Human Focused Action Localization in Video

Alexander Kläser[1], Marcin Marszałek[2],
Cordelia Schmid[1], and Andrew Zisserman[2]

[1] INRIA Grenoble, LEAR, LJK
{klaser,schmid}@inrialpes.fr
[2] Engineering Science, University of Oxford, UK
{marcin,az}@robots.ox.ac.uk

Abstract. We propose a novel *human-centric* approach to *detect and localize* human actions in *challenging* video data, such as Hollywood movies. Our goal is to localize actions in time through the video and spatially in each frame. We achieve this by first obtaining generic spatio-temporal human tracks and then detecting specific actions within these using a sliding window classifier.

We make the following contributions: (i) We show that splitting the action localization task into spatial and temporal search leads to an efficient localization algorithm where generic human tracks can be reused to recognize multiple human actions; (ii) We develop a human detector and tracker which is able to cope with a wide range of postures, articulations, motions and camera viewpoints. The tracker includes detection interpolation and a principled classification stage to suppress false positive tracks; (iii) We propose a track-aligned 3D-HOG action representation, investigate its parameters, and show that action localization benefits from using tracks; and (iv) We introduce a new action localization dataset based on Hollywood movies.

Results are presented on a number of *real-world* movies with crowded, dynamic environment, partial occlusion and cluttered background. On the Coffee&Cigarettes dataset we significantly improve over the state of the art. Furthermore, we obtain excellent results on the new *Hollywood–Localization* dataset.

Keywords: Action recognition, localization, human tracking, HOG.

1 Introduction

Our objective is to *localize* human actions both in space (the 2D image region) and time (the temporal window) in videos. Early work on action recognition in video used sequences with prominent actions, mainly static cameras, simple backgrounds and full bodies visible, as in the KTH [1] and Weizmann [2] datasets, e.g. [3,4,5]. This enabled action classifiers to be explored with variation in the actors and actions, but without the added complexity of change of viewpoint, scale, lighting, partial occlusion, complex background etc. However, following recent work [6,7,8,9] where video material from movies is used, the field has

K.N. Kutulakos (Ed.): ECCV 2010 Workshops, Part I, LNCS 6553, pp. 219–233, 2012.
© Springer-Verlag Berlin Heidelberg 2012

moved onto less controlled and much more challenging datasets. Our work is aimed at this more challenging movie material.

As is well known from the results of the PASCAL Visual Object Classes challenges [10], *localization* is much more demanding than *classification*. In the case of action classification, which is often the aim in action research, sequences with pre-defined temporal extent are labeled as belonging to one of n action classes – experiments on the KTH, Weizmann and Hollywood2 [11] datasets generally report such classifications rather than localization. However, for video search and annotation applications, action localization is far more useful than classification as it enables the temporal sequence containing the action to be delimited and also the actor carrying out the action to be identified, when there are multiple actors in a shot.

We propose an approach which explicitly splits the spatio-temporal action localization into first detecting and tracking humans, which determines the spatial localization of the action, followed by a temporal action classification of the tracks, which detects and localizes the action in time. To this end we make contributions in two areas: (i) a generic human tracking method for uncontrolled videos, which outputs high quality tracks by adding a classification stage to suppress false positives; and (ii) a track adapted 3D (space and time) descriptor, inspired by the HOG descriptor [12], which enables a temporal sliding window classifier to reliably recognize and localize actions.

We show that using human tracks gives benefits on three fronts: first, the localization performance improves over the state of the art; second, the complexity of search is reduced (since search restricted to a track is less costly than an exhaustive search of the complete spatio-temporal volume); and, third, learning new actions is far more efficient – since the tracks are agnostic about the actions, they can be reused for any action, and training the classifier for new actions is cheap.

While the idea of combining tracking and classification for action localization is not new, previously it has mainly been applied to video restricted to a static camera [13,14] or simple background with limited clutter, as for example football or ice hockey fields [15,16]. In such a context methods such as background subtraction or image differencing can localize the actors. Furthermore, recognition tasks often focus on periodic and continuous actions (e.g., handwaving or running) or only perform temporal, but not spatial, localization [17].

A few recent approaches address the problem of localizing natural actions in realistic cluttered videos. Laptev and Pérez [7] localize actions by training an action-pose specific human detector (e.g. for the moment of drinking) in combination with a spatio-temporal video block classifier. Willems *et al.* [18] also improve efficiency of detection, in their case by using visual words that are discriminative for an action to propose spatio-temporal blocks for subsequent classification. Ke et al. [19] matches spatio-temporal voxels to manually created shape templates. As will be seen our method substantially outperforms [7,18].

The datasets and evaluation method used throughout the paper are described in section 2. In particular we introduce a new dataset for training and testing

action localization – *Hollywood–Localization*. Section 3 describes the tracking-by-detection method we use to obtain human tracks. Given the track, we determine *if* the action occurs and *when* (temporal localization) using a sliding window classifier based on the new spatio-temporal track-adapted 3D-HOG descriptor (section 4). In section 5 we compare our approach to previous methods and descriptors using the *Coffee&Cigarettes* (*C&C*) dataset [7], and present results for our new dataset *Hollywood–Localization*. We demonstrate that our use of a generic human tracker does not reduce performance over action specific methods; indeed our performance exceeds previous localization results. Furthermore, we show that the human tracks can be used for multiple actions, including: drinking, smoking, phoning and standing-up.

2 Datasets and Evaluation Method

We use two movie datasets in this work: *C&C* (in which we additionally annotate the smoking action) and our new *Hollywood–Localization* dataset. The new dataset and the corresponding annotations will be made available online if the paper is accepted.

Coffee&Cigarettes. The film *C&C* consists of 11 short stories, each with different scenes and actors. The dataset *C&C* introduced by Laptev and Pérez [7] consists of 41 drinking sequences from six short stories for training and 38 sequences from two other short stories for testing. Additionally, the training set contains 32 drinking samples from the movie *Sea of Love* and 33 drinking samples recorded in a lab. This results in a total of 106 drinking samples for training and 38 for testing.

We evaluate additionally on smoking actions. The *C&C* dataset also provides annotations for smoking, however, no results for localization have been reported in [7]. The smoking training set contains 78 samples: 70 samples from six short stories of *C&C* (the ones used for training the *drinking* action) and 8 from *Sea of Love*. 42 samples from three other short stories of *C&C* are used for testing.

We use the evaluation protocol of [7] in our experiments: an action is correctly detected if the predicted spatio-temporal detection has an overlap with the ground truth annotation $O(X, Y) \geq 0.2$. The overlap between a ground truth cuboid Y and a track segment X is given by $O(X, Y) = (X \cap Y)/(X \cup Y)$. Once an annotated sample has been detected, any further detection is counted as a false positive.

Hollywood–Localization. To evaluate the performance of our approach on challenging video data, we introduce the *Hollywood–Localization* dataset based on sequences from Hollywood movies [11]. In total we annotated 130 clips containing the action *answer phone* and 278 clips with the action *standing-up*. The same number of randomly selected clips not containing the action are used as negatives in each case. We keep the training/test movies split from [11] which roughly divides the samples into two halves.

Since *Hollywood–Localization* actions are much more dynamic, a cuboid is no longer an adequate representation for the ground truth. Therefore, the ground

truth we provide specifies an action by its temporal start and end frames, and a spatial localization rectangle for one of the intermediate frames. For evaluation we adapt the *C&C* protocol. The overlap in time is computed as $O_t(X,Y) = O(X_t,Y_t)$, and in space as $O_s(X,Y) = O(X_s,Y_s)$, where X_t and Y_t are the temporal extents of the track X and the annotation Y, and X_s and Y_s are the corresponding spatial rectangles in the annotated action frame. The final overlap is computed as $O'(X,Y) = O_t(X,Y) \times O_s(X,Y)$ and the accuracy threshold is set to 0.2 as for *C&C*.

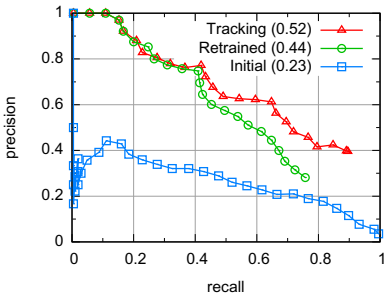
3 Human Detection and Tracking

To detect (i.e. localize) and track human actors we use the tracking-by-detection approach [20,21,22,23] that has proved successful in uncontrolled video. This involves detecting humans in every frame, and then linking the detections using a simple general purpose tracker. We illustrate the method here for human upper body detections using a HOG [12] descriptor and sliding window linear SVM classifier. The method is equally applicable to other human detectors – such as faces or whole bodies (pedestrians). Following [21], we use KLT [24] as the tracker. Since tracking and detection are not the main focus of this work we only concentrate on the novel aspects here. In particular, the *interpolation* of missed detections, and a *classification* stage for the final tracks in order to reduce false positives.

3.1 Upper Body Detection and Association by Tracking

For human actions in movies, an upper body detector [22,7] is suitable for medium and long shots. Based on Dalal and Triggs [12], the upper body detector is trained in two stages. In the **initial stage**, positive and negative windows are extracted from the *Hollywood–Localization* training movies. For this purpose we have annotated heads in keyframes and automatically extended them to upper bodies. Each annotation window is jittered [25] and flipped horizontally amounting to over 30k positive training samples in total. We sample about 55k negative training windows that do not overlap significantly with the positive annotations. For the second **retraining stage**, we follow the strategy of Dalal and Triggs [12] and look for high ranked false positives using the initial stage detector. We retrieve additional 150k false positives from the *Hollywood–Localization* training movies, and also add over 6k jittered positives and 9k negatives from the *C&C* training set.

Figure 1 (left) compares the precision-recall plots obtained for the two stages of the detector. We evaluate the predicted upper bodies using ground truth annotation for 137 frames of *C&C* [7] not used for training, for a total of 260 upper bodies. A person is considered to be correctly localized when the predicted and ground truth bounding box overlap (intersection to union) ratio is above 0.5. Re-training improves the precision for low recalls but with some loss of recall (blue initial and green retrained lines). However, the recall is largely recovered by



	recall					
	0.99	0.95	0.90	0.85	0.80	0.70
RBF-SVM	0.18	**0.42**	**0.60**	**0.68**	**0.73**	**0.78**
Lin-SVM	**0.19**	0.41	0.58	**0.68**	**0.73**	**0.78**
AvgScore	0.18	0.21	0.27	0.35	0.38	0.50
Occlusion	0.14	0.19	0.23	0.24	0.24	0.25
Length	0.14	0.15	0.18	0.22	0.26	0.27
No filtering	0.14	0.14	0.14	0.14	0.14	0.14

Fig. 1. (left) Upper body detector evaluated on *frames* from the *C&C* sequences not used for training. Average precision is given in parentheses. Note how precision is improved with detector retraining, and both precision and recall with tracking. **(right)** Precision of *tracks* for various filtering methods at recall rates of interest on *C&C* stories not used for training. Note the huge improvement obtained by classifying on a set of track properties, rather than using the properties individually.

the interpolating tracker (red line) which fills in missing detections (as described in section 3.2).

Upper body detections are associated between frames using a KLT [24] feature tracker. In a similar manner to Everingham *et al.* [26], the number of KLT features passing through two detections (both forwards and backwards in time) is used to compute a connectivity score between them, and detections are then linked by agglomerative clustering.

3.2 Interpolation and Smoothing

Detections can be missing in some frames, and hence the tracks formed by agglomerative clustering can have temporal gaps. To construct *continuous* tracks, it is necessary to fill in these gaps (otherwise the subsequent computation of the action descriptor is more difficult). Furthermore, the position and scale of the upper body detections can be noisy. In order to provide a stable reference frame for the subsequent action classification, we smooth (and complete by interpolation) the estimated detection window by optimizing over the track parameters $\{\mathbf{p}_t\}$:

$$\min_{\{\mathbf{p}_t\}} \sum_{t \in T} \left(||\mathbf{p}_t - \bar{\mathbf{p}}_t||^2 + \lambda^2 ||\mathbf{p}_t - \mathbf{p}_{t+1}||^2 \right) \tag{1}$$

where $\mathbf{p}_t = (x_t, y_t, w_t, h_t)$ denotes the position, width and height of a bounding box at time instance t for a track T, $\bar{\mathbf{p}}_t = (\bar{x}_t, \bar{y}_t, \bar{w}_t, \bar{h}_t)$ are the detections and λ is a temporal smoothing parameter. Note that if a detection is missed, then the appropriate term $\bar{\mathbf{p}}_t$ is removed from the cost function for that frame. Optimizing (1) results in a linear equation with a tri-diagonal matrix, which can be solved efficiently by Gaussian elimination with partial pivoting. Setting $\lambda = 4$ for 25Hz videos results in a virtual "steadi-cam" with no adverse oversmoothing.



Fig. 2. Upper body detections (top row) and tracks (bottom row) after classification post-processing for a sample test sequence of *C&C*. The bounding box colours indicate different tracks. Note the improvement due to the tracking where false positives have been removed, as well as the high accuracy despite motion, articulations and self-occlusions.

Figure 1 (left) shows the gain from smoothing and completing detections to form tracks. Exploiting the temporal consistency (tracking) significantly improves the recall of the retrained human detector.

3.3 Classification Post-processing

Since the upper body detector considers only a single frame, background clutter can generate many false positives. Some of these are quite stable and survive tracking to produce erroneous human tracks that should be removed.

We take a principled approach and in a final stage train a classifier to distinguish correct from false tracks. To this end, we define 12 track measures based on track length (since false tracks are often short); upper body SVM detection score (false detections normally have a lower score than true ones); scale and position variability (those often reveal artificial detections); and occlusion by other tracks (patterns in the background often generate a number of overlapping detections). For these measures we compute a number of statistics (min, max, average) where applicable and form a 12-dimensional feature vector used to classify the track. We obtain ground-truth for the tracks using 1102 annotated keyframes from *Hollywood–Localization* training movies (a track is considered positive if it coincides with an actor in the annotated keyframe, and negative otherwise) and train an SVM classifier (linear and RBF). The SVM is then used to classify the tracks.

Figure 1 (right) compares different methods used to remove erroneous tracks resulting from background clutter. The detection score turns out to be crucial for recognizing true human tracks. Nevertheless, training an SVM classifier on all 12 track measures significantly improves recognition precision compared to any heuristics on the individual measures. Using either a linear or a non-linear SVM, the precision at a useful recall of 0.8 improves from 0.14 to 0.73, i.e., the number of false positives is reduced by more than two thirds. The benefits to both precision and recall are evident in figure 1 (left).

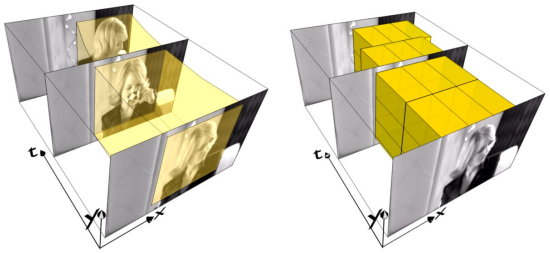


Fig. 3. The HOG-Track descriptor: (left) the human tracker detects and tracks a human upper body; (right) the *HOG-Track* descriptor divides the track into temporal slices. Each slice is aligned with the bounding box of its centre frame and is divided into a spatial grid of cuboid cells.

Overall, the proposed human detection and tracking method copes with a rich set of articulations, viewing angles and scales, as illustrated in figure 2, and results significantly improve over the individual human detections. Missed actors arise from unusual shots with camera roll, face close-ups or distant views. In crowded scenes, background actors might be missed, but most of the foreground characters are detected.

4 Action Localization

Given a set of human tracks, the goal is to determine which tracks contain a given action and to localize the action within the track. Our approach is based on a temporal sliding window, that is, we search for a range of frames which contains the action. Due to the tracks, the spatial extent of the action is already fixed. Consequently, we only need to delimit the beginning and length of an action (a two dimensional search space). This is in contrast with an exhaustive search, which needs to determine also the 2D image region corresponding to the human, i.e., its position and scale in the case of a sliding window approach.

Actions are represented by a spatio-temporal window descriptor. Our descriptor extends the HOG image descriptor [12] to spatio-temporal volumes, and goes beyond a rigid spatio-temporal cuboid [7,18], as it adjusts piecewise to the spatial extent of the tracks. This introduces a more flexible representation, where the description will remain centred on the deforming human action. This descriptor is termed *HOG-Track*, and is described in section 4.1. For temporal localization we use a state-of-the-art two stage sliding window classifier [27,28] on the tracks.

4.1 HOG-Track Descriptor

The *HOG-Track* action descriptor divides a track segment into cells. As in the original HOG [12], there are cells in the 2D spatial domain, but additionally the track segment is divided into temporal slices. These slices are aligned with a human track, as illustrated in figure 3. In more detail, a given track segment is defined by a temporal sequence of bounding boxes. This sequence is divided

into equally long temporal slices and the spatial image region corresponding to the slice is given by the bounding box of its centre frame. This ensures that our descriptor follows the variation of spatial position of a human within the spatio-temporal volume of the video.

Each slice is split into a spatial grid of cuboid cells as illustrated in figure 3 and each cell is represented by a histogram of spatio-temporal (3D) gradient orientations. Orientation is quantized over an icosahedron—a regular polyhedron with 20 faces. Opposing directions (faces of the icosahedron) are identified into one bin, i.e., there are a total of 10 orientations. Each gradient votes with its magnitude into the neighbouring bins, where weights are distributed based on interpolation.

For better invariance to position, we design spatially adjacent cells to have an overlap of 50%. All cell descriptors in a slice are L2 normalized per slice, and the final descriptor concatenates all cell descriptors. The parameters of the descriptor (the spatial grid and temporal slice granularity) are determined by cross-validation, as described in section 5. On the drinking and smoking actions the training performance is optimized for a spatial grid of 5×5 and 5 temporal slices. The dimensionality of the resulting descriptor is 10 orientation bins $\times$ 5^2 spatial cells $\times$ 5 temporal slices = 1250. This configuration is used in all our experiments.

4.2 Action Classification and Localization

Our temporal sliding window approach extracts descriptors at varying locations and scales. To classify these descriptors, we use a state-of-the-art two stage approach [27,28] which rejects most negative samples with a linear SVM, and then uses a non-linear SVM with an RBF kernel to better score the remaining samples.

When training the sliding window classifier, the ground-truth annotations are matched to the tracks and the action part of the track is used for training. The *HOG-Track* is computed for this temporal section, i.e., the temporal slices are aligned with the ground-truth begin and end time stamps of the action. The spatial regions are obtained from the track bounding box of the centre frame of each slice. Training is very similar to the detector training of section 3: additional positives are generated here by jittering the original positives in time, duration, and spatial scale. Initial negative samples are obtained by randomly sampling positions with varying lengths in the tracks, which do not overlap with any positive annotations, and in a re-training stage hard negatives are added to the training set. The C parameter and weight for positive samples are determined on the training set using a leave-one-video-out cross-validation. The second stage classifier uses a *non-linear* SVM with an RBF kernel and is trained on the same training data as the linear one. Again, we optimize the parameters via cross-validation.

At test time, a sliding window is used to localize actions. Seven temporal window scales are evaluated starting from a minimum length of $l = 30$ frames, and increasing by a factor of $\sqrt{2}$. The window step size is chosen as one fifth

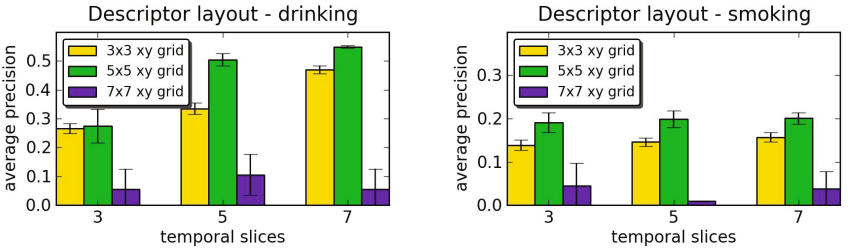


Fig. 4. HOG-Track descriptor evaluation: for a varying number of spatial cells and temporal slices for drinking and smoking actions on the *C&C* test dataset averaged over three runs.

of the current scale. The HOG-Track descriptor for each window is classified with the linear SVM. Non-maxima suppression then recursively finds the global maximum in a track and removes all neighbouring positive responses with an overlap greater than 0.3. The remaining detections are re-evaluated with the non-linear SVM classifier. As will be seen next, this second re-scoring stage improves classification results considerably.

5 Experimental Results

5.1 Coffee&Cigarettes

Tracks for Action Localization. Our action localization method depends on correct track positions in space and time. When training the sliding window classifier, the ground-truth is matched to the tracks and the corresponding tracks are used for training. We only keep samples that have an overlap of at least 0.5. This results in a loss of around 10% of the training samples. During testing an action can not be detected if the track is not localized. This reduces the maximum possible recall by again around 10%.

Descriptor Evaluation. In order to determine a suitable layout of our *HOG-Track* descriptor, we evaluate its parameters using cross-validation on the training set. Best results are obtained for 5 or 7 temporal slices; we use 5 as it results in a lower dimensional descriptor. The performance is quite sensitive to the number of spatial cells, best results are obtained for 5×5. This behaviour translates also to the test set which is illustrated in figure 4. The performance is averaged over three independent runs.

Localization Results & Comparison to State of the Art. Figure 5 presents precision-recall curves for localizing drinking and smoking actions in *C&C*. The detectors are trained on the training part of each dataset and evaluated on the corresponding test sets. Figure 5 (left) evaluates the detection results for localizing *drinking* actions. Under the same experimental setup, the linear classifier (50.8%) substantially outperforms the state-of-the-art, i.e., Willems *et al.* [18] (45.2%) and Laptev and Pérez [7] (43.4%). The non-linear classifier further improves the results (54.1%). Note the excellent precision (100%) up to a recall of

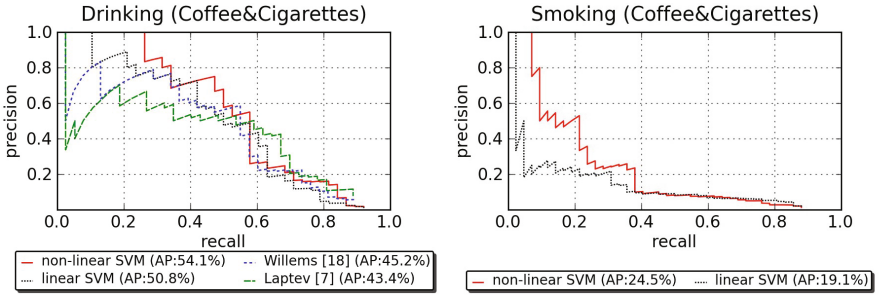


Fig. 5. Precision-recall curves on the *C&C* test set. Human actions evaluated: drinking (left) and smoking (right). We compare our linear and non-linear detectors and report state-of-the-art results where applicable.

ca. 30%. Figure 6 (top) illustrates the corresponding top 5 drinking localizations ordered by their SVM score. Note the variety of camera viewpoints and lighting.

Figure 5 (right) evaluates the detection results for localizing *smoking* actions. The non-linear classifier turns out to be crucial, improving the performance by +5.4% to 24.5% in terms of AP. The noticeably lower performance for smoking (when compared to drinking) can be explained by the large intra-class variability of this action. Temporal boundaries of a smoking action can in fact be only loosely defined and smoking often happens in parallel with other activities (like talking or drinking). Furthermore, a cigarette is smaller and less distinctive than a cup. Previous action analysis on this dataset [7,18] did not include smoking, so no comparisons can be given. The top 5 smoking localizations are shown in figure 6 (bottom). Interestingly, the false positive ranked 4th includes rapid vertical motion of the hand towards head and mouth.

Since drinking and smoking actions seem to be visually similar, it is interesting to assess the discriminative power of both classifiers. For this, we measure the performance of a drinking classifier for the task of localizing smoking and vice versa. Table 1 displays the *confusion* between the actions drinking and smoking. In both cases the performance is very low (around 5% AP) which shows that both



Fig. 6. The five highest ranked drinking (top) and smoking (bottom) detections on *C&C*. For drinking the first false positive (FP) is ranked 11th.

Table 1. Performance (AP) of drinking and smoking classifiers when localizing drinking and smoking actions. Note that the classifiers do not confuse the actions.

	Drinking action	Smoking action
Drinking detector	54.1%	5.3%
Smoking detector	5.0%	24.5%

classifiers are able to learn discriminative models that can distinguish visually similar, yet different actions successfully.

Comparison with Other Action Descriptors. To show the importance of computing the *HOG-Track* descriptor on the spatial extent of humans determined by tracks, we conduct experiments with a number of baseline classifiers. We keep the experimental setup and descriptor parameters the same.

First, we extract the HOG descriptor for the entire video frame, i.e., ignore the tracks. In this case the evaluation criterion only measures the overlap in time, as we do not determine the spatial extent. The average precision for the linear baseline classifier on the *C&C* drinking dataset is 8.1% (vs 50.8% with tracks) and for the non-linear one it is 17.1% (vs 54.1%). Clearly, such baseline is able to localize drinking actions to some extent, but its performance is inferior without the spatial localization provided by the tracks.

Next, we evaluate the importance of adapting the *HOG-Track* descriptor to tracks. We compute the descriptor for a spatio-temporal cuboid region tangent to the track. Precisely, we align the centre of the cuboid with the track, but do not "bend" it along the track. The performance for the linear classifier on drinking is 28.9% (vs 50.8% with adaptation) and this improves to 47.3% (vs 54.1%) with the non-linear classifier. This confirms the importance of descriptor adaptation.

Finally, we further evaluate the cuboid representation by performing an exhaustive (i.e., not using tracks) spatio-temporal search for an action. The non-linear classifier achieves an AP of 25.8% (vs 54.1%) for drinking. Figure 7 compares all these different methods. We also include results for the exhaustive cuboid search carried out by Laptev and Pérez [7]. Overall, using tracks to drive the action localization significantly outperforms the other approaches.

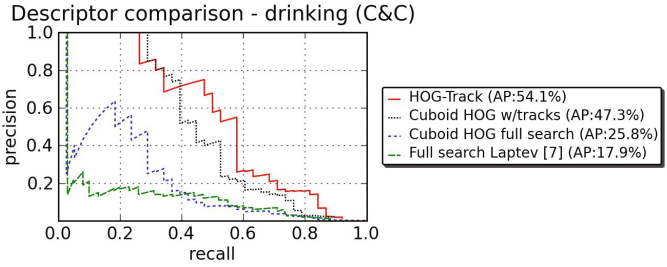


Fig. 7. Precision-recall curves comparing HOG-Track to other action descriptors on *C&C* for the action drinking

Complexity. In the following we investigate the theoretical and practical time complexity of our localization approach. We also discuss memory requirements and compare to an exhaustive "sliding cuboid" baseline.

For the theoretical analysis, without loss of generality we assume a linear one-against-rest classifier. We consider the number of multiplications in classifier evaluation (i.e., computing the dot product in the linear case) as the complexity measure. In a standard sliding window scheme the classifier is evaluated once for each window. Consequently, the total recognition cost will linearly depend on (a) the number of actions considered, (b) the number of windows evaluated, and (c) the dimensionality of the descriptor. The complexity of the "sliding cuboid" baseline can therefore be written as $O(a \cdot s_x^2 s_t \cdot r_x^2 r_t)$ where a is the number of actions, s_x/s_t denote spatial/temporal size of the problem (video), and r_x/r_t correspond to spatial/temporal resolution (dimensionality) of the descriptor.

Our approach combines a spatial sliding window human classifier and a temporal detector. Its complexity can be written as $O(s_x^2 s_t \cdot r_x^2 + a \cdot t s_t \cdot r_x^2 r_t)$ where t corresponds to the number of tracks in the video. Note that the above expression is normally dominated by the spatial search (left term). Compared to the exhaustive approach, we gain from having an action-agnostic classifier (no factor a) and using a simpler detector first (no factor r_t). The temporal search (right term) is fast since it searches only one dimension and $t \ll s_x^2$.

In practice, the difference in the runtime is even more significant due to limited memory. Computing the video descriptor does not allow for many optimizations which are possible for a single frame/image – like precomputing or caching the gradient histograms for instance. This in practice adds another factor to the sliding cuboid complexity. It does not affect our method since in our case the complexity is dominated by human detection, where memory requirements are not a problem.

The theoretical analysis above is confirmed in practice. Processing about 25 minutes of video using our method takes about 13 hours in total on a standard workstation. Human detection takes under 10 hours, tracking humans adds 3 hours, action localization is performed in under 10 minutes. For comparison, running an exhaustive cuboid search on the same data takes over 100 hours.

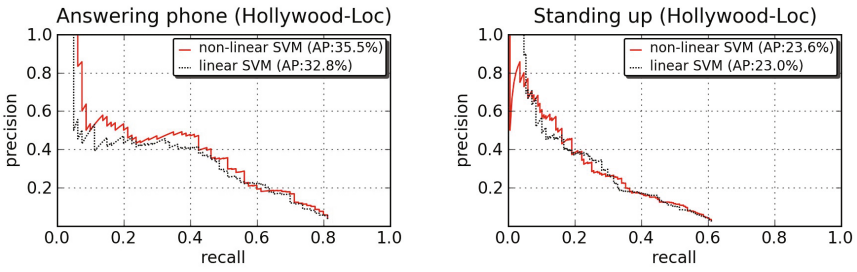


Fig. 8. Precision-recall curves for the actions answering phone and standing-up of the *Hollywood–Localization* test set

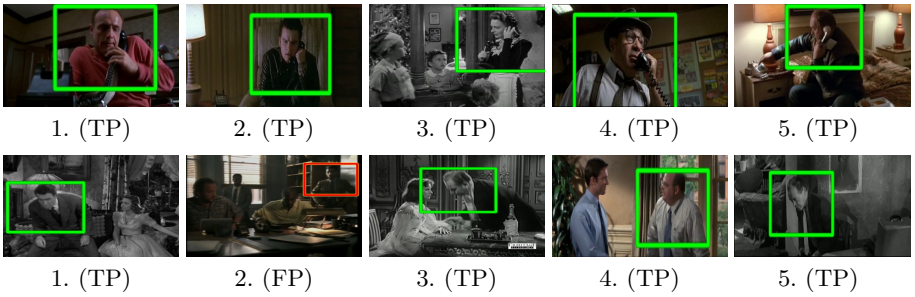


Fig. 9. The five highest ranked phoning (top) and standing-up (bottom) actions detected on *Hollywood–Localization*. For phoning the first FP is ranked 6th.

5.2 Hollywood–Localization

For this dataset we use the same parameters throughout as those used for *C&C*. Figure 8 (left) evaluates the detection results for localizing *phoning* actions in our *Hollywood–Localization* dataset. Due to the much larger variety of the videos (Hollywood movies), this dataset is much more challenging than *C&C*. The difficulty of the task is further increased by the fact that negative samples contain, without exception, other dynamic human actions. Some of those actions, like eating for example, might share similar motion patterns. Nevertheless, the recognition performance is satisfactory. In almost 40 minutes of video we can correctly localize over 80% of phoning actions and retrieve the top ones with high precision. The top 5 phoning localizations on the test set are shown in figure 9 (top). The true positive detections cover a large variety of poses and scenes. The top false positives mostly involve a rapid vertical hand movement.

Figure 8 (right) evaluates the detection results for localizing *standing-up* actions, and figure 9 (bottom) shows the top 5 detections. This action differs from the previous three as it does not involve the hand moving towards the head. The results are promising; the recall is worse than for all the other classes, but the precision is satisfactory.

6 Conclusion

We have demonstrated the value of using human tracks for visual action localization. In each dataset the same tracks support localization of different types of actions. This allows natural human actions to be effectively recognized in challenging environments.

A track introduces a separation between the human foreground and background of a scene, and either or both may provide information. In this paper we have proposed a robust model for foreground regions. In the future, given this separation, appropriate descriptors and classifiers can then be learnt for the foreground and background regions. For example, if the camera is panning to follow a person, then the motion from the background can be suppressed. However, for some actions it will be the background (the context) or background motion that is more informative, e.g. perhaps in the case of a person standing up.

Acknowledgements. This work was partially funded by the European research project CLASS, the MSR/INRIA joint project, and the ERC grant VisRec no. 228180.

References

1. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR (2004)
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV (2005)
3. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005)
4. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: ICCV (2007)
5. Schindler, K., van Gool, L.: Action snippets: How many frames does human action recognition require? In: CVPR (2008)
6. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
7. Laptev, I., Perez, P.: Retrieving actions in movies. In: ICCV (2007)
8. Mikolajczyk, K., Uemura, H.: Action recognition with motion-appearance vocabulary forest. In: CVPR (2008)
9. Rodriguez, M., Ahmed, J., Shah, M.: Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In: CVPR (2008)
10. Everingham, M., van Gool, L., Williams, C., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge. In: Workshop in Conj. with ICCV (2009)
11. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR (2009)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
13. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: ICCV (2009)
14. Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. In: CVPR (2009)
15. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: ICCV (2003)
16. Lu, W.L., Little, J.J.: Simultaneous tracking and action recognition using the pca-hog descriptor. In: CRV (2006)
17. Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: ICCV (2009)
18. Willems, G., Becker, J.H., Tuytelaars, T., van Gool, L.: Exemplar-based action recognition in video. In: BMVC (2009)
19. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: ICCV (2007)
20. Cour, T., Jordan, C., Miltsakaki, E., Taskar, B.: Movie/Script: Alignment and Parsing of Video and Text Transcription. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 158–171. Springer, Heidelberg (2008)
21. Everingham, M., Sivic, J., Zisserman, A.: Hello! My name is... Buffy – automatic naming of characters in TV video. In: BMVC (2006)
22. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR (2008)

23. Leibe, B., Schindler, K., van Gool, L.: Coupled detection and trajectory estimation for multi-object tracking. In: ICCV (2007)
24. Shi, J., Tomasi, C.: Good features to track. In: CVPR (1994)
25. Laptev, I.: Improvements of object detection using boosted histograms. In: BMVC (2006)
26. Everingham, M., Sivic, J., Zisserman, A.: Taking the bite out of automatic naming of characters in TV video. Image and Vision Computing 27, 545–559 (2009)
27. Harzallah, H., Jurie, F., Schmid, C.: Combining efficient object localization and image classification. In: ICCV (2009)
28. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV (2009)