Explicit modeling of human-object interactions in realistic videos

Alessandro Prest, Vittorio Ferrari, Cordelia Schmid

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

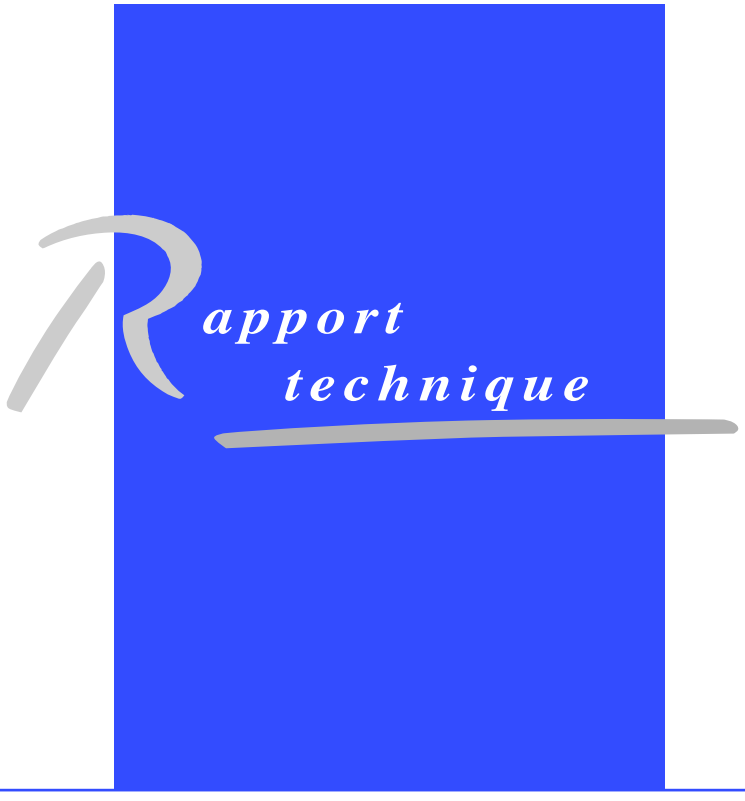
Explicit modeling of human-object interactions in realistic videos

Alessandro Prest — Vittorio Ferrari — Cordelia Schmid

N° 0411

Septembre 2011

Vision, Perception and Multimedia Understanding



ISSN 0249-0803 ISRN INRIA/RT--0411--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

Explicit modeling of human-object interactions in realistic videos

Alessandro Prest*, Vittorio Ferrari†, Cordelia Schmid‡

Theme : Vision, Perception and Multimedia Understanding
Équipes-Projets LEAR

Rapport technique n° 0411 — Septembre 2011 — 25 pages

Abstract: We introduce an approach for learning human actions as interactions between persons and objects in realistic videos. Previous works typically represent actions with low-level features such as image gradients or optical flow. In contrast, we explicitly localize in space and track over time both the object and the person, and represent an action as the trajectory of the object wrt to the person position. Our approach relies on state-of-the-art approaches for human [28] and object detection [11] as well as tracking [3]. We show that this results in human and object tracks of sufficient quality to model and localize human-object interactions in realistic videos. Our human-object interaction features capture relative trajectory of the object wrt the human.

Experimental results on the *Coffee & Cigarettes* [22] and the video dataset of [17] show that (i) our explicit human-object model is an informative cue for action recognition; (ii) it is complementary to traditional low-level descriptors such as 3D-HOG extracted over human tracks. When combining our human-object interaction features with 3D-HOG features [20], we show to improve over their separate performance as well as over the state of the art.

Key-words: Action Recognition, Human-Object Interaction, Video Analysis.

* A. Prest is with the Computer Vision Laboratory at ETH Zurich and with the LEAR team at INRIA Grenoble.
† V. Ferrari is with the Computer Vision Laboratory at ETH Zurich.
‡ C. Schmid is with the LEAR team at INRIA Grenoble.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Résumé :

Mots-clés :

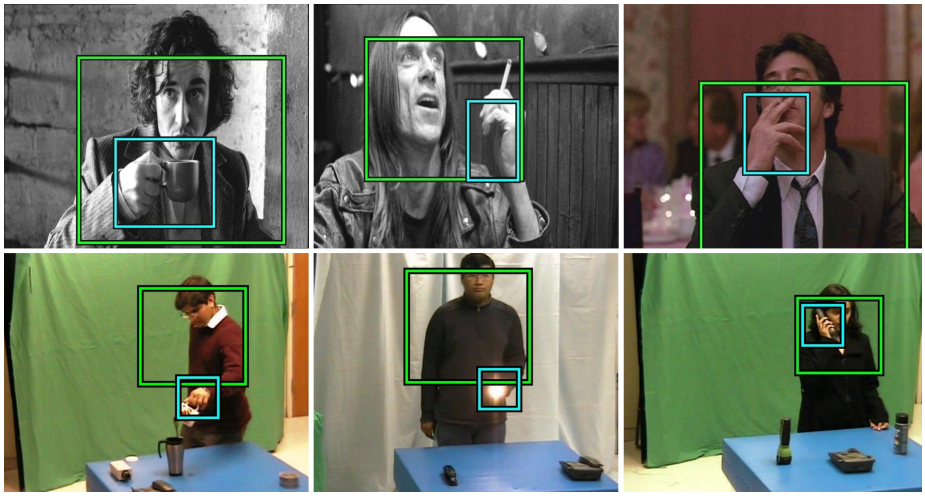


Figure 1: **Human-object interactions**. *Top row: one drinking and two smoking instances from the Coffee & Cigarettes dataset [22]. Second row: examples from the dataset of [17]. Human and object locations automatically obtained by our method are indicated in green and cyan respectively.*

1 Introduction

Human action recognition is an open problem in computer vision. It is important for a wide range of applications, such as video indexing and surveillance. It is challenging due to the high variety of human appearances and poses within an action class. In this paper we focus on actions defined by the interaction between a person and an object, such as drinking and smoking.

Many previous approaches represent actions by distributions of low-level descriptors such as bags of space-time interest points [6, 21, 32] or describe the action as a distribution over point motion features localized in space and time on the human [10, 20, 22, 24, 35].

In this paper, we propose a novel approach that has an explicit notion of the action object and represents an action by spatio-temporal descriptors dedicated to human-object interactions. These include the relative motion of the object with respect to the human, which is typically highly distinctive for the action. Measuring these features involves automatically localizing the human and the action object and tracking them over time as shown in fig. 1. Our method is especially designed to do this in realistic videos, such as feature films. It does not involve any component that depends on background subtraction, which makes it suitable for any camera and background motion. Moreover, the method builds on state-of-the-art object detection techniques [11, 28] operating in single frames, and robustly links detections over time even across many frames where the object was missed, again using a state-of-the-art approach [3]. Finally, our technique takes advantage of the temporal continuity in video to reduce the amount of supervision needed to learn an appearance model of the action object as well as the interaction model. As a result, it can be trained with a modest amount of annotation: for each video clip of the action class we only need a

spatio-temporal cuboid on the person and a bounding-box on the action object in *one* frame.

We evaluate our method on the highly challenging task of spatio-temporal action localization on the *Coffee & Cigarettes* dataset [22], and on the simpler task of action classification on the dataset of [17]. Our experiments demonstrate that (i) our human-object interaction model enables action localization and classification already on its own (sec. 7); (ii) it captures information complementary to existing low-level descriptors such as 3D-HOG computed over human tracks [20], i.e., their combination performs better than either alone and improves over the state of the art on *Coffee & Cigarettes* [20]. Moreover our approach matches the performance of Gupta et al. [17] on their dataset while using less supervision for training. In the rest of the paper we refer to this dataset as the Gupta video dataset.

The rest of this paper is organized as follows. Sec. 3 first gives an overview of our method, and then sections 4 to 6 explain its components in detail. Sec. 4 explains our algorithm to robustly detect and track humans and objects in realistic videos. For this we employ state-of-the-art methods for detecting humans [28] and objects [11], as well as for tracking them over time [3]. This is a necessary step towards our human-object interaction model, which is the main contribution of this paper (sec. 5). In sec. 6 we build a complete action recognition classifier by combining our interaction model with traditional low-level cues, and finally present experiments in sec. 7.

2 Related work

Many existing approaches for action recognition rely on simple measurements such as optical flow or spatio-temporal gradients extracted from video clips. An example are the popular bags of spatio-temporal features, initially introduced in [6, 32, 41]. These techniques extract spatio-temporal features over video clips, quantize them and use a frequency histogram to represent the clips. Recent extensions model the temporal structure of actions as a composition of smaller sub-parts [14, 21, 25]. Furthermore, they determine the temporal extent of video clips optimal for a bag-of-features representation in realistic movies [7, 31].

Another line of work describes the human tracks based on low-level features such as optical flow [8] or based on the silhouette of the humans [1, 40, 15]. Specifically, [40, 15] propose human-centered approaches for action recognition based on spatio-temporal volumes (STV) obtained by accumulating silhouette information over time. They then extract information such as speed, direction and shape to characterize the STV. In [1] they extract silhouettes from a single view and aggregate differences between subsequent frames of an action sequence resulting in a binary motion energy image. Temporal information is included through a motion history image. The method proposed in [8] operates on sports footage. They compensate camera movement by tracking the person and calculate optical flow in person-centered tracks.

All of the above mentioned human-centric approaches operate either with static cameras, i.e., human can be located based on background subtraction, or with simple backgrounds from which human can be extracted easily, as for example football or ice hockey fields.

More recent human-centric approaches [22, 24, 20, 30] deal with action localization in realistic video. Laptev and Perez [22] aggregate local spatio-temporal features over time into a spatio-temporal grid. They use keyframe priming to refine the output of their method. In [24] authors also adopt a human-centric approach where vocabularies of local motion and shape features are combined with a voting approach. The method proposed in [20] localizes actions in space and time by first extracting human tracks and then detecting specific actions within the tracks using a sliding window classifier. Actions are described by track-aligned 3D-HOG features. These features are shown to be complementary to our human-object interaction descriptors and are incorporated in our final classifier.

Weakly-supervised approaches by Ikizler et al. [18, 19] attempt to decrease the amount of supervision necessary for training action classifiers. Training videos for learning actions are obtained inexpensively from YouTube [19]. Their approach is robust to the low-quality video as well as complex scenes necessary for such video material.

All the above approaches are based on low-level features either describing the video clips or the human track. More sophisticated approaches model instead human-human or human-object interactions. In [5, 26] interactions between humans are recognized in a hierarchical manner using higher level features such as body parts. These approaches operate on constrained data that permits preprocessing steps such as background subtraction and body part segmentation. In contrast, a recent work [27] represents human-human interactions based on a robust human-centred descriptor, which uses head orientation and captures hand and arm movements through spatial and temporal features collected in the neighborhood of each detected person.

Most closely related to our approach is probably the work of Gupta et al. [17]. They model the action object and the human-object motion for classifying interactions between humans and objects. However, the motion features used in their approach are more fragile: they rely on hand trajectories to model how objects are reached and grasped. In particular, the velocity profile of the reaching hand and the time interval between a reach and a grasp motion proved to be powerful features in their experiments. Nevertheless, these fine-grained features rely on motion extracted based on background subtraction, which limits its applicability to static cameras and backgrounds (as opposed to uncontrolled video such as feature films). Moreover, [17] requires substantial annotation effort for training, including the location of the person, of its hands, and a pixelwise segmentation of the action object in all video frames.

Our work is also related to methods for modeling human-object interactions in static images [28, 38, 39]. However, these approaches do not take advantage of motion characteristics of actions. Furthermore, [38] operates in a constrained setup where human location is given even at test time, and [39] expects the full human body pose to be always visible.

3 Overview of our method

In this section we present an overview of our approach to action recognition, based on explicitly modeling the human-object interaction (fig. 2). We summa-

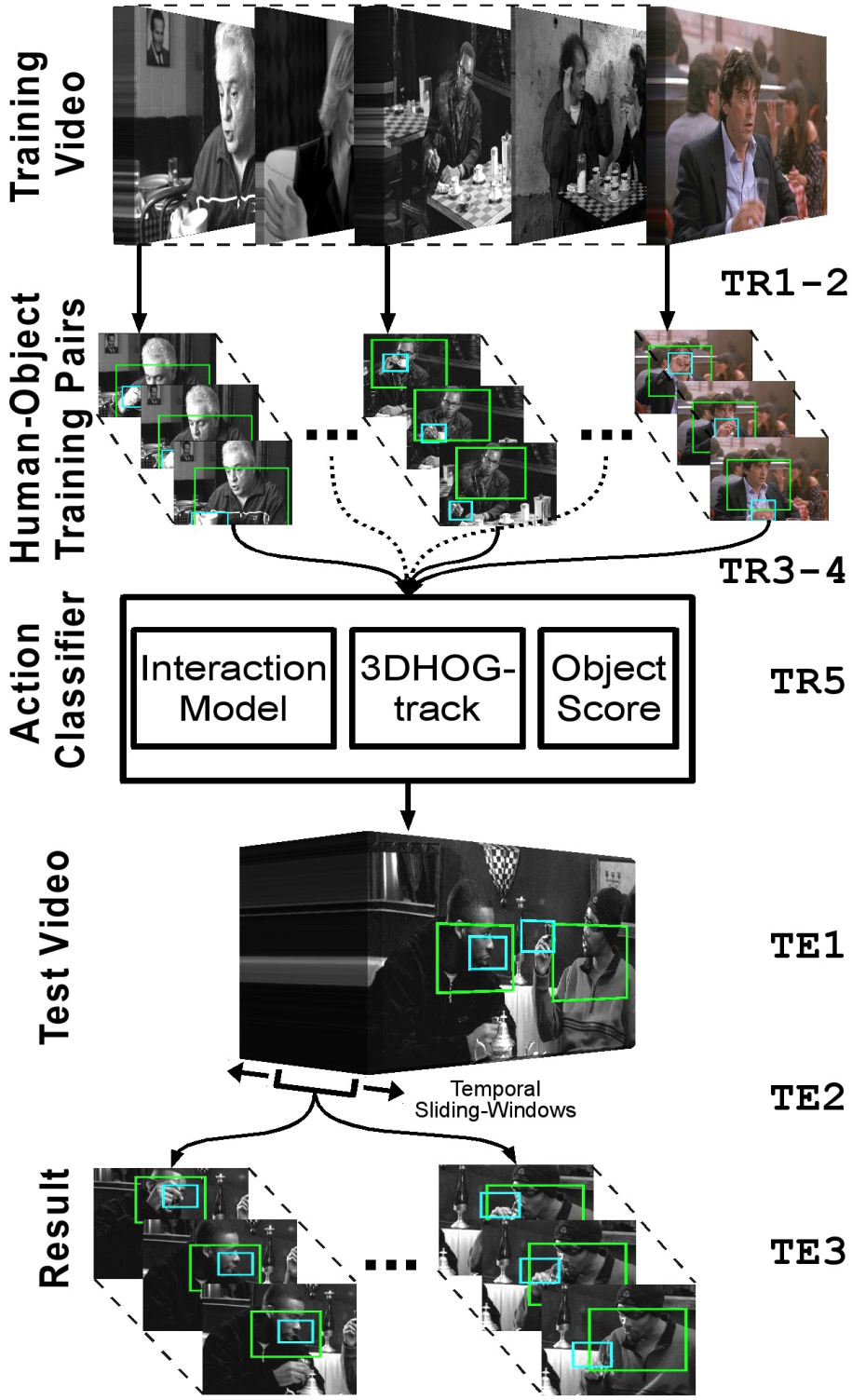


Figure 2: **Overview of our method.** *We show the training pipeline (*TR1 − 5*) and the test pipeline (*TE1 − 3*). See text for details.*

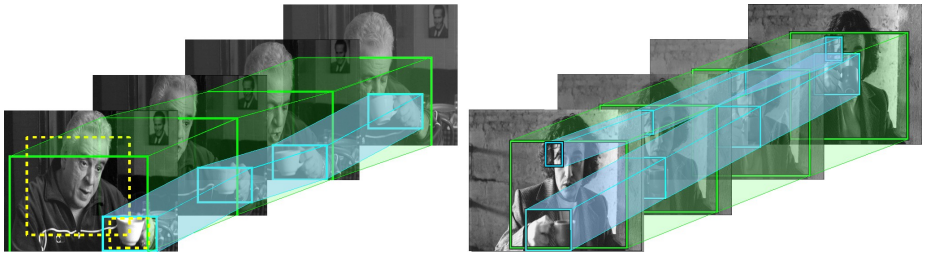


Figure 3: **Tracking at training and test time.** *(Left) The training stage* TR1 *tracks the annotated bounding-boxes (dashed yellow) throughout the temporal extent of the annotation cuboid (persons in green and objects in cyan). (Right) The test stage* TE1 *detects both humans and objects automatically and tracks them throughout the video. For illustration we show here only two object tracks, out of many more (a positive one covering the cup, and a negative one on the actor's face).*

rize the stages of the pipelines for training the model for an action class (sec. 3.1) and for localizing it in space and time in a novel test video (sec. 3.2).

3.1 Training

Input. In order to train the model for an action class, our method takes as input: (i) a long video including instances of the action class; (ii) spatio-temporal cuboids, constant in the spatial dimension. Each annotation cuboid defines the location in time and space of a human performing an instance of the action class; (iii) for each annotation cuboid, the location of the action-object is annotated in *one frame* within the temporal extent of the cuboid. In the following we describe each step of the training (TR) procedure, marked as TR1 – 5.

TR1. We localize and track the humans in the training video. We first apply the human detector of [28] independently on each frame and then link the resulting detections over time into tracks (sec. 4). For each annotation cuboid, we select the track which best overlaps with it and cut it to the precise temporal extent. This results in our set of *positive human tracks*. There is exactly one such track for each cuboid.

As the overall goal of our work is to learn the relative motion between humans and objects that is characteristic for the action class, we also need to track action-objects. For each annotation cuboid, we track the object starting from the single annotated frame forward and backward in time until either end of the temporal extent of the cuboid (sec. 4). These form the *positive object tracks* (fig. 3, left). Again, there is exactly one such track for each cuboid. For each cuboid we now associate its human and object track into a *positive human-object pair*.

TR2. We use the object windows in all frames of all positive object tracks as positive samples for training an action-object detector using the recent method of [11] (sec. 4.1).

TR3. We use the detector from TR2 on the negative parts of the training video (i.e. parts not overlapping in time with any cuboid), and then run our tracker to link the resulting detections over time, obtaining *negative object tracks*. These

Figure 4: **The LDOF-MS tracker at test time**. *In the first frame (left) the cup is automatically detected by an object detector. In the subsequent frames no detections are found on the cup. LDOF-MS produces an object track (dashed window) by propagating the detection from the first frame according to point tracks (white segments). In the last frame (right) the cup is again detected. LDOF-MS adds this detection to the track and uses it to update the confidence score, but not the location of the track.*

are valuable 'hard negatives'. We now form *negative human-object pairs* by associating negative human and object close in space and time (sec. 5.2).
TR4. For each human-object pair we compute an interaction descriptor capturing the relative location, relative area and relative motion of the object wrt the human (sec. 5.1). Moreover, we also compute the low-level 3DHOG-track descriptors [20] for each human track in a pair. As a third descriptor, we use the score of the object detector trained in TR2 on the object track in a pair (sec. 6).
TR5. We use the descriptors from positive and negative human-object pairs to train a discriminative action classifier (sec. 6).

3.2 Testing

Input. Given an input test video we localize the action class in space and time. Note the complexity of the task: we localize a short action in a full length movie.
TE1. We compute human tracks on the test video with the same technique as in TR1 (sec. 4). However, as we now have no cuboid annotations, we retain all human tracks for the later stages. We also compute candidate object tracks by first running the single-frame action-object detector learned in TR2, and then running our tracker to link the resulting detections over time (fig. 3 right). We then associate human and object tracks into human-object pairs (sec. 5.2).
TE2. These raw pairs are unlikely to precisely cover the temporal extent of the action. In order to obtain an appropriate temporal extent of the action, we use a multi-scale temporal sliding window to produce multiple *candidates* with different temporal extents for each test pair (sec. 5.3).
TE3. For each candidate pair, we compute the three descriptors as in TR4 and score it with the action classifier trained in TR5. As a last step, we suppress multiple detections of the same action instance: we remove any candidate with significant overlap in space and time with a higher-scored candidate.

4 Tracking humans and objects

Our approach for modeling human-object interactions depends on the availability of human and object tracks in the same time period. For robustness, it

is important to ensure the highest possible recall for both human and object tracks, as missing either of the two prevents the system from recognizing the action.

It is an elusive goal to design robust detectors and trackers to deal with difficult, small objects such as cigarettes or cups. Instead, we propose a tracking-by-detection approach that can be run on top of weak single-frame detectors, and produces a large number of candidate tracks in order to miss as few positive tracks as possible, see sec. 7.1.1 for an experimental evaluation. Then, in sec. 5 we introduce a highly discriminative descriptor that allows to mine for relevant human-object track pairs out of this pool of candidates.

4.1 Detection

Humans. Detecting humans in the C&C dataset is particularly hard due to their variety of appearance, pose, viewpoints and lighting conditions. The previous work of Klaser et al. [20] used a human detector based on HOG features [4] trained on C&C to learn the specific features of this dataset.

We take a more general approach by employing the generic part-based human detector presented in [28]. This detector combines four part detectors dedicated to different regions of the human body (including full-body, upper-body, and face). It was trained from external still images without using any C&C images [28, sec. 2]. Two of the four components of this combined detector are taken from the popular person detector of [11].

Objects. Detecting small objects such as cups and cigarettes is an even harder task than detecting humans. In addition to being small, these objects present a high degree of pose and appearance variability. For this task we rely on the detection approach of Felzenswalb et al. [11], which demonstrated state-of-the-art results on the PASCAL VOC object detection challenge [9]. We use the windows from the positive object tracks obtained in TR2 as positive training data. As negative training data we randomly sample windows from Caltech-101 [12].

4.2 Tracking

Tracking is needed at various stages of our approach. During training we need to track each action object starting from the initialization in a single annotated frame (TR1). This is a traditional bottom-up tracking task [37, 36, 16]. Furthermore, during TR1 we need to link over time human detections obtained automatically in individual frames. This is instead a tracking-by-detection task [2, 13, 20]. During testing (TE1) tracking-by-detection is needed again for both humans and objects (as at this point we have an object detector from TR2).

Previous works [2, 13, 20, 29] have been successful in tracking people in realistic videos by linking the output of a person detector run independently on each frame (tracking-by-detection). However, tracking small objects such as cups or cigarettes in this manner is much harder because detectors tend to miss the object in many frames. As a consequence, the object motion is typically broken into many short tracks. As another problem, tracking-by-detection does not work when we do not have a detector yet, i.e. when the object to be tracked

is given only as a bounding-box in a single frame. This is instead the realm of application of bottom-up trackers [37, 36, 16].

We propose here a general-purpose tracking method to robustly track multiple targets in an integrated manner that encompasses both the bottom-up and the tracking-by-detection scenarios. Inspired by [33], our algorithm takes as input any number of detection windows on the targets, and propagates them forward and backward in time based on point-tracks. During this process, multiple windows that spatially meet in a frame are automatically merged in a single output track.

Our tracker, referred to by *large-displacement optical flow [3] – median shift* (LDOF-MS), works as follows:

1. *Input.* A sequence of frames $\{s, \ldots, e\}$ and a set of detections $\mathcal{D}^i$ for each frame $i \in \{s, \ldots, e\}$. At least one detection in one frame is required for the algorithm to run (bottom-up tracking). If more are provided, the algorithm will try to link them over time (tracking-by-detection). Any in-between situation is supported, e.g. where some targets have a single initialization window and others have a sparse set of windows output by a detector. For producing point tracks, we run the large displacement, dense optical flow tracker [3] over the entire sequence.

2. *Initialization.* Let f be the first frame for which a detection is available. For each detection $\mathcal{D}_j^f \in \mathcal{D}^f$ create a new track $\mathcal{T}_j$, and add it to overall track set $\mathcal{T}$.

3. *Forward pass.* Loop over frames i from f to e

   (a) Loop over tracks $\mathcal{T}_j \in \mathcal{T}$

      i. *Update location.* The position of $\mathcal{T}_j^{i+1}$ of track $\mathcal{T}_j$ in frame $i+1$ is the position of $\mathcal{T}_j^i$ shifted by the median displacement between frame i and $i+1$ of the point tracks inside window $\mathcal{T}_j^i$.

      ii. *Include a detection.* If a detection $\mathcal{D}_k^{i+1}$ in frame $i+1$ substantially overlaps with $\mathcal{T}_j^{i+1}$, then it is assigned to $\mathcal{T}_j^{i+1}$. The detection $\mathcal{D}_k^{i+1}$ is then removed from $\mathcal{D}^{i+1}$. This step has no other effect for the moment. The detections assigned to a track will be used in step 6) to compute its confidence score.

   (b) *Add new tracks.* For each detection $\mathcal{D}_k^{i+1}$ that was not included into an existing track in step 3.(a).ii, we start a new track and add it to $\mathcal{T}$.

4. *Backward pass.* Store away the current tracks. Restart the process from step 2, this time over the reversed sequence from f to s.

5. *Concatenate forward-backward tracks.* Assemble the final tracks by concatenating the tracks from forward pass to the (reverse) tracks from backward pass.

6. *Confidence scores.* The confidence of a track is the average over the scores of the windows it contains, where the windows scores are normalized between 0 and 1. Windows which are not supported by any detection (see the two central images in fig. 4) are given a score of 0, thus penalizing the overall average.

An important problem this tracker addresses is that detectors of small objects such as cigarettes and cups tend to produce sparse detections in time. As we observed in *Coffee & Cigarettes*, it is common to have tens of frames without detecting the object. LDOF-MS links detections even in this situation, see figure 4 for an illustration. Moreover, it can be used as a bottom-up tracker by providing a single initialization window in one frame. This is due to the fact that the tracker actively updates the position of a window over time according to the median motion of the point tracks inside it. Therefore, once a track is initialized, it does not require later detections to survive, as opposed to [13, 20, 34]. Finally, note how LDOF-MS tracks any number of detections in parallel without substantial increase in computation time.

Robust point tracks. For obtaining point tracks in step 1, we rely on the recent work on large-displacement optical flow (LDOF) [3]. LDOF is a variational technique that integrates discrete point matches, namely the midpoints of regions, into a continuous energy formulation. The energy is optimized by a coarse-to-fine scheme to estimate large displacements also for small scale structures. As opposed to traditional optical flow, this algorithm tracks points over multiple frames, not only over two.

5 Modeling human-object interactions

In this section we model the interaction between a human track $\mathcal{H}$ and an object track $\mathcal{O}$ in terms of relative position and motion features (stages TR4 and TE3). These features are computed for a human-object track pair, which have been formed before. Positive human-object pairs are formed easily at training time, as there is only one possible pair for each annotation cuboid (TR1). Instead, forming negative training pairs, and all pairs at test time, requires a dedicated procedure which we describe in sec. 5.2. In sec. 5.1 we start by presenting our interaction descriptor, which we compute for any human-object pair.

5.1 Interaction descriptor

In the following we describe the relative location, area and motion of the object track wrt the human track in the time interval $[t_{min}, t_{max}]$ in which they both exist (i.e. the intersection of their temporal extents). Note that both $\mathcal{H}$ and $\mathcal{O}$ have a window $\mathcal{H}^t$ and $\mathcal{O}^t$ in every frame $t \in [t_{min}, t_{max}]$, as our tracker never skips a frame (sec. 4.2).

At every frame t in the interval $[t_{min}, t_{max}]$ we compute three features:

1. *Relative location.* The relative location $\mathrm{l}(\mathcal{H}^t, \mathcal{O}^t)$ of the object window $\mathcal{O}^t$ wrt to the human window $\mathcal{H}^t$ in frame t

$$\mathrm{l}(\mathcal{H}^t, \mathcal{O}^t) = ((\mathcal{O}^t_x - \mathcal{H}^t_x)/\mathcal{H}^t_W \ , \ (\mathcal{O}^t_y - \mathcal{H}^t_y)/\mathcal{H}^t_H) \tag{1}$$

   where subscripts indicate a window's center x, y, width W and height H.

2. *Relative area.* The area of $\mathcal{O}^t$ relative to $\mathcal{H}^t$

$$\mathrm{a}(\mathcal{H}^t, \mathcal{O}^t) = \mathrm{area}(\mathcal{H}^t)/\mathrm{area}(\mathcal{O}^t) \tag{2}$$

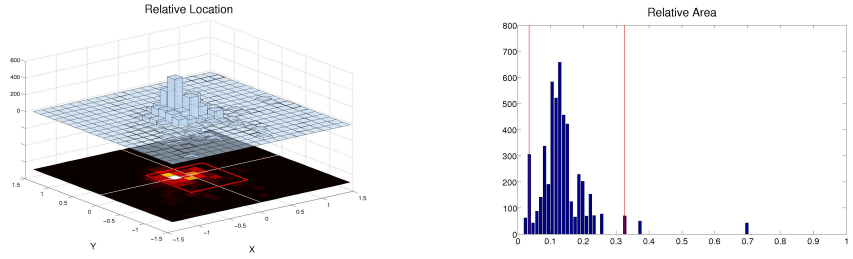


Figure 5: **Learning the interaction ranges.** *Histograms of relative location and relative area accumulated over all positive training human-object pairs. In red are shown the learned ranges. The left plot shows that the location of the cup is typically in the middle of the human window along the horizontal axis and slightly above it along the vertical axis.*

3. *Relative motion.* The relative motion of the object wrt to the human is an important cue for distinguishing actions. We define this as the 2D vector

$$\mathrm{m}(\mathcal{H}^t, \mathcal{O}^t) = \mathrm{l}(\mathcal{H}^t, \mathcal{O}^t) - \mathrm{l}(\mathcal{H}^{t-1}, \mathcal{O}^{t-1}) \tag{3}$$

the difference between the relative location $\mathrm{l}(\mathcal{H}^t, \mathcal{O}^t)$ in frame t and $\mathrm{l}(\mathcal{H}^{t-1}, \mathcal{O}^{t-1})$ in frame $t-1$. We represent this vector by its magnitude and direction.

We compute an interaction feature at every frame of a human-object pair and then aggregate them into a single descriptor of fixed dimensionality as follows. For each feature we accumulate its values over the time interval in a histogram. We independently $L1$-normalize each histogram and then concatenate them to obtain the final interaction descriptor. The 2D *relative location* and *relative motion* cues are quantized into 16-dimensional histograms each and *relative area* is quantized into 4 dimensions. This results in a total of 36 dimensions. Interestingly, we did not observe any improvement by using a higher dimensionality.

5.2 Forming human-object pairs

We describe here how to associate human and object tracks when collecting negative human-object pairs during training (stage TR3) and when forming pairs during testing (stage TE2). A simple approach would be to take all temporally overlapping pairs of human and object tracks. However, this would lead to a huge number of pairs, which would make action detection very slow. Instead, we perform here a preselection stage, where we associate pairs based on two interaction features from sec. 5.1.

Learning interaction ranges. Previous works on human-object interactions [17, 28, 39, 38] have shown the importance of limiting the spatial range of an action-object wrt a human. We learn the interaction range for the *relative location* and *relative area* features. After the training step TR1, we have a set of positive human-object track pairs. For each frame in every human-object pair, we compute the two interaction features, see fig. 5 for their distribution. For each feature, we then select the range of the feature such that 90% of the mass of the distribution is contained in it. Note how this threshold operates at

a frame level thus discarding 10% of the outlying mass of the distribution and preserving relevant geometric information from the remaining frames.
Forming pairs. The ranges learned for the spatial interaction features are used to select spatially consistent pairs from the set of temporally overlapping ones. Fig. 5 illustrates the feature distributions and learned ranges for the drinking action from *Coffee & Cigarettes*.

5.3 Temporal chunking at test time.

In the above pairing scheme, the temporal extent of a test pair is simply the time interval during which both tracks exist. Instead, we would like to focus on the temporal segment where the action takes place. For this reason we introduce a multi-scale temporal sliding-window mechanism for the test human-object pairs. For our experimental results we use three temporal scales, which are learned from the training cuboids. Given the temporal duration of these cuboids, k-means determines three clusters. The durations corresponding to the cluster centers are used as temporal scales. The step size is fixed to 10 frames in all our experiments. The output of this procedure is a large number of overlapping test pairs which are then scored by our action classifier TE3 (sec. 6). As a final step, we apply non-maxima suppression in order to suppress multiple detections of the same action instance: we remove any candidate with significant overlap in space and time with a higher-scored candidate.

6 Action classifier

This section presents how to train the action classifier (stages TR4 and TR5). We train multiple classifiers based on different features capturing complementary aspects of actions. The goal of each classifier is to decide whether a human-object track pair $(\mathcal{H}, \mathcal{O})$ is an instance of the action class. In a final step, we combine the output of all classifiers into a single action classifier. This is used to score candidate track pairs during testing (stage TE3).
Human-object interaction classifier. The training stage TR4 outputs an interaction descriptor (sec. 5.1) for each training $(\mathcal{H}, \mathcal{O})$ pair. We train an SVM classifier with an intersection kernel [23] to separate descriptors from positive and negative pairs.
Action-object classifier. For each training pair $(\mathcal{H}, \mathcal{O})$, we collect the score of the object detector in each frame of the object track $\mathcal{O}$. The maximum value over the track is taken as the output of this classifier. Given that the object might be hard to recognize in many frames due to viewpoint changes and localization inaccuracy, the maximum value gives the track a high score as long as at least one frame has a high score.
3DHOG-track classifier. We compute the 3DHOG-track features [20] on the human track $\mathcal{H}$. This feature extends the HOG image descriptor to videos by extracting 3D HOG descriptors for spatio-temporal subvolumes of the track. It goes beyond a rigid spatio-temporal cuboid [22, 35], as it adjusts piecewise to the spatial extent of the tracks. This introduces a more flexible representation, where the descriptor remains centered on the action. The 3DHOG-track feature is complementary to our human-object interaction descriptor, as it captures low-level appearance and motion information. Experimental results demonstrate

their complementarity (sec. 7.1.2). We train a non-linear SVM classifier with RBF kernel to separate positive and negative training track pairs.

Combined action classifier. We linearly combine the output of the three above classifiers by training a linear SVM on the 3D vector of outputs from positive and negative training pairs. At test time, stage TE3, we use this classifier to score all test pairs (obtained as in sec. 5.2).

7 Experimental results

We present an evaluation of our method on two existing dataset of human-object interactions: *Coffee & Cigarettes* [22] (sec. 7.1) and the Gupta video dataset [17] (sec. 7.2). These datasets are complementary. *Coffee & Cigarettes* focuses on accurate spatio-temporal localization of two actions in a full-length realistic movie. In contrast, the Gupta video dataset has more action classes, but the videos are taken in a controlled laboratory environment and each video clip contains only a single action. Furthermore, the task is multi-class classification of the clips, i.e., the actions are approximatively localized and localization in space and time are not evaluated. We also investigate the performance of human and object tracks and human-object track pairs on *Coffee & Cigarettes* (sec. 7.1.1).

7.1 Evaluation on Coffee & Cigarettes

The film *Coffee & Cigarettes* consists of 11 short stories, each with different scenes and actors. The C&C dataset [20, 22] focuses on the actions drinking and smoking.

For drinking, the training set contains 41 video clips from 6 short stories. Additionally, it contains 32 samples from the movie *Sea of Love* and 33 samples recorded in a lab. This results in a total of 106 positive drinking samples for training. We collect 50000 negative samples (human-object pairs) from the 6 training short stories by selecting sequences which do not overlap with any of the positive samples.

For testing, instances of the drinking action are localized in 2 short stories not used for training, i.e., in 24 minutes of video, which contain 38 drinking samples corresponding to a total of 1.8 minutes.

The smoking training set contains 78 samples: 70 samples from 6 short stories of C&C (the ones used for training the drinking action) and 8 from *Sea of Love*. Analogously to the drinking action, we use 50000 human-object pairs from the 6 short stories of C&C not overlapping with any annotation as negative training samples. For testing, instances of the smoking action are localized in 3 short stories not used for training, i.e., in 21 minutes of video, which contain 42 smoking samples corresponding to a total of 2.3 minutes. Note the difficulty of spatio-temporal detection of such short actions in realistic full-length videos.

The training annotations [22] come in the form of cuboids $\mathcal{A}$ which define the location in time and space of humans performing the action. For each training cuboid we complement these original annotations with a bounding-box delimiting the action-object in *one frame*.

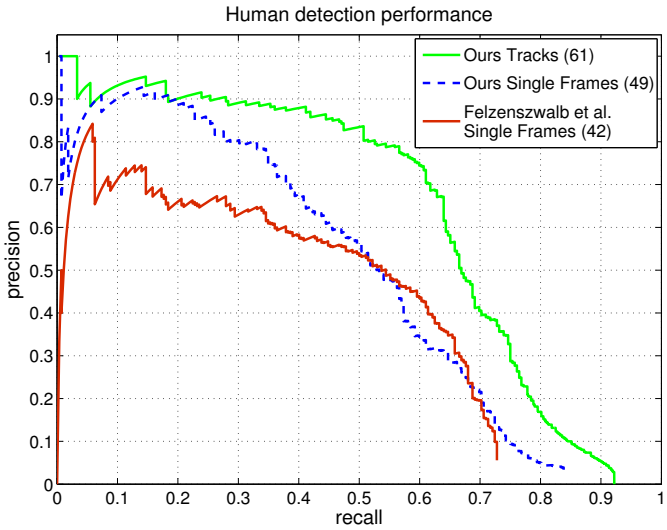


Figure 6: **Human detection performance**. *See text for details.*

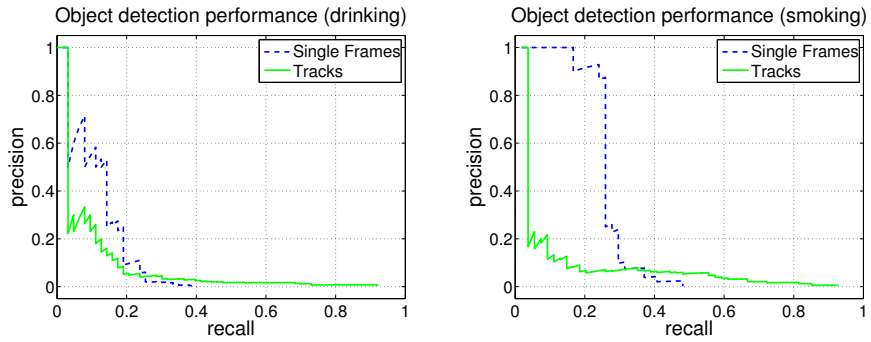


Figure 7: **Object detection performance**. *See text for details.*

7.1.1 Evaluating human and object tracks

Humans. In order to compare the human detection and tracking performance of our method with the one from Klaeser et al. [20] we evaluate on their dataset. This dataset is composed of 137 frames of C&C [22], for which a total of 260 ground-truth bounding-boxes are available. These frames are extracted from sequences of the movie that are not part neither of the training nor the test set. Unlike the original C&C annotations that provide the location of humans performing the action, this dataset contains the location of *every* human in an image. A person is considered to be correctly localized when the predicted and ground-truth bounding-boxes overlap more than the PASCAL VOC criterion (i.e. Intersection-over-Union above 50%). Performance is summarized by average precision (AP).

Fig. 6 compares three methods. The two *Single frames* methods run a human detector on each test image independently: (i) the popular human detector

	Drinking	**Smoking**
$\|\mathcal{H}\|$	8924 (94%)	12558 (93%)
$\|\mathcal{O}\|$	49319 (92%)	71737 (93%)
$\|(\mathcal{H},\mathcal{O})\|$	418980 (90%)	1619284 (89%)

Table 1: *Number of tracks and recall (in parentheses) for humans $\mathcal{H}$, objects $\mathcal{O}$ and human-object pairs $(\mathcal{H},\mathcal{O})$ on the Coffee & Cigarettes test set.*

of [11], trained on the PASCAL 2007 VOC training set [9]; and (ii) our detector [28], which complements [11] with additional detectors specialized for the face and upper-body regions (sec. 4.1). We can observe that the combination of different human part detectors [28] is beneficial on this difficult *Coffee & Cigarettes* dataset, improving over [11] by 7% AP.

The *Tracks* method links the detections output by our human detector using the tracker presented in sec. 4.2. For this evaluation, detections are first computed on each frame in a short temporal interval around a test image, and then linked using the tracker. However, evaluation is only done on the 137 test frames, as for the *Single frames* methods. The associated score is the one of the track, i.e., the average detection score over the track.

Our *Tracks* method outperforms substantially our single-frame method, confirming the contribution of our LDOF-MS tracker (+12% AP). Moreover, it also achieves 9% higher AP compared to the human tracker of [20]. This is remarkable, as [20] was trained specifically on C&C, while our detector is trained using only external material ([28, sec. 2]).

Objects. We evaluate object detection performance on frames selected from the test part of C&C. We sample either one or two frames from every positive sample depending on its temporal length. This results in 54 frames for drinking and 47 for smoking. We also evaluate on negative images (i.e. not containing the object): for each class we select a number of negative images that reflects the proportion between positive and negative frames in the test set. This results in 500 negative images for drinking and 349 for smoking. As discussed in sec. 4.1, we train the object detection model of [11] from all windows in the positive object tracks automatically obtained and negative images from Caltech-101. The only manual annotation used was a bounding-box in *one frame* of each action instance.

Fig. 7 compares the performance of the object detectors in the *Single Frames* and the *Tracks* modes. The *Tracks* mode, although introducing some additional false-positives, doubles the maximum recall compared to the *Single Frames* mode, and detects more than 90% of all object instances. This fits the goal stated at the beginning of sec. 4: to produce a pool of candidate tracks which misses as few true object instances as possible. The lower performance of the *Tracks* method in terms of Average Precision is inherent to the context we operate in: we deal with detections which are sparse in time (typically less than 30% of a positive track's frames are supported by a detection) and every frame where a detection is missing penalizes the overall score of the track. As a result, the average track score loses significance. The track score could certainly be made more robust to outliers, but this was not necessary in our context, which requires maximum recall. Note also that this is not a problem when a reliable detector is available, as is the case for human detection (fig. 6).

Human-object pairs. In order to localize an action with our human-object interaction model, the human as well as the object track need to be present. To miss as few as possible human-object pairs performing the action, we keep all human and object tracks, i.e., we operate at the maximum recall level, see the right-most datapoints in fig. 6 and 7. The corresponding numbers are reported in tab. 1. Note that the number of tracks and the recall are reported for the final test datasets, i.e., the two and three short stories used to evaluate drinking and smoking localization.[1]

Given this set of human and object tracks, we form human-object pairs based on the approach described in sec. 5.2, i.e., use preselection based on relative location and area. This results in 418980 track pairs and a recall of 90% for drinking and 1619284 track pairs and a recall of 89% for smoking, see last row of table 1. This shows that the recall is sufficiently high, to make localization of most action instances possible.

This is not the case, if we keep only the 50% highest scoring human and object tracks, i.e., the overall number of human-object pairs is reduced by approximatively a factor four and recall drops to 43% respectively 39% for drinking and smoking.

In the next section we will show that our interaction descriptor is sufficiently powerful to discard the large number of track pairs which do not contain the action.

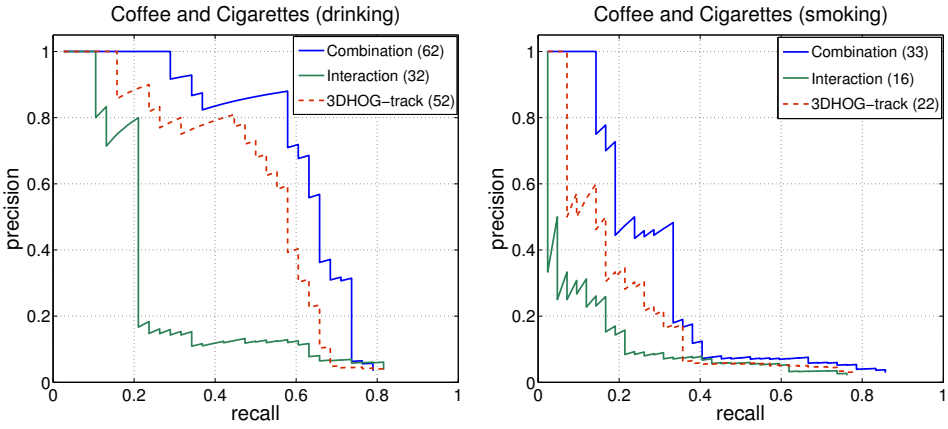


Figure 8: **Precision-recall curves for C&C.** *Performance for spatio-temporal localization of the actions drinking (left) and smoking (right). For each method we present its average precision (AP) in parenthesis.*

7.1.2 Evaluating action detection (localization in space and time)

We now evaluate the performance of our approach for spatio-temporal localization of the actions drinking and smoking on the *Coffee & Cigarettes* dataset and compare to the state of the art. We adopt the evaluation protocol of [22]: an action is correctly detected if the predicted spatio-temporal detection overlaps

[1] This explains the difference in recall for human tracks wrt figure 6, where the evaluation is performed on a different subset of C&C.

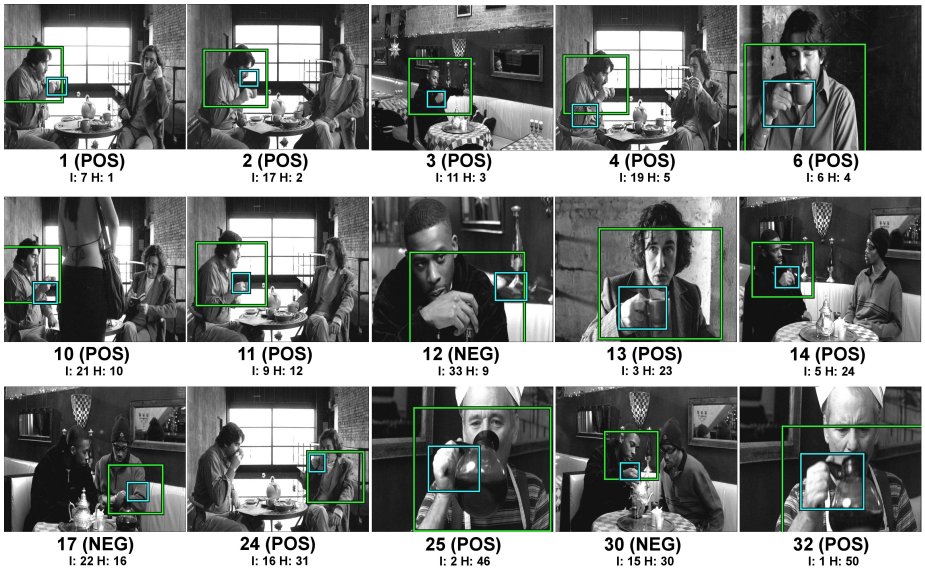


Figure 9: **Drinking results.** *Human-object pairs localized in test videos. The ordering corresponds to the ranking of the combined action classifier. We also show the rank of the individual classifiers separately (I: interaction classifier, H: 3DHOG-track classifier). These results show that the interaction and 3DHOG-track classifiers complement each other. Samples 13 and 14 have a relatively low 3DHOG-track score, whereas the interaction classifier successfully captures the discriminative motion of the object track. In contrast, for samples 2 and 4 the object track is incorrect, resulting in a lower interaction score rank, whereas the 3DHOG-track classifier correctly scores these samples highly. It is interesting, how our method finds object tracks also on unconventional objects such as the jug in samples 25 and 32, which receive top scores by the interaction classifier. For these examples 3DHOG-track fails due to the unusual object appearance. This confirms the ability of the interaction classifier to generalize the appearance of objects and describe their relative motion wrt to the human. Failure cases of the interaction classifier are often due to other objects moving in a similar way as action objects. For example in sample 30 the actor is pouring water from a teapot, resulting in a trajectory similar to the drinking action. For the 3DHOG-track classifier, a typical failure case is when low-level features perform poorly, as is the case in scenes with difficult lighting conditions, as in sample 14, or when the object has an unusual appearance, as in 25 and 32. Other failures by 3DHOG-track (classifying a negative as positive) are due to the actor being in a pose similar to the action, but not performing it, as in sample 12. Note that the interaction classifier receives a relatively low score as there is no motion.*

Figure 10: **Smoking results.** *Human-object pairs localized in test videos. The ordering corresponds to the ranking of the combined action classifier. We also show the rank of the individual classifiers (I: interaction classifier, H: 3DHOG-track classifier). In many cases the interaction and 3DHOG-track classifiers agree and assign both a high score to a positive sample. Complementary scores are obtained for samples 16, 23, 30 and 49: the interaction classifier correctly penalizes these negative samples without correct object motion, whereas 3DHOG-track is unable to distinguish them and assigns high scores. Note that sample 49 is a true negative, as it represents a person holding a cigarette and not smoking. For sample 19 the object track does not cover the correct object, thus the interaction classifier gives a low score, whereas the 3DHOG-track classifier assigns a high score. (*) For the smoking action we point out that the imprecise temporal extent of the annotations sometimes leads to an incorrect evaluation: samples 7 and 8 show a person smoking, but do not meet the spatio-temporal overlap threshold with the annotations.*

	Drinking	Smoking
Interaction classifier	**31.60**	**16.20**
Object classifier	4.30	5.50
3DHOG-track classifier	52.20	21.50
Combination	**62.10**	**32.80**
Laptev et al. [22]	43.40	-
Willems et al. [35]	45.20	-
Klaeser et al. [20]	54.10	24.50

Table 2: **Average precision for spatio-temporal localization on C&C.** *First three rows: our individual classifiers. Fourth row: our full method combing the three classifiers. Last three rows: competing methods ([22, 35] do not report AP for smoking).*

at least 20% with the ground-truth cuboid. The overlap between a ground-truth annotation cuboid $\mathcal{A}$ and a human-object pair $(\mathcal{H}, \mathcal{O})$ is given by $(\mathcal{A} \cap \mathcal{H})/(\mathcal{A} \cup \mathcal{H})$ (i.e. for evaluating our method we use the human track within a pair, as this corresponds to the standard protocol).

Fig. 8 shows precision-recall curves for drinking and smoking actions obtained with our combined method ('Combination') and the individual classifiers. Table 2 reports the average precision (AP) and compares to the state of the art [22, 35, 20]. The classifier based on the score of the object detector (second row) performs very poorly, which confirms that a human-object interaction cannot be defined purely based on the appearance of the object involved. The human-object interaction model we propose achieves good performance already when used on its own (first row). This shows how the relative location and motion of the object wrt the human is a distinctive feature characterizing the human-object interaction. More importantly, combining it with the low-level 3DHOG-track descriptor [2] improves on both and leads to a significant improvement over the state-of-the-art [20] (+8% AP). This demonstrate that our interaction model is complementary to traditional low-level descriptors. Fig. 9 and 10 show some of the top-scored human-object pairs according to the combined action classifier.

7.2 Multi-class classification on Gupta video dataset

The Gupta video dataset [17] contains 60 video clips with 10 actors performing 6 different actions, i.e. drinking from a cup, spraying from a bottle, answering a phone call, making a phone call, pouring from a cup and lighting a flashlight. For each action, the videos are split into 5 training and 5 test videos. Unlike the C&C dataset, these videos are shot in controlled conditions inside a laboratory with a static camera and a static background of uniform color. Furthermore, the video clips are restricted to the temporal extent of the action. Fig. 11 shows frames extracted from the Gupta video test set. Since the annotations used in [17] are not available online, we have re-annotated the dataset to the same level as in 7.1: for each video one cuboid on the human performing the action and a bounding-box delimiting the object in *one frame*.

We train an action classifier for each of the six actions using as negative examples the training videos from the other classes. If two actions share the same object we merge the object tracks from the training videos and learn a single detector in step TR2 (this happens for cup and phone). Given a test video, we evaluate the action classifier score for each of the six actions and return as class label the one with the highest score. Note that the sliding window mechanism of sec. 5.3 is not required, as the video clips are already temporally segmented to the extent of the action. For evaluation we measure the percentage of test videos for which the algorithm predicts the correct label, as in [17].

Table 3 shows the multi-class action classification results. Remarkably, the proposed interaction model already achieves 80% accuracy on its own and outperforms the 3DHOG-track. This demonstrates how our explicit modeling of

[2] Our reimplementation of the 3DHOG-Track classifier achieves a slightly lower performance than the one reported by [20] (52/22 vs 54/25). This might be because [20] uses a finer temporal sliding window for the test tracks (7 scales vs our 3).

	Gupta video
Interaction classifier	**80.00**
Object classifier	36.60
3DHOG-track classifier	63.30
Combination	**93.30**
Gupta et al. [17]	93.00

Table 3: **Average classification accuracy on the Gupta video dataset.**

Figure 11: **Human-object pairs localized on the Gupta video test set with the combined classifier.** *Each row shows one frame of the five test sequences of a class. Actions in this dataset follow precise motion patterns: each row displays samples selected to follow the temporal pattern. The two only misclassified samples are indicated with the incorrect class label overlaid.*

the object motion trajectory is a strong cue for action classification. The interaction model performs better on this dataset than on C&C, because the objects are easier to track in these simpler imaging conditions.

The performance obtained with our combined action classifier is on par with the result from [17]. Note that [17] explicitly takes advantage of the static camera and background used in these videos, rendering it unsuitable for more

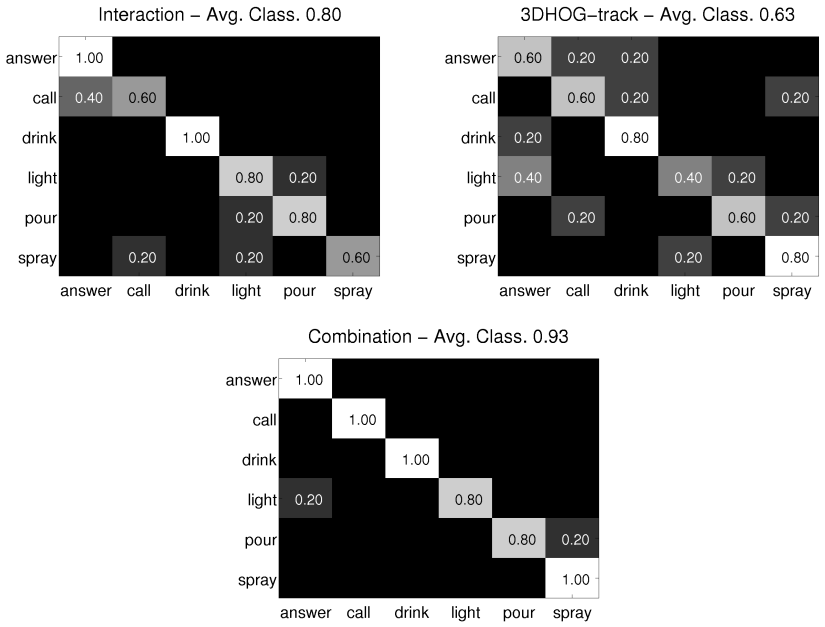


Figure 12: **Confusion matrices on the Gupta video dataset.** *(Top-left) performance of the interaction classifier; (top-right) 3DHOG-track classifier; (bottom) combined action classifier.*

complex videos such as C&C. Moreover, our method needs substantially less manual annotation for training than [17], which requires the location of the person's hand and a pixelwise segmentation of the object in every frame of all training videos.

Figure 12 presents the confusion matrices, showing that most errors made by the interaction classifier are due to the similarity of the action 'lighting torch' with 'pouring water' and 'spraying'. These were distinguished in [17] based on the color of the action-object, a feature which is not used here. Misclassifications between 'answering' and 'calling' are due to their similar motion. In the case of the combined classifier, there are only two misclassified samples, i.e, "light" is misclassified as "answer" and "pour" as "spray", see figure 11. These could probably be removed if colour information was used.

8 Conclusion

This paper introduces an approach for learning human-object interactions in videos. It explicitly tracks both the human and the action-object and represents the interaction as the relative position and motion of the object wrt the human. Experimental results confirm that human-object interactions, when explicitly captured by our method, are a rich source of information for action recognition and localization in video. Furthermore, we show that the proposed interaction model captures information complementary to existing low-level descriptors. Moreover, when combining the two, our approach improves over the state of the

art [20] on *Coffee & Cigarettes* and achieves the same results of [17] on Gupta video, despite using substantially less supervision for training.

References

[1] A.F. Bobick and J.W. Davis, *The recognition of human movement using temporal templates*, PAMI (2001).

[2] M. Breitenstein, F. Reichlin, and L. Van Gool, *Robust tracking-by-detection using a detector confidence particle filter*, ICCV, 2009.

[3] T. Brox and J. Malik, *Large displacement optical flow: Descriptor matching in variational motion estimation*, PAMI (2011).

[4] N. Dalal and B Triggs, *Histogram of oriented gradients for human detection*, CVPR, 2005.

[5] Ankur Datta, Mubarak Shah, Niels Da, and Niels Da Vitoria Lobo, *Person-on-person violence detection in video data*, ICPR, 2002.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, *Behavior recognition via sparse spatio-temporal features*, VS-PETS, 2005.

[7] O Duchenne, I Laptev, J Sivic, F Bach, and J Ponce, *Automatic annotation of human actions in video*, ICCV, 2009.

[8] Alexei Efros, Alexander Berg, Greg Mori, and Jitendra Malik, *Recognizing action at a distance*, ICCV, 2003.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[10] A. Fathi and G. Mori, *Action recognition by learning mid-level motion features*, CVPR, 2008.

[11] Pedro F. Felzenszwalb, Ross B. Girshick, D. McAllester, and D. Ramanan, *Object detection with discriminatively trained part based models*, PAMI (2009).

[12] R. Fergus and P. Perona, *Caltech object category datasets*, http://www.vision.caltech.edu/html-files/archive.html, 2003.

[13] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, *Progressive search space reduction for human pose estimation*, CVPR, 2008.

[14] A. Gaidon, Z. Harchaoui, and C. Schmid, *Actom sequence models for efficient action detection*, CVPR, 2011.

[15] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, *Actions as space-time shapes*, PAMI (2007).

[16] Helmut Grabner, Christian Leistner, and Horst Bischof, *Semi-supervised on-line boosting for robust tracking*, ECCV, 2008.

[17] A. Gupta, A. Kembhavi, and L.S. Davis, *Observing human-object interactions: Using spatial and functional compatibility for recognition*, PAMI (2009).

[18] Nazli Ikizler and David A. Forsyth, *Searching for complex human activities with no visual examples*, IJCV (2008).

[19] N. Ikizler-Cinbis, G. Cinbis, and S. Sclaroff, *Learning actions from the web*, ICCV, 2009.

[20] Alexander Kläser, Marcin Marszałek, Cordelia Schmid, and Andrew Zisserman, *Human focused action localization in video*, International Workshop on Sign, Gesture, and Activity (SGA) in conjunction with ECCV, 2010.

[21] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, *Learning realistic human actions from movies*, CVPR, 2008.

[22] I. Laptev and P. Perez, *Retrieving actions in movies*, ICCV, 2007.

[23] S. Maji, A.C. Berg, and J. Malik, *Classification using intersection kernel support vector machines is efficient*, CVPR, 2008.

[24] K. Mikolajczyk and H. Uemura, *Action recognition with motion-appearance vocabulary forest*, CVPR, 2008.

[25] Juan Carlos Niebles, Chih-Wei Chen, , and Li Fei-Fei, *Modeling temporal structure of decomposable motion segments for activity classification*, ECCV, 2010.

[26] Sangho Park and J. K. Aggarwal, *Simultaneous tracking of multiple body parts of interacting persons*, CVIU (2006).

[27] Alonso Patron, Marcin Marszalek, Andrew Zisserman, and Ian Reid, *High five: Recognising human interactions in TV shows*, BMVC, 2010.

[28] Alessandro Prest, Cordelia Schmid, and Vittorio Ferrari, *Weakly supervised learning of interactions between humans and objects*, PAMI (to appear).

[29] D. Ramanan, D. A. Forsyth, and A. Zisserman, *Tracking people by learning their appearance*, PAMI (2007).

[30] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah, *Action mach: a spatio-temporal maximum average correlation height filter for action recognition*, CVPR, 2008.

[31] Scott Satkin and Martial Hebert, *Modeling the temporal extent of actions*, ECCV, 2010.

[32] C. Schuldt, I. Laptev, and B. Caputo, *Recognizing human actions: A local SVM approach*, ICPR, 2004.

[33] J. Sivic, M. Everingham, and A. Zisserman, *Person spotting: video shot retrieval for face sets*, CIVR, 2005.

[34] Josef Sivic, Mark Everingham, and Andrew Zisserman, *'who are you? – learning person specific classifiers from video*, CVPR, 2009.

[35] G. Willems, J. H. Becker, T. Tuytelaars, and L. van Gool, *Exemplar-based action recognition in video*, BMVC, 2009.

[36] Zheng Wu, Margrit Betke, Jingbin Wang, and Vassilis Athitsos, *Tracking with dynamic hidden-state shape models*, ECCV, 2008.

[37] Changjiang Yang, Ramani Duraiswami, and Larry Davis, *Efficient mean-shift tracking via a new similarity measure*, CVPR, 2005.

[38] B. Yao and L. Fei-Fei, *Grouplet: A structured image representation for recognizing human and object interactions*, CVPR, 2010.

[39] ———, *Modeling mutual context of object and human pose in human-object interaction activities*, CVPR, 2010.

[40] Alper Yilmaz and Mubarak Shah, *Actions sketch: a novel action representation*, CVPR, 2005.

[41] L. Zelnik-Manor and M. Irani, *Event-based analysis of video*, CVPR, 2001.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-0803