions. However, this dependence can be a bottleneck in many scenarios either because of efforts involved or inherent ambiguity during annotation. In the future, present day crowdsourcing solutions will be impractical due to high associated costs and ever increasing amount of data. Moreover, this also ignores vast amount of freely available weakly structured data. As a result, recent works in object detection have turned towards utilizing weakly labelled data [5, 6, 7, 8, 9, 10, 11, 12], particularly videos [13, 14, 15]. Critically, these methods assume that motion or appearance of objects are sufficient descriptors for segmenting them with relative ease, which is indeed the case for large active objects such as flying airplanes and walking tigers. The assumption is further strengthened by the abundance of labelled videos on the Internet which are characteristically object- or action-centric.

Most present day models for object detection work well for objects that cover a significant part of the image. The concept of representing objects by parts [16] and scoring their relative locations in a star model [17] brought significant improvements in modelling larger objects such as airplanes, boats, cars, horses etc. However, modelling daily objects such as markers, remotes or plates is still largely unresolved [18]. Exploiting weakly labelled data for such objects is further complicated by the scarcity of *clean* data because such objects do not form popular subjects for generating and sharing videoclips.

On the other hand, labelled videos involving human activity, like pouring milk or eating cereal are abundantly available. Such data, however, violates the principle assumption since the prevelant themes of the video are now human body parts and background clutter instead of objects of interest; thus resulting in the failure of contemporary methods as demonstrated in our experiments. The problem is further complicated by varing appearance and pose of objects undergoing interactions coupled with low resolutions and frequent occlusions. As a result, appearance-only approaches are limited in capacity to detect such objects, as shown in our experiments.

A preliminary version of this work appeared in [19] in which we address the problem of weakly supervised learning for medium or small sized objects from action videos where humans interact with them. We propose a method composed of two stages as shown in Figure 1. The first stage tackles the issue of objects of interest not corresponding to dominant motion segments. Instead, we generate seeds by sampling superpixels that are likely to overlap with objects from a generative model encoding human-

[*] Corresponding author
*Email addresses:* srikanth@iai.uni-bonn.de (Abhilash Srikantha), gall@iai.uni-bonn.de (Juergen Gall)

object interaction. Object candidates are then generated by tracking the seeds to form spatio-temporal tubes as illustrated in Figure 2. To tackle the rich variety of object appearance and motion, tracking is made robust by sampling from a pool of algorithms and parameters. The second stage tackles the issue of appearance features alone being insufficient to describe objects. To this end, we propose an object similarity measure that depends not only on appearance and size but also on functionality derived from relative motion with respect to the human.

In the present work, we generalize the assumption of extracting a single tube from each video as in [19]. The generalization facilitates extracting (possibly) numerous tubes overlapping with the object as a final solution, resulting in increased economy of tapping information from the data. Also, due to inherent clutter and noise, having flexibility to choose no tube from a video can potentially improve homogeneity within inferred tubes. In this regard, we incorporate the above improvements as a greedy iterative approach into the inference procedure.

We demonstrate the robustness of our approach on three demanding datasets, namely one RGB dataset [20] and two RGB-D datasets [21, 22]. Each dataset is recorded with a different type of sensor viz. time of flight [21], color camera [20] and structured light sensor [22]. Automatically extracted human pose in each dataset also varies in the number of detected body parts and in the quality of joint localization. We demonstrate that the proposed method is successful in detecting objects from videos of activities on all three datasets. Further, we provide a detailed evaluation of the generalized inference with regard to the quality of inferred tubes and the impact of various potentials.

## 2. Related Work

Object detection encapsulates determining whether an image contains instances of a certain object category and their locations. Optionally, additional information e.g. about part-locations [16], object pose [23, 24] and occlusion [25, 26, 27, 28] has been inferred. The fundamental challenge is to effectively model inter and intra class appearance and shape variation of objects. To this day, this is usually achieved by designing a parametric model.

These models can be broadly classified into three categories. The first category of algorithms extract local (SIFT, HOG) or global (GIST, Fischer) image features and represent objects as *bag of words* (BoW) through statistical classifiers [29, 30, 31]. Although shown to work well for classification, the approach is suboptimal for locating objects. The approach [32] alleviates the problem by efficiently building dictionaries of visual words in a framework that is jointly optimized for classifying and regressing object centers. The second category of algorithms detect the presence of objects by fitting rich object models such as *deformable part models* (DPMs). This process can reveal useful hidden information such as object part locations [16] and occlusion [25, 26, 27], but DPMs are usually trained from images with known locations of objects or even their parts [33]. The third category of algorithms are *convolutional neural networks* (CNNs) which learn feature representations of objects [34, 35, 36].

To this day, the parameters of the model are learned through a set of traning instances using statistical machine learning techniques. The various learning methods can therefore be characterized by the extent of supervision involved during learning. At one end of the spectrum, fully supervised methods require careful annotation of object locations in the form of bounding boxes [16, 34, 32], segmentations [37] or even object part locations [33, 38], which is costly and can frequently introduce inconsistency and ambiguity. On the other hand, unsupervised learning methods that do not require any supervision aim at finding similar objects in a set of unlabelled images [7, 39] or videos [40]. They are, however, often limited to frequently occuring and visually consistent objects and are easily susceptible to background clutter. The stringent requirements regarding cleanliness of input data has been relaxed by using exemplar samples [41] or by employing pretrained object detectors [42, 43, 44]. On similar lines, cosegmentation [5, 6, 8, 9] approaches identify object instances up to a bounding box or segmentation on a collection of images with an object class label. Further, [45] segments objects in videos by clustering long term point trajectories. However, it assumes similarity between trajectories from object regions and does not investigate relationships between videos.

Weakly supervised learning lies at the middle of the spectrum by providing annotations at a higher level of abstraction thereby reducing the annotation effort. This is an important scenario for many practical applications because weak labels are more readily available e.g. in form of text tags [46], movie transcripts [47, 48], geographical meta-data [49] and captions [50]. Weakly labelled videos are exploited in [15, 51, 13, 14, 19].

In the context of object detection, the common practice has been to model object location with latent variables while jointly learning an appearance model. Most approaches impose certain assumptions for successful application e.g. [12, 11, 10, 52, 14] assume a single predominant object in the input data and [14, 15] assume rigid or articulated objects with motion distinctive from its background. These assumptions guide the latent variables such that the solution extracts object instances despite object deformations and background. In practice, however, the quality of a solution depends on the similarity measure used. For instance, [52, 10, 12] obtain a solution set that is most consistent in terms of shape and color, [14, 15] exploit motion and appearance consistency within the input data and [11] exploits symmetry constraints of objects in a multiclass framework. The solution is mostly obtained by multiple instance learning [10] or by minimizing an energy on a fully connected graph [14, 19]. Most methods fail to exploit training data completely as they only select one instance per image or video. This is a suboptimal choice because all other instances of that object in the image are ignored therefore failing to tap its true potential. This limitation is dealt in [11] by introducing a latent SVM formulation that exploits presence of multiple object instances in an image. On similar lines, the present method is a generalization of [19] where the assumption of selecting strictly one instance per video is relaxed in the framework of exploiting human context for building models of small and medium sized objects.

The theme of scene understanding driven by human context has gained recent attention owing to advances in techniques and commercial SDKs for human pose estimation [21, 22, 53, 54, 55, 56, 57, 58, 59, 60, 61]. In [59, 61], image regions are segmented based on observed human trajectories in office and street environments. While several works [62, 63, 64] investigate combining object detection and action recognition, the works [21, 22, 58, 60] employ affordance cues as higher level representation for video understanding. In [58], both object detection and activity recognition are improved by jointly representing objects and their functionality. Unsupervised clustering of objects based on their motion relative to humans is performed in [21]. Further, human activity is recognized based on object functionality in the context of hand-object interactions in [60] or based on high level attribute co-occurance statistics in [65, 66]. In [22], activities and object affordances are learned simultaneously, while [67] deals with appearance based object detection based on weak action-object labels in egocentric videos.

Human models have also been used to hallucinate their interactions with given scenes. In [55], scene locations that can afford the action *sittable* are learned through geometric relations between the scene and a human pose representing the action. A similar approach is incorporated for 3D scene labelling in [57, 53] and extracting scene geometry in [54] by modelling relations between objects and human pose. An opposite approach is followed by [56] where human poses are inferred based on scene geometry.

## 3. Learning object models from activities

Figure 1 illustrates the pipeline for detecting instances of an object class in a set of RGB-D or RGB videos. Input to the pipeline is a collection of videos that is labelled with the involved activity of human-object interaction. E.g., the label *drinking coffee* indicates the presence of a mug. We also assume that the 2d or 3d human pose has already been extracted. This is readily feasible because of freely available SDKs for RGB-D data and due to significant progress in 2d pose estimation in the recent past. No further restrictions are imposed on the nature of input videos in that they may contain a multitude of activities, persons and/or objects. For instance, the labels *eating cereal* and *stacking bowls* are different activities that, among many other objects, commonly involve a bowl.

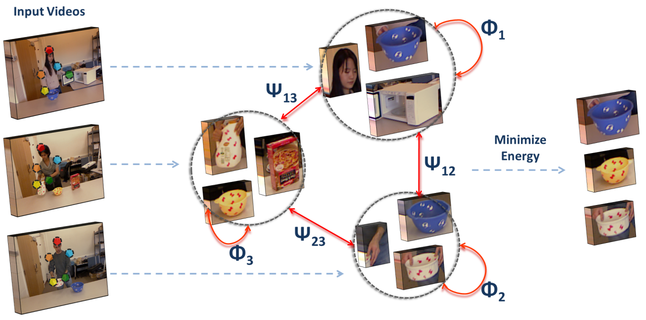


Figure 1: Processing pipeline: Input is a set of action videos with human pose. Multiple sequences of object proposals (tubes) are generated from each video. By defining a model that encodes the similarity between tubes in terms of appearance and object functionality, instances of the common object class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

The first step involves generating several object proposals per video. An object proposal is modelled as a spatio-temporal region in the video, also called a tube. Multiple tubes are sampled from a video using a simple graphical model representing human-object interactions. This procedure is explained in Section 3.1. While the purpose is to extract tubes that significantly overlap with the objects of interest, this is hardly true in practice as they overlap mostly with background clutter or body parts; thereby lacking object information. To this end, given a collection of tubes from all videos, we select a subset of tubes best describing the object from each video. This is realized by minimizing an energy functional that comprises unary and pairwise potentials. Unary potentials evaluate the presence or absence of an object in a tube and pairwise potentials evaluate the similarity of objects between two tubes. All potentials incorporate appearance and functionality as described in Section 3.3.

### *3.1. Generating tubes*

Extracting dominant motion segments as in [68, 14] is a naive way of generating tubes. Such methods cannot generate meaningful tubes in the present context as dominant motion segments mostly correspond to body parts. Instead, we generate a tube $T_v$ from video $v$ by tracking a frame based superpixel $S$ over time. Owing to the rich variety of objects and actions, we found no unique universal setting for either superpixel selection or tracking that yielded tubes of good quality. We therefore model this uncertainty by randomly selecting a tracking algorithm $\tau$ from a pool of tracking algorithms. In other words, we obtain a set of tubes by sampling from the probabilistic graphical model defined over the tubes, given by

$$p(T_v, \tau, S) = p(T_v|\tau, S)p(\tau)p(S). \tag{1}$$

In practice, we use a pool of two tracking algorithms that are selected with uniform probability i.e. $p(\tau) = 0.5$. The first method is based on propagating a superpixel based on median optical flow [69] into the next or previous frame. The second method is based on mean shift [70]. While the first method successfully tracks medium sized rigid objects, it is easily misled by fast motion, background clutter or small objects. The second method is more robust to fast motion but gets misled by occlusions during human-object interactions. Since either case is not robust for long term tracking, we limit the length of each tube to a maximum of 300 frames.

For generating superpixels $S$, we modify [71] to incorporate depth as an additional feature. Since the relevance of depth information depends on material properties, object size and object characteristics, we found it useful to use a pool of data channels. In practice, the pool is defined as $\sigma \in \{RGB, D, RGB-D\}$. Each configuration in the pool represents the data using which superpixels are generated. The probability of selecting a superpixel also depends on frame $f$ and a spatial prior that depends on the frame $p(l|f)$. We obtain a superpixel by sampling from

$$p(S, f, l, \sigma) = p(S|f, \sigma)p(l|f)p(f)p(\sigma). \tag{2}$$

We set a uniform prior over $\sigma$. $p(f)$ is a temporal prior that represents the probability of close human-object interaction in frame $f$. While a high level representation of humans and objects can be utilized to model this probability, we use a uniform distribution. In other words, we assume that human-object interaction occurs in all frames. As for the spatial prior $p(l|f)$, we incorporate human pose information. To this end, we select the joint with the highest variance in location, computed within a temporal neighborhood of 15 frames. We then model $p(l|f)$ as an isotropic uniform distribution at joint location $j$ at frame $f$ with radius 400mm in case of RGB-D videos. Since human pose from RGB does 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 practice, we use $\gamma = 0.2$.

Sampling a tube from Equation (1) corresponds to sampling a superpixel and a tracking method. To sample a superpixel $S$ from Equation (2), we sample a configuration $\sigma$ to generate a superpixel segmentation of a randomly selected frame $f$ among which one superpixel $S$ is chosen based on the spatial prior $p(l|f)$. This is then tracked over time using a randomly sampled tracking algorithm $\tau$ as per Equation (1) to generate a tube $T_v$. The procedure is illustrated in Figure 2. As for the number of tubes generated per video, we set it to 30 for all our experiments.

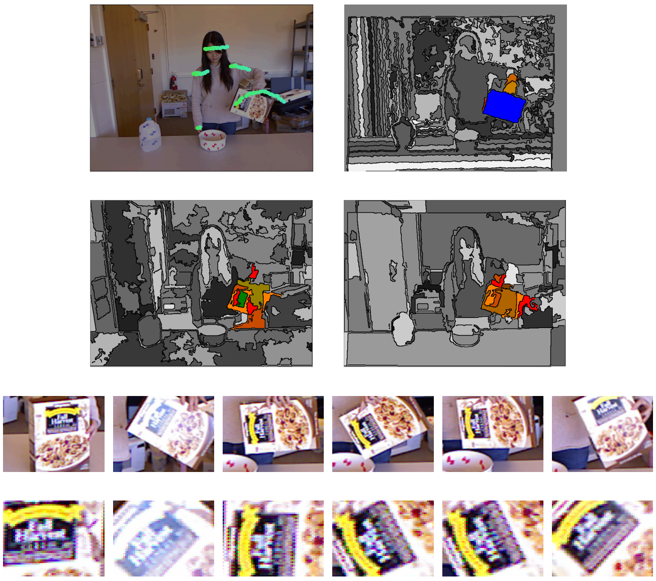


Figure 2: Illustrating the tube generation process. Images of the top half: The first image shows joint trajectories. The most active joint is used to compute the spatial prior for selecting superpixels. The three images next to it show three superpixel representations computed using depth (D), color (RGB) and both (RGB-D). Colored superpixels are within the specified distance of the most active joint. Each of the last two rows visualizes a tube $T_v$ sampled from the blue and green superpixel $S$ respectively.

### 3.2. Generating object hypotheses

Given a set of candidate tubes $\mathcal{T}_v$ in each video $v$, the goal of [72, 14, 19] is to select one tube per video that contains the object class and is tight around the object. This has been formulated in [19] as an energy minimization problem defined jointly over all $N$ videos. Let $l_v \in \mathcal{L}_v = \{1, \dots, |\mathcal{T}_v|\}$ be a label that selects one tube out of a video. The energy of all selected tubes $(l_1, \dots, l_N)$ is defined as

$$E(l_1, \dots, l_N) = \sum_{v=1}^{N} \left( \Phi(l_v) + \sum_{w=v+1}^{N} \Psi(l_v, l_w) \right) \tag{3}$$

where the unary potentials $\Phi$ measure the likelihood of a single tube being a tight fit around an object. The binary potentials $\Psi$ measure the homogeneity in object appearance and functionality of a pair of tubes.

The constraint of selecting exactly one tube per video, however, assumes that there is at least one tube containing the object and limits the amount of information extracted from the data. In some cases, a video might contain more than one object instance or might not contain the object at all. We therefore extend the approach proposed by [19] and reformulate Equation (3) to select a varying number of tubes from each video. To this end, we search for a set of tubes $S_v \subseteq \mathcal{L}_v$ for each video, which can also be an empty set. The energy of a configuration $\mathcal{S} = (\mathcal{S}_1, \dots, \mathcal{S}_N)$ is then defined as

$$\begin{split} E\left(\mathcal{S}\right) = \sum_{v=1}^{N} \Bigg\{ \sum_{j=1}^{|\mathcal{S}_v|} \Phi\left(l_v^j\right) + \sum_{w=v+1}^{N} \sum_{j=1}^{|\mathcal{S}_v|} \sum_{k=1}^{|\mathcal{S}_w|} \Psi\left(l_v^j, l_w^k\right) \\ + \alpha \left(1 - \frac{\gamma^{|\mathcal{S}_v|} e^{-\gamma}}{|\mathcal{S}_v|!}\right) \Bigg\}. \end{split} \tag{4}$$

The first two terms $\Phi$ and $\Psi$ are the same as in Equation (3), but they are computed over all selected tubes $\mathcal{S}_v$ for each video. The last term is a prior on the number of expected tubes with object instances per video, modelled by a Poisson distribution $P_\gamma(|\mathcal{S}_v|)$. Since we minimize Equation (4), we use $1 - P_\gamma(|\mathcal{S}_v|)$. The parameter $\gamma$ represents the expected number of object-overlapping tubes. The impact impact of this prior is controlled by $\alpha$. If we use the constraint that $|S_v| = 1$ for all videos $v$, minimizing Equation (4) is equivalent to minimizing Equation (3) since the last term reduces to a constant.

To minimize Equation (4), we use an iterative, greedy approach. To this end, we extend the label set by an auxiliary label, i.e., $\hat{\mathcal{L}}_v = \{0, 1, \ldots, |\mathcal{T}_v|\}$. Let $\mathcal{S}_v^{t-1}$ denote the selected tubes for each video at the end of iteration $t-1$. In the next iteration, we then either select no tube, which corresponds to $\hat{l}_v^t = 0$, or one tube per video. We exclude the already selected tubes as $\hat{l}_v^t \in \hat{\mathcal{L}}_v^t = \hat{\mathcal{L}}_v \setminus \mathcal{S}_v^{t-1}$ and the energy for iteration $t$ is defined by

$$E(\hat{l}_1^t, \ldots, \hat{l}_N^t | \mathcal{S}^{t-1}) = \sum_{v=1}^{N} \left( \Phi(\hat{l}_v^t) + \sum_{w=v+1}^{N} \sum_{k=1}^{|\mathcal{S}_w^{t-1}|} \Psi(\hat{l}_v^t, l_w^k) \right. \\ \left. + \sum_{w=v+1}^{N} \Psi(\hat{l}_v^t, \hat{l}_w^t) \right) \tag{5}$$

where

$$\Phi(\hat{l}_v^t{=}0) = \alpha \left( 1 - \sum_{n=0}^{|\mathcal{S}_v^{t-1}|} \frac{\gamma^n e^{-\gamma}}{n!} \right) \tag{6}$$
$$\text{and} \quad \Psi(0, \hat{l}_w) = \Psi(\hat{l}_v, 0) = 0.$$

In Equation (5), the constant terms

$$\sum_{v=1}^{N} \sum_{j=1}^{|\mathcal{S}_v^{t-1}|} \Phi\left(l_v^j\right) \quad \text{and} \quad \sum_{v=1}^{N} \sum_{w=v+1}^{N} \sum_{j=1}^{|\mathcal{S}_v^{t-1}|} \sum_{k=1}^{|\mathcal{S}_w^{t-1}|} \Psi\left(l_v^j, l_w^k\right) \tag{7}$$

are omitted. The Poisson prior $P_\gamma(|\mathcal{S}_v|)$ is expressed in the greedy approach by Equation (6). In other words, the cost of selecting no tube corresponds to the probability that the video contains more than $|\mathcal{S}_v^{t-1}|$ tubes with object instances. Using $\hat{\Phi}(\hat{l}_v^t) = \Phi(\hat{l}_v^t) + \sum_w \sum_k \Psi(\hat{l}_v^t, l_w^k)$, we can rewrite Equation (5) as

$$E(\hat{l}_1^t, \ldots, \hat{l}_N^t | \mathcal{S}^{t-1}) = \sum_{v=1}^{N} \left( \hat{\Phi}(\hat{l}_v^t) + \sum_{w=v+1}^{N} \Psi(\hat{l}_v^t, \hat{l}_w^t) \right). \tag{8}$$

Accumulating binary potentials into the unaries as in Equation (8) encourage tubes selected in the present iteration to be similar to those in the past. This can cause undesirable effects as errors in the present iteration are propagated to the next. In this regard, independently optimizing each iteration can be advantageous as verified in our experiments and is formulated as

$$E(\hat{l}_1^t, \ldots, \hat{l}_N^t | \mathcal{S}^{t-1}) = \sum_{v=1}^{N} \left( \Phi(\hat{l}_v^t) + \sum_{w=v+1}^{N} \Psi(\hat{l}_v^t, \hat{l}_w^t) \right). \tag{9}$$

**Algorithm 1** Greedy inference procedure

```
procedure INFER($\mathcal{S}_1, \ldots, \mathcal{S}_N$)
    Initialize $\mathcal{S}_v^0 = \emptyset, \hat{\mathcal{L}}_v = \{0, 1, \ldots, |\mathcal{T}_v|\} \forall\, 1 \leq v \leq N$
    Precompute unaries $\Phi(\hat{l}_v)$ and binaries $\Psi(\hat{l}_v, \hat{l}_w)$
    Iterator $t = 0$
    Continue = $True$
    while Continue do
        $t = t + 1$
        Update set of possible labels as $\hat{\mathcal{L}}_v^t = \hat{\mathcal{L}}_v \setminus \mathcal{S}_v^{t-1}$
        Obtain $(\hat{l}_1^t, \ldots, \hat{l}_N^t)$ by minimizing Equation (9)
        Update $\mathcal{S}_v^t = \mathcal{S}_v^{t-1} \cup \hat{l}_v^t$ if $\hat{l}_v^t \neq 0$
        Continue = $True$ iff $\hat{l}_v^t \neq 0$ for any $v$
    end while
    return $\{\mathcal{S}_1^t, \ldots, \mathcal{S}_N^t\}$
end procedure
```

We use Tree-Reweighted Message Passing [73] for minimizing Equation (8) or (9) and update the solution set for each video by $\mathcal{S}_v^t = \mathcal{S}_v^{t-1} \cup \hat{l}_v^t$ if $\hat{l}_v^t \neq 0$. The optimization procedure terminates if $\hat{l}_v^t = 0$ for all videos $v$. The greedy approach is described in Algorithm 1. While this does not necessarily converge to the global minimum of Equation (4), it produces satisfying results as we show in our experiments.

The proposed formulation can also be used to infer instances of object classes from videos or images without human context since it can be combined with any type of unary and pairwise potentials. In this work, however, we focus on explicit modeling of human context for the task and therefore introduce potentials that model appearance similarity as well as functionality of the object class.

### *3.3. Unary potentials $\Phi$*

Unary potentials are used to measure the quality of tube $l_v$ in video $v$. It is composed of four aspects each of which aim to select tubes tightly bound to objects and interacted with. They are described as follows.

#### *3.3.1. Appearance Saliency*

Appearance saliency is a commonly used objectness measure since the appearance of an object generally stands out from the background. We base the saliency of the $k^{th}$ frame of a tube on two distributions. While the first captures the RGB or RGB-D distribution computed on region $I_k$ inside the tube, the latter captures the distribution from the region $S_k$ equal to and surrounding $I_k$. Saliency for frame $k$ is then computed as the $\chi^2$ distance between the two. Assuming tube saliency factorizes over individual frames, we have

$$\Phi^{app}(l_v) = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{1}{2}\sum_i \frac{(I_{k,i} - S_{k,i})^2}{I_{k,i} + S_{k,i}}\right). \tag{10}$$

The effect of the unary potential is that it penalizes tubes that are loosly or partially bound around objects. In either case, appearance inside and outside the tube is more similar than for a tightly bound case.

### *3.3.2. Pose-object Relation*

This is a measure to evaluate if the tube is being interacted with by the human. Given the frame $k$, we propose to evaluate the 2d or 3d Euclidean distance between the locally active end effector joint $j_k$ of the human pose and the center $c_k$ of the tube in that frame. To make the measure robust to pose estimation errors and interactions spanning short time durations e.g. interaction with a bowl during *eating cereal*, we perform $\alpha = 0.3$ trimmed mean filtering. Assuming that the measure factorizes over individual frames, we have

$$\Phi^{Pose}(l_v) = \frac{1}{K}\sum_{k=\alpha\cdot K}^{(1-\alpha)\cdot K} \|c_{D(k)} - j_{D(k)}\| \tag{11}$$

where D is a look up table to index over the sorted list of distances.

### *3.3.3. Body part avoidance*

Body part avoidance guides the energy functional towards meaningful solutions in the weakly supervised setting. The need is highlighted in the case of body parts which are consistently present in all videos, thereby guiding the optimization to these trivial solutions. Without the aid of this term, background regions corresponding to body parts such as faces and hands, which occur in all videos, will be selected instead of objects. We model appearance of the body as mixture comprising models for skin, upper and lower bodies. The potential is then defined as

$$\begin{aligned} \Phi^{body}(l_v) &= \max\left\{\bar{p}_{skin}(I), \bar{p}_{upper}(I), \bar{p}_{lower}(I)\right\}, \\ \text{with} \quad \bar{p}_x(I) &= \frac{1}{K}\sum_k p_x(I_k) \end{aligned} \tag{12}$$

where $I_k$ is the color histogram of the tube at frame $k$. We use 5-component Gaussian mixture models for both upper and lower bodies, learned directly using pixels around relevant joints of the estimated pose. We use a generic model for skin [74].

### *3.3.4. Size prior*

A prior on the size of an object is an important cue that can be inferred relative to the human size in human-object interaction scenarios. In other words, there are bounds on the physical size of an object a human can interact with. E.g. interactions with phone, plate and markers are possible, but not with the floor or the cap of a marker. Such video level priors can be helpful when tubes are very small, rendering other potentials unreliable. The prior on the object size is modelled as a Gaussian distribution given as

$$\Phi^{size}(l_v) = \exp\left(\frac{(w_{l_v} - 2w_h)^2 + (h_{l_v} - 2h_h)^2}{2\sigma_h^2}\right) \tag{13}$$

where $(w_h, h_h)$ and $(w_{l_v}, h_{l_v})$ are average width and height of the hand and tube respectively and $\sigma_h$ is 1.5 times the average size of the hand.

### *3.3.5. Unary potential*

The final unary potential is formed by linearly combining the four terms as

$$\begin{aligned} \Phi(l_v) = {} & \lambda_1\Phi^{app}(l_v) + \lambda_2\Phi^{pose}(l_v) \\ & + \lambda_3\Phi^{body}(l_v) + \lambda_4\Phi^{size}(l_v) \end{aligned} \tag{14}$$

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

## *3.4. Pairwise potentials $\Psi$*

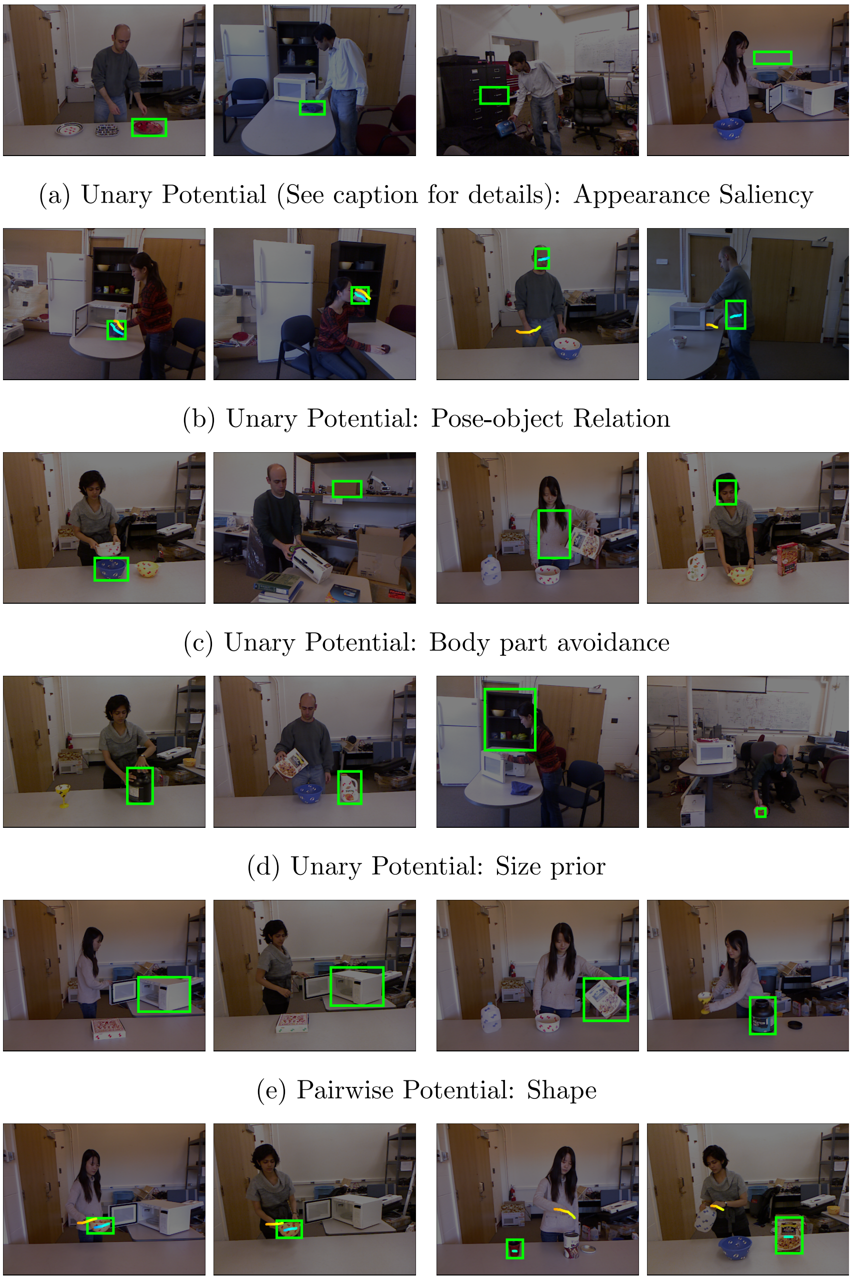
The pairwise potential measures the similarity between two tubes $l_v$ and $l_w$ and is composed of two terms. The first term measures the inter-tube appearance similarty and the second term measures the similarity of their motion during interaction.

### *3.4.1. Shape*

We follow [14] and measure the appearance similarity between two tubes based on PHOG [75]. The appearance of a tube is described by a multiresolution histogram of gradients computed over 50 uniformly sampled frames in the tube. Further, the two sequences are first aligned using dynamic time warping to account for varying object appearance during interaction. The warping is performed using the joint locations of the head, shoulders and hands as features. Since the alignment between two distinct action sequences is meaningless, we retain the original tubes if the alignment error exceeds a certain threshold. The pairwise potential $\Psi^{shape}(l_v, l_w)$ defined as the median $\chi^2$ distance between PHOG features from corresponding frames $k$ of tubes $l_v$ and $l_w$ is given as

$$\Psi^{shape}(l_v, l_w) = \operatorname*{median}_k \left\{\frac{1}{2}\sum_i \frac{\left(P_{\omega_v(k),i} - P_{\omega_w(k),i}\right)^2}{P_{\omega_v(k),i} + P_{\omega_w(k),i}}\right\} \tag{15}$$

where $\omega_u$ 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 the $k^{th}$ frame of tube $l_u$ after warping.



(a) Unary Potential (See caption for details): Appearance Saliency

(b) Unary Potential: Pose-object Relation

(c) Unary Potential: Body part avoidance

(d) Unary Potential: Size prior

(e) Pairwise Potential: Shape

(f) Pairwise Potential: Functionality

Figure 3: Illustrating the unary and pairwise potentials. Bounding boxes in the first two columns favour the energy in Equation (4) by decreasing it in comparison with those on the last two columns. (a)–(d) correspond to unary potentials and illustrate two distinct favorable and an unfavorable cases each. (e)–(f) correspond to pairwise potentials and illustrate a single favorable and unfavorable case. Temporal paths of the most active joint location and the bounding box are marked in yellow and cyan respectively.

|  | [14] | modif-[14] | Equation (3) | Equation (8) | Equation (9) |
|---|---|---|---|---|---|
| ETHZ-Action | 0.063 | 0.249 | 0.447 | 0.439 | **0.471** |
| CAD-120 | 0.039 | 0.246 | 0.410 | 0.393 | **0.423** |
| MPII-Cooking | 0.023 | 0.221 | 0.342 | 0.333 | **0.348** |

Table 1: Average class-IoU of the proposed model for the three datasets. The Equation (9) which infers multiple tubes per video outperforms Equation (3) which extracts a single tube per video and [14] which relies on motion segments and object appearance and ignores object functionality.

### *3.4.2. Functionality*

Assuming relative trajectories of objects with respect to the human correlate with object functionality, we measure the relative Euclidean distance between the center of the tube and the human. After having preprocessed the tubes as for the shape potential, we sample 50-pairs of corresponding frames uniformly. Given frame $k$, we compute the distance between the center $c_{u(k)}$ of the tube $l_u$ and the head position $h_{u(k)}$ and normalize it by the distance between the head and the locally active end effector $j_{u(k)}$:

$$d_{u(k)} = \frac{\|h_{u(k)} - c_{u(k)}\|}{\|h_{u(k)} - j_{u(k)}\|}. \tag{16}$$

While the normalization accounts for lack of 3d information in 2d human poses, it also compensates for varying body sizes in 3d human poses. Given the dynamic time warping functions $\omega_*$, the potential $\Psi^{func}(l_v, l_w)$ is then the median of these differences:

$$\Psi^{func}(l_v, l_w) = \underset{k}{\text{median}} \left\{ |d_{\omega_v(k)} - d_{\omega_w(k)}| \right\}. \tag{17}$$

**Pairwise potential** The final pairwise potential is formed by linearly combining the two terms as

$$\Psi\left(l_v, l_w\right) = \lambda_5 \Psi^{shape}\left(l_v, l_w\right) + \lambda_6 \Psi^{func}\left(l_v, l_w\right) \tag{18}$$

where weighting parameters $\lambda_i$ are learned together with the weights of the unary potential (14) from a validation set.

## 4. Experiments

We evaluate the proposed method on two RGB-D datasets and one RGB dataset[1], which represent a rich variety of modalities: ETHZ-Activity [76], CAD-120 [22] and MPII-Cooking [20]. The ETHZ-Activity is an RGB-D dataset composed of a time of flight and color camera with a resolution of $170 \times 144$ and $640 \times 480$ respectively. The dataset contains 6 different actors each performing high level activities with 12 objects totalling to 143 video sequences. An 8-joint upper body 3d human pose is extracted using a model based method. While interactions are mostly restricted to a single object, there is significant intra-class variation in object appearance due to the interaction. The 12 objects range from being medium sized e.g. *brush* and *teapot* to small sized e.g. *marker* and *videogame*. A typical frame illustrating the relative size of an object is shown in Figure 4.

CAD-120 is an RGB-D dataset captured using the Kinect sensor. Therefore both color and depth images have a resolution of $640 \times 480$. The dataset contains 4 actors performing 10 different high level activities totalling to 120 video sequences. The OpenNI SDK is used to extract human pose consisting of 15 3d whole body joint locations with binary confidence flags for each joint. Noise in the pose is more pronounced for limb joints i.e. hands and legs. Some activities involve multiple instances of the same object e.g. *stacking objects* or multiple objects e.g. *taking medicine* that indicates presence of *medicinebox* and *cup*. It must be noted that the classes *book* and *remote* appear in only three video sequences each.

The MPII-Cooking is a high resolution ($1624 \times 1224$) RGB dataset. It contains 2 high level activities performed by 12 different actors totalling to 65 video sequences. The extracted human pose consists of 8 2d joint locations for the arms. Therefore, in the pairwise potential $\Psi^{func}(l_v, l_w)$ in Equation (17), we replace the location of the head by the mean location of both shoulders. This is a challenging dataset where objects evolve in appearance and frequently undergo occlusions. E.g. *bread* evolves from being a layer of dough to an arrangement of vegetables during the course of preparing a pizza.

For evaluation, objects in all datasets are labelled by drawing tight bounding boxes for every 10th frame and interpolating intermediate bounding boxes.

We also compare with a weakly supervised approach [14] and an unsupervised approach [76]. The method in [76] discovers objects by clustering trajectories of human joint locations. The method in [14] uses motion segments to

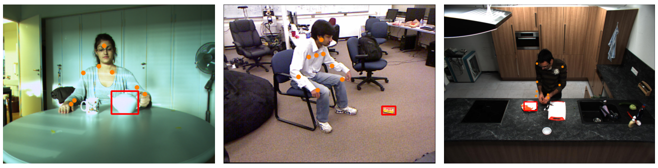
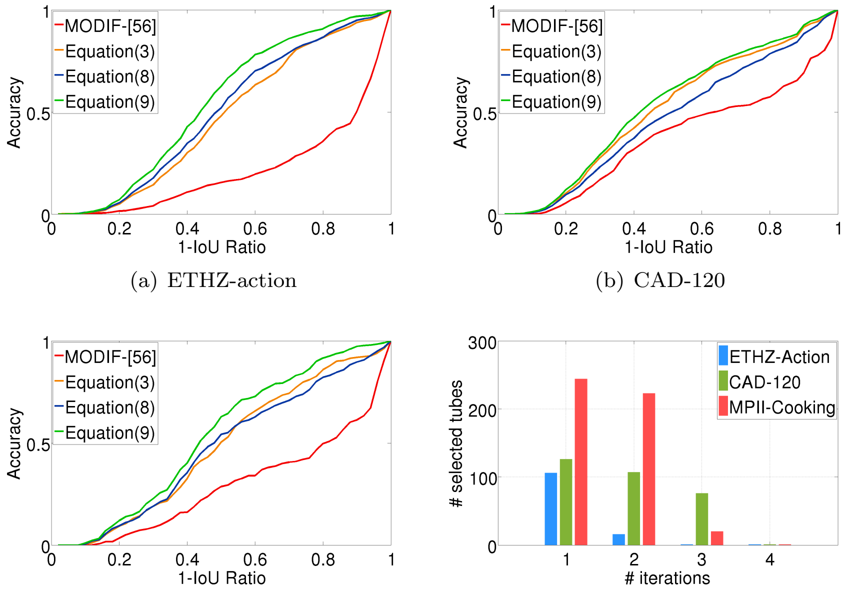


Figure 4: Illustrating human-object interaction from ETHZ-Action dataset, CAD-120 dataset and MPII cooking dataset with human pose overlayed in orange and objects with a red bounding box.

[1] Annotations for all three datasets can be found at `http://ps.is.tue.mpg.de/person/srikantha`



(a) ETHZ-action

(b) CAD-120

(c) MPII-Cooking

(d) Algorithm (1)

Figure 5: (a–c) The accuracy is measured as the fraction of bounding boxes with IoU ratio greater 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 presented is averaged over all classes. (d) Number of selected tubes inferred in each iteration. After the third iteration, the approach has converged because no new tubes are added to the set of selected tubes $\mathcal{S}_v^t$.

generate object proposals which are then fed into an energy functional similar to Equation (3). The unary and pairwise potentials are inspired purely by appearance features. While unary potentials are composed of objectness [77], intra-tube shape consistency and bounding-box heuristics, pairwise potentials are based on inter-tube shape consistency. Their solution involves extracting one tube per video which best represents the latent object. In a similar framework [19] generates object proposals as described in Section 3.1 and uses the potentials as in Section 3.3.

### 4.1. Inference

The output of the system is a collection of tubes that best describe an object class common to all input videos. detected instances of object classes are shown in Figure 8. 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.

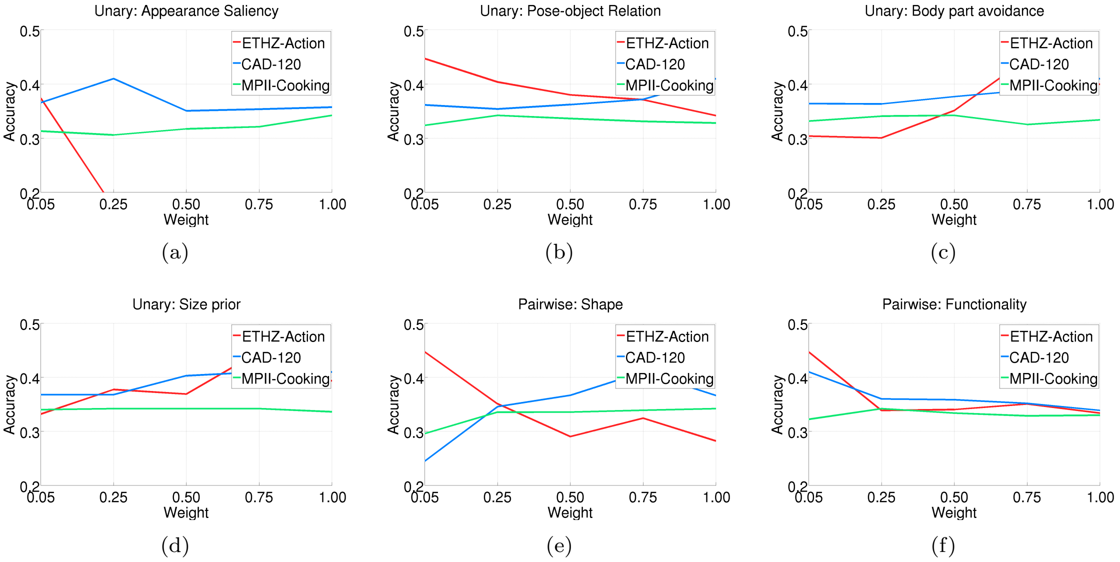
To learn the parameters $\alpha$, $\lambda$ and $\gamma$ of the energy model, we use ground-truth object annotations of one randomly chosen object class per dataset as validation: *puncher* (ETHZ-Action), *milkbox* (CAD-120) and *whisker* (MPII-Cooking). We perform a grid-search in $\{0.05, 0.25, 0.50, 0.75, 1.00\}$ to set these parameters and take the configuration that maximizes class-IoU for the validation class. We therefore exclude validation classes from all performance evaluations that follow.

### 4.2. Comparison

In the context of detecting objects from videos with activities, the experiments show that naive motion based segmentation as in [14] and object proposal method [78] fail at varying levels of severity. Improved performance is shown in [19] due to improved object proposals as generated by Section 3.1 and the inclusion of object functionality in Equation (3). We show that extending the approach [19] to select a varying number of tubes from each video improves the quality of inferred tubes and the subsequent object detection performance. We find the framework presented in Equation (8) to be prone to noise thereby often yielding suboptimal solutions and the independence assumption incorporated in Equation (9) helps alleviate this limitation. We further present details of the experiments below.

Firstly, we compare an object proposal technique [78] against the proposed tube generation process. We consider every $10^{th}$ frame in the ETHZ-Action dataset for this experiment. The recall of [78] for ($10^2$, $10^3$, $10^4$) proposals per image was (0.19, 0.58, 0.67) respectivly, against 0.65 for 30 tube proposals as generated in Section 3.1.

Regarding overall accuracy, we compare a method for learning from weakly labelled videos [14] with an approach that optimizes Equation (3). The average class-IoU for all three datasets is presented in Table 1. Optimizing



(a) (b) (c)

(d) (e) (f)

Figure 6: Sensitivity of parameters for Equations (14) and (18). Accuracy is measured by average class-IoU.

Equation (3) outperforms [14] significantly. The poor performance of [14] is due to the inferior quality of object proposals which are extracted based on dominant motion segments, which overlap mostly with human body parts instead of objects. We therefore evaluate the method by replacing object proposals with those generated from Section 3.1 but retaining the energy functional proposed in [14]. We denote the modified approach as modif-[14] in Table 1. While modif-[14] performs significantly better than its baseline [14], it still underperforms when compared to Equation (3).

Equations (8), (9) extend Equation (3) by inferring multiple tubes. While the former lags behind the baseline Equation (3) on all three datasets, the latter performs favorably in ETHZ-Action and CAD-120 datasets and comparably in the MPII-Cooking dataset. To reason about the superior performance of Equation (9) against that of Equation (8), we calculated the energy as in Equation (4) for the solutions obtained by both methods. We found that energies pertaining to Equation (9) were lower in 9 out of 12 classes in ETHZ-Action and 6 out of 9 classes in CAD-120 dataset. A possible reasoning for this could be that assuming independence between iterations as in Equation (9) can better handle noise without propagating it into further iterations.

To further evaluate the quality of inferred tubes, we define class-accuracy as the fraction of bounding boxes with an IoU ratio greater than a given threshold. Figure 5 shows class-accuracy averaged over all classes for decreasing IoU ratios. Because of the inferior performance of [14], we show the accuracy for modif-[14]. As can be seen, modif-[14] underperforms in all three datasets verifying the sub-optimally of related potentials. The introduction of new potentials as in [19] shows improvements, the biggest of which is for the ETHZ-Action dataset at 1-IoU=0.8 where the former performs at 0.36 and the latter at 0.86. Although introducing multi-tube inference as in Equation (8) results in reduced performance, the independence assumption in Equation (9) is favorable on all three datasets. Significant improvements are found in ETHZ-Action and MPII-Cooking at 1-IoU=0.5 with around 10% increase in accuracy from the performance of Equation (3). At IoU=0.5, the accuracies of Equations (9), (3) and modif-[14] are (0.62, 0.48, 0.16) for ETHZ-Action, (0.60, 0.56, 0.42) for CAD-120 and (0.63, 0.53, 0.29) for the MPII-Cooking dataset respectively.

The number of tubes selected in each iteration for the datasets is shown in Figure 5(d). For the ETHZ-Action dataset, all tubes are selected after two iterations. For the other two datasets, the approach converges after three iterations. Using the multiple instance inference of Equation (9), a total of 124, 310 and 488 tubes are selected for the ETHZ-Action, CAD-120 and MPII-Cooking dataset respectively. As a comparison, single instance inference of Equation (3) selects only 106, 126 and 244 tubes for the datasets.

Regarding running times, the CPU only implementation takes about 1 hour to extract 30 tubes per video and about 5 hours to precompute unary and pairwise potentials. The inference procedure is fast and takes about 15 seconds for a collection of 20 videos.

### *4.3. Evaluating parameter sensitivity*

In our experiments, we have estimated the parameters of Equations (14) and (18) on a validation set as described in Section 4.1. In order to show the sensitivity of parameters, we vary each of the learned weights and measure average class-IoU as shown in Figure 6(a)–(f). As can be seen, while varying almost any potential has minimal

|  | modif-[14] | Equation (3) | APP | APP+SIZ | FUN | APP+FUN | FUN+SIZ |
|---|---|---|---|---|---|---|---|
| ETHZ-Action | 0.249 | 0.447 | 0.192 | 0.305 | 0.292 | 0.312 | 0.390 |
| CAD-120 | 0.246 | 0.410 | 0.168 | 0.191 | 0.147 | 0.202 | 0.350 |
| MPII-Cooking | 0.221 | 0.342 | 0.079 | 0.149 | 0.229 | 0.235 | 0.288 |

Table 2: Studying the contribution of various potential groups in Equation (3). Average class-IoU is presented for (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.

|  | $\Phi^{app}$ | $\Phi^{pose}$ | $\Phi^{body}$ | $\Phi^{size}$ | $\Psi^{shape}$ | $\Psi^{func}$ |
|---|---|---|---|---|---|---|
| ETHZ-Action | -3.27 | -11.40 | -6.09 | -13.17 | -2.43 | -3.00 |
| CAD-120 | -9.48 | -0.85 | -6.38 | -9.19 | -10.06 | -11.71 |
| MPII-Cooking | -10.33 | -7.47 | -7.47 | -3.54 | -9.79 | -34.00 |

Table 3: Percentage change in average class-IoU performance when any given potential is discarded from Equation (9).

effect on MPII-Cooking, the effects are more drastic for ETHZ-Action. The performance on CAD-120 is sensitive to variations in $\Phi^{body}$ and $\Psi^{shape}$.

### *4.4. Impact of Potentials*

We group the potentials into three categories to study the nature of contributions from the designed potentials. They are: APP consisting of potentials that are inherent to object appearance $\{\Phi^{app}, \Psi^{shape}\}$, SIZ denotes the size prior $\{\Phi^{size}\}$ and FUN consisting of potentials derived from human-object interaction $\{\Phi^{pose}, \Phi^{body}, \Psi^{func}\}$. Table 2 presents the performance of Equation (3) under various group combinations.

The foremost observation is that the group APP underperforms in comparison with modif-[14] for all datasets. This is an expected fall in performance due to the difference in the representation of appearance information by both methods. The performance improves upon adding the size prior (APP+SIZ). The importance of incorporating human-object interaction is seen when the functionality terms (FUN) outperform modif-[14] and APP on both ETHZ-Action and MPII-Cooking datasets. Further, combination of (FUN+APP) outperforms individual settings indicating that both groups encode complementary information. Finally, the pair of (FUN+SIZ) performs best amongst all proper subset combinations attaining more than 80% of the maximum recorded performance. This indicates that all potential groups are important for achieving maximum performance.

We now study the effect of discarding a single potential from the model in Equation (9). Corresponding percentage change in class-IoU performance is presented in Table 3. It can be observed that eliminating any potential causes a drop in performance. Appearance based features have minimal impact on the ETHZ-Action dataset as they are not reliable for small objects. Discarding $\Psi^{func}$ most adversely affects the MPII-Cooking dataset due to closer interaction between human and objects in comparison with the other two datasets. On the other hand, discarding $\Phi^{pose}$ has the least impact on the CAD-120 dataset. This is because the inferred human pose is noisy due to missing joints and poor localization accuracy. In fact a, qualitative evaluation confirmed that the pose quality for CAD-120 is the lowest among the three datasets. $\Phi^{body}$ and $\Phi^{size}$ reduce the performance for all three datasets. Due to the small size of the objects in ETHZ-Action, $\Phi^{size}$ has the biggest impact on this dataset. Studying the unaries and pairwise potentials using Equations (3), (8) showed similar trends.

Further, we study the robustness of pose-related potentials with respect to strong pose estimation noise on the CAD-120 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 0.423 (see Table 1). The performance, however, is still higher than without using these potentials (see APP+SIZ in Table 2).

### *4.5. Evaluating object models*

We now evaluate the quality of inferred tubes from Equation (9) in terms of object detection performance. Training and testing data are obtained by defining splits on each dataset such that they share no common actors. For training, we consider data from 3 out of 4 actors in CAD-120, 5 out of 6 actors in ETHZ-Action and 9 out of 12 actors in the MPII-Cooking dataset. The rest of the data i.e. *Subject-1* for CAD-120, *actor-14* for ETHZ-Action and $\{$*s18,s19,s20*$\}$ for MPII-Cooking is used for testing.

For object detection, we use a Hough forest [32] with 5 trees. Each tree is trained until a maximal depth of 25 and with 50,000 positive and 50,000 negative patches (drawn uniformly from the background). We do not make use of depth data for this experiment. For comparison, we use manually annotated bounding boxes of training images, i.e. every $10^{th}$ frame of the training sequences. This is denoted as ‘GTr.’ in Table 4. The ‘Infer’ training data is based on an equal number of frames from the automatically extracted tubes by Equation (9).

| Class | GTr. | Infer | Class | GTr. | Infer | Class | GTr. | Infer | Class | GTr. | Infer |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ETHZ-Action** | | | | | | | | | | | |
| brush | 45.1 | 38.0 | calcul. | 100.0 | 100.0 | camera | 83.5 | 73.0 | remote | 49.4 | 36.7 |
| mug | 38.0 | 30.2 | headph. | 69.8 | 63.7 | marker | 39.7 | 39.7 | teapot | 63.2 | 59.2 |
| videog. | 78.3 | 77.6 | roller | 99.6 | 66.1 | phone | 0.05 | 11.9 | **Avg.** | 60.6 | 54.2 |
| **CAD-120** | | | | | | | | | | | |
| book | 11.2 | 03.2 | medbox. | 58.3 | 53.3 | bowl | 24.5 | 24.5 | mwave. | 71.4 | 71.0 |
| box | 24.4 | 21.5 | plate | 16.2 | 14.3 | cup | 14.8 | 12.9 | remote | 14.1 | 08.3 |
| | | | cloth | 20.1 | 18.6 | **Avg.** | 29.4 | 25.3 | | | |
| **MPII-Cooking** | | | | | | | | | | | |
| bowl | 69.2 | 64.4 | spiceh. | 100.0 | 100.0 | bread | 25.5 | 13.2 | squeez. | 61.5 | 61.5 |
| plate | 43.4 | 43.4 | tin | 33.0 | 26.4 | grater | 02.2 | 01.2 | **Avg.** | 47.8 | 44.3 |

Table 4: Average precision (%) for different datasets comparing object models built from ground truth data (GTr.) and inferred data (Infer) from Equation (9).

The results show that optimal performance is achieved for categories like *calculator*, *marker* in ETHZ-Action, *bowl*, *microwave* in CAD-120 and *spiceholder*, *squeezer* in MPII-Cooking. On the other hand, a loss in performance is observed for many categories due to weaker supervision. This is due to noisier extracted tubes in comparison with manually annotated data. Nevertheless, performances of the object detectors trained on weakly supervised videos achieve 89.4% (ETHZ-Action), 86.1% (CAD-120) and 92.6% (MPII-Cooking) of that from full supervision.

We now compare the object detection performance when training data is obtained from Equations (3), (8), (9) in Figure 7. It can be observed that object detectors based the Equation (9) generally outperform those from Equations (3) and (8). Particularly, the average precision is improved when compared to Equation (3) in all three datasets from 53.2% to 54.2% for ETHZ-Action, 24.4% to 25.3% for CAD-120 and 35.3% to 44.3% in the MPII-Cooking dataset. However, there is a small loss in performance for a few classes such as *camera, headphone* in ETHZ-Action and *book, remote* in the CAD-120 dataset.

We also compare with [76] which is an unsupervised approach that segments and clusters videos based on pose features. [76] 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-Action is 24.85% in comparison to 54.20% of our approach.

### *4.6. Refining objectness using object detectors*

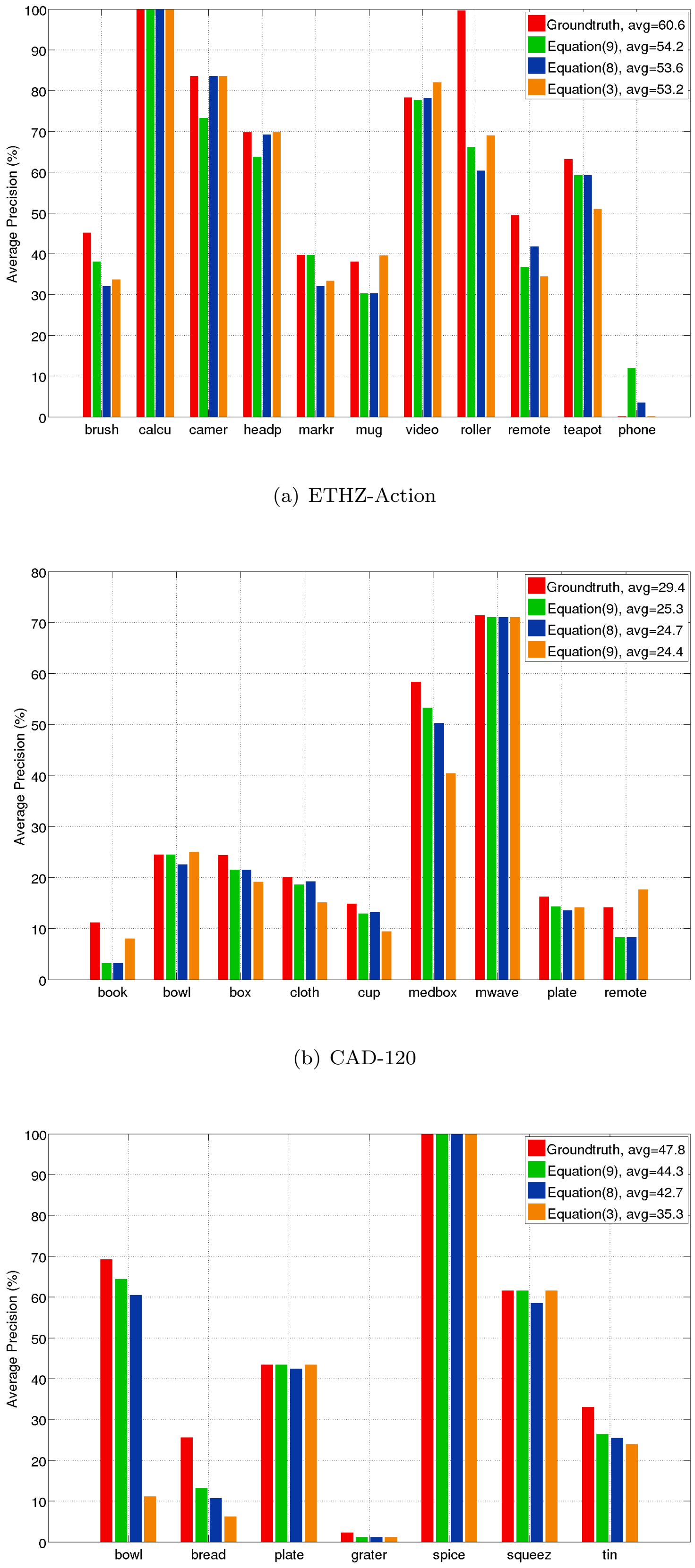
Approaches like [79, 80] propose a weakly supervised method where a detector is initialized using a few seed examples and later refined by incorporating new detections. In order to evaluate if iterating between training the detector and inferring training data from videos improves accuracy, we apply the object detector (Section 4.5) to the tubes and use the detector confidence as a fifth unary potential in Equation (14). The detector confidence is obtained by max pooling frame-wise detection confidences over any given tube. The process is iterated until the set of selected tubes does not change anymore.

Repeating the experiments as described in Sections 4.1–4.2 with the augmented model, the procedure for ETHZ-Action and CAD-120 terminated after the first iteration without any improvement in average class-IoU measure. However, the procedure for MPII-Cooking terminated after two iterations and yielded a marginal improvement from 0.342 to 0.343 (cf. Table 2). The object detection performance remained unchanged for all datasets.

## 5. Conclusion

In this work, we have addressed the problem of detecting instances of small and medium sized objects from weakly labelled activity videos. Our experiments show that approaches relying entirely on object motion or appearance fail for this task. Although using only object appearance is shown to be insufficient, coupling it with object functionality leads to greatly improved performance. An interesting aspect is that the results reveal the complementary nature of functionality and appearance related potentials for detecting objects. In order to maximize utilization of data, we propose a framework for inferring multiple object instances from each video which is solved using a greedy approach. The superior quality of these tubes are verified by the experiments. The generalization capabilities of our approach are 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 86% and 92% of the performance of a fully supervised approach for object detection.

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



(a) ETHZ-Action

(b) CAD-120

(c) MPII-Cooking

Figure 7: Average precision (%) for object detection on different datasets given training data from groundtruth and from Equations (9), (8) and (3).

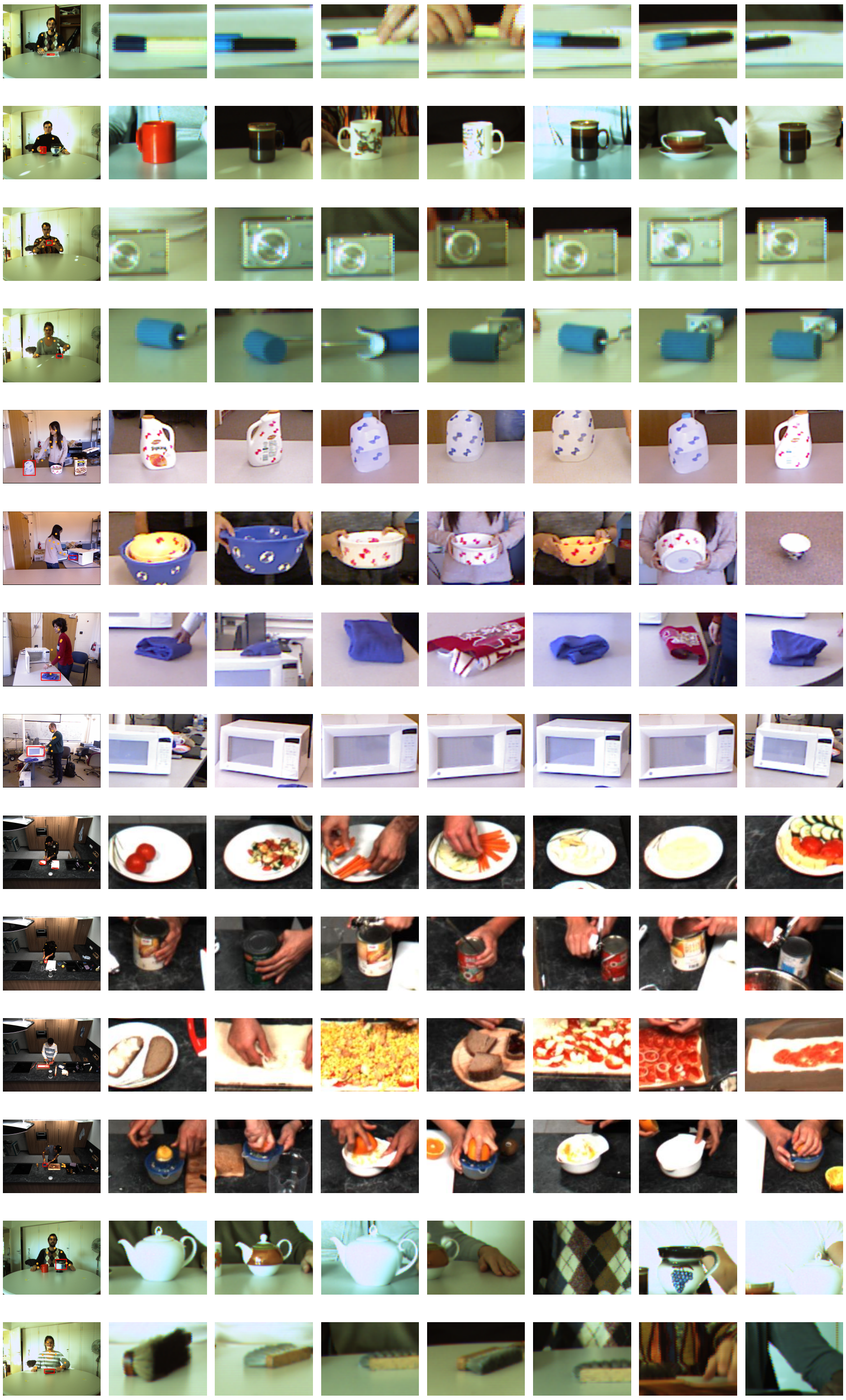


Figure 8: Detected instances of the object classes as in Equation (3): *Marker, Mug, Camera, Roller, Milkbox, Bowl, Cloth, Microwave, Plate, Tin, Bread, Squeezer* and Failure cases *Teapot, Brush.* The first image in each row shows relative object size by illustrating a typical action scene with overlayed human pose and a bounding box around the object of interest. Since the objects are relatively small, images are best viewed by zooming in.

# References

[1] M. Everingham, L. Van Gool, C. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, IJCV 88 (2010) 303–338.

[2] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: CVPR, 2014, pp. 1891–1898.

[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS, 2012, pp. 1097–1105.

[4] X. Zhu, C. Vondrick, D. Ramanan, C. Fowlkes, Do we need more training data or better models for object detection?., in: BMVC, Vol. 3, 2012, p. 5.

[5] M. B. Blaschko, A. Vedaldi, A. Zisserman, Simultaneous object detection and ranking with weak supervision, in: NIPS, 2010, pp. 235–243.

[6] O. Chum, A. Zisserman, An exemplar model for learning object classes, in: CVPR, 2007, pp. 1–8.

[7] Y. J. Lee, K. Grauman, Learning the easy things first: Self-paced visual category discovery., in: CVPR, 2011, pp. 1721–1728.

[8] M. Rubinstein, A. Joulin, J. Kopf, C. Liu, Unsupervised joint object discovery and segmentation in internet images, in: CVPR, 2013, pp. 1939–1946.

[9] J. M. Winn, N. Jojic, Locus: Learning object classes with unsupervised segmentation., in: ICCV, 2005, pp. 756–763.

[10] R. G. Cinbis, J. Verbeek, C. Schmid, Multi-fold mil training for weakly supervised object localization, in: CVPR, 2014, pp. 2409–2416.

[11] H. Bilen, M. Pedersoli, T. Tuytelaars, Weakly supervised object detection with posterior regularization, in: BMVC, 2014.

[12] H. Bilen, M. Pedersoli, T. Tuytelaars, Weakly supervised object detection with convex clustering, in: CVPR, 2015, pp. 1081–1089.

[13] C. Leistner, M. Godec, S. Schulter, A. Saffari, M. Werlberger, H. Bischof, Improving classifiers with unlabeled weakly-related videos., in: CVPR, 2011, pp. 2753–2760.

[14] A. Prest, C. Leistner, J. Civera, C. Schmid, V. Ferrari, Learning object class detectors from weakly annotated video, in: CVPR, 2012, pp. 3282–3289.

[15] D. Ramanan, D. A. Forsyth, K. Barnard, Building models of animals from video., PAMI 28 (8) (2006) 1319–1334.

[16] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: CVPR, 2008, pp. 1–8.

[17] B. Leibe, A. Leonardis, B. Schiele, Robust object detection with interleaved categorization and segmentation, IJCV 77 (1-3) (2008) 259–289.

[18] B. Pepik, R. Benenson, T. Ritschel, B. Schiele, What is holding back convnets for detection?, in: GCPR, 2015, pp. 517–528.

[19] A. Srikantha, J. Gall, Discovering object classes from activities, in: ECCV, 2014, pp. 415–430.

[20] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele, A database for fine grained activity detection of cooking activities, in: CVPR, 2012, pp. 1194–1201.

[21] J. Gall, A. Fossati, L. Van Gool, Functional categorization of objects using real-time markerless motion capture., in: CVPR, 2011, pp. 1969–1976.

[22] H. Koppula, R. Gupta, A. Saxena, Learning human activities and object affordances from rgb-d videos, IJRR 32 (8) (2013) 951–970.

[23] C. Desai, D. Ramanan, Detecting actions, poses, and objects with relational phraselets, in: ECCV, 2012, pp. 158–172.

[24] M. Sun, S. Savarese, Articulated part-based model for joint object detection and pose estimation, in: ICCV, 2011, pp. 723–730.

[25] E. Hsiao, M. Hebert, Occlusion reasoning for object detectionunder arbitrary viewpoint, PAMI 36 (9) (2014) 1803–1815.

[26] T. Gao, B. Packer, D. Koller, A segmentation-aware object detection model with occlusion handling, in: CVPR, 2011, pp. 1361–1368.

[27] B. Pepikj, M. Stark, P. Gehler, B. Schiele, Occlusion patterns for object class detection, in: CVPR, 2013, pp. 3286–3293.

[28] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, A. Yuille, Detect what you can: Detecting and representing objects using holistic models and body parts, in: CVPR, 2014.

[29] B. Leibe, A. Leonardis, B. Schiele, Combined object categorization and segmentation with an implicit shape model, in: Workshop on statistical learning in computer vision, ECCV, Vol. 2, 2004, p. 7.

[30] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: ECCV, 2010, pp. 143–156.

[31] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, IJCV 73 (2) (2007) 213–238.

[32] J. Gall, A. Yao, N. Razavi, L. Van Gool, V. Lempitsky, Hough forests for object detection, tracking, and action recognition, PAMI 33 (11) (2011) 2188–2202.

[33] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, in: CVPR, 2011, pp. 1385–1392.

[34] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: CVPR, 2014, pp. 580–587.

[35] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: NIPS, 2015, pp. 91–99.

[36] R. Girshick, Fast r-cnn, in: ICCV, 2015, pp. 1440–1448.

[37] P. Yadollahpour, D. Batra, G. Shakhnarovich, Discriminative re-ranking of diverse segmentations, in: CVPR, 2013, pp. 1923–1930.

[38] T. Brox, L. Bourdev, S. Maji, J. Malik, Object segmentation by alignment of poselet activations to image contours, in: CVPR, 2011, pp. 2225–2232.

[39] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, W. Buntine, Unsupervised object discovery: A comparison, IJCV 88 (2) (2010) 284–302.

[40] S. Schulter, C. Leistner, P. M. Roth, H. Bischof, Unsupervised object discovery and segmentation in videos, in: BMVC, 2013, pp. 1–12.

[41] Y. Wu, J. Lim, M.-H. Yang, Online object tracking: A benchmark, in: CVPR, 2013, pp. 2411–2418.

[42] H. Pirsiavash, D. Ramanan, C. C. Fowlkes, Globally-optimal greedy algorithms for tracking a variable number of objects, in: CVPR, 2011, pp. 1201–1208.

[43] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L. Van Gool, Online multiperson tracking-by-detection from a single, uncalibrated camera, PAMI 33 (9) (2011) 1820–1833.

[44] Y.-X. Wang, M. Hebert, Model recommendation: Generating object detectors from few samples, in: CVPR, 2015, pp. 1619–1628.

[45] Y. J. Lee, J. Kim, K. Grauman, Key-segments for video object segmentation, in: ICCV, 2011, pp. 1995–2002.

[46] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, in: ICCV, 2009, pp. 309–316.

[47] A. Gaidon, M. Marszalek, C. Schmid, Mining visual actions from movies, in: BMVC, 2009, pp. 125–1.

[48] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: CVPR, 2009, pp. 2929–2936.

[49] C. Doersch, S. Singh, A. Gupta, J. Sivic, A. Efros, What makes paris look like paris?, ACM Transactions on Graphics 31 (4).

[50] V. Ordonez, G. Kulkarni, T. L. Berg, Im2text: Describing images using 1 million captioned photographs, in: NIPS, 2011, pp. 1143–1151.

[51] B. Ommer, T. Mader, J. Buhmann, Seeing the Objects Behind the Dots: Recognition in Videos from a Moving Camera, IJCV 83 (2009) 57–71.

[52] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Is object localization for free?–weakly-supervised learning with convolutional neural networks, in: CVPR, 2015, pp. 685–694.

[53] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, A. Efros,

Scene semantics from long-term observation of people, in: ECCV, 2012, pp. 284–298.

[54] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, J. Sivic, People watching: Human actions as a cue for single-view geometry, in: ECCV, 2012, pp. 259–274.

[55] H. Grabner, J. Gall, L. Van Gool, What makes a chair a chair?, in: CVPR, 2011, pp. 1529–1536.

[56] A. Gupta, S. Satkin, A. A. Efros, M. Hebert, From 3d scene geometry to human workspace, in: CVPR, 2011, pp. 1961–1968.

[57] Y. Jiang, H. Koppula, A. Saxena, Hallucinated humans as the hidden context for labeling 3d scenes, in: CVPR, 2013, pp. 2993–3000.

[58] H. Kjellström, J. Romero, D. Kragic, Visual object-action recognition: Inferring object affordances from human demonstration, CVIU (2010) 81–90.

[59] P. Peursum, G. West, S. Venkatesh, Combining image regions and human activity for indirect object recognition in indoor wide-angle views, in: ICCV, 2005, pp. 82–89.

[60] A. Pieropan, C. H. Ek, H. Kjellstrom, Functional object descriptors for human activity modeling, in: ICRA, 2013, pp. 1282–1289.

[61] M. W. Turek, A. Hoogs, R. Collins, Unsupervised learning of functional categories in video scenes, in: ECCV, 2010, pp. 664–677.

[62] R. Filipovych, E. Ribeiro, Recognizing primitive interactions by exploring actor-object states, in: CVPR, 2008, pp. 1–7.

[63] A. Gupta, L. Davis, Objects in action: An approach for combining action understanding and object perception, in: CVPR, 2007, pp. 1–8.

[64] D. Moore, I. Essa, M. Hayes, Exploiting human actions and object context for recognition tasks, in: ICCV, 1999, pp. 80–86.

[65] M. Jain, J. van Gemert, C. Snoek, What do 15,000 object categories tell us about classifying and localizing actions?, in: CVPR, 2015, pp. 46–55.

[66] M. Jain, J. C. van Gemert, T. Mensink, C. G. Snoek, Objects2action: Classifying and localizing actions without any video example, in: ICCV, 2015, pp. 4588–4596.

[67] A. Fathi, X. Ren, J. Rehg, Learning to recognize objects in egocentric activities, in: CVPR, 2011, pp. 3281–3288.

[68] T. Brox, J. Malik, Object segmentation by long term analysis of point trajectories, in: ECCV, 2010, pp. 282–295.

[69] T. Brox, J. Malik, Large displacement optical flow: descriptor matching in variational motion estimation, PAMI 33 (3) (2011) 500–513.

[70] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, PAMI 24 (5) (2002) 603–619.

[71] P. F. Felzenszwalb, D. P. Huttenlocher, Efficient graph-based image segmentation, IJCV 59 (2) (2004) 167–181.

[72] T. Deselaers, B. Alexe, V. Ferrari, Localizing objects while learning their appearance., in: ECCV, Vol. 6314, 2010, pp. 452–466.

[73] V. Kolmogorov, Convergent tree-reweighted message passing for energy minimization, PAMI 28 (10) (2006) 1568–1583.

[74] M. Jones, J. Rehg., Statistical color models with application to skin detection, IJCV 46 (1) (2002) 81–96.

[75] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: ACM Int. Conf. on Image and Video Retrieval, 2007, pp. 401–408.

[76] A. Fossati, J. Gall, H. Grabner, X. Ren, K. Konolige (Eds.), Consumer Depth Cameras for Computer Vision, Springer, 2013, Ch. Human Body Analysis.

[77] B. Alexe, T. Deselaers, V. Ferrari, What is an object?, in: CVPR, 2010, pp. 73–80.

[78] S. Manen, M. Guillaumin, L. Van Gool, Prime object proposals with randomized prim's algorithm, in: ICCV, 2013, pp. 2536–2543.

[79] P. Siva, T. Xiang, Weakly supervised object detector learning with model drift detection, in: ICCV, 2011, pp. 343–350.

[80] I. Misra, A. Shrivastava, M. Hebert, Watch and learn: Semi-supervised learning for object detectors from video, in: CVPR, 2015, pp. 3593–3602.