# Retrieving actions in movies

Ivan Laptev and Patrick Pérez
IRISA / INRIA Rennes, Campus universitaire de Beaulieu
35042 Rennes Cedex France
{ilaptev | perez}@irisa.fr

## Abstract

*We address recognition and localization of human actions in realistic scenarios. In contrast to the previous work studying human actions in controlled settings, here we train and test algorithms on real movies with substantial variation of actions in terms of subject appearance, motion, surrounding scenes, viewing angles and spatio-temporal extents. We introduce a new annotated human action dataset and use it to evaluate several existing methods. We in particular focus on boosted space-time window classifiers and introduce "keyframe priming" that combines discriminative models of human motion and shape within an action. Keyframe priming is shown to significantly improve the performance of action detection. We present detection results for the action class "drinking" evaluated on two episodes of the movie "Coffee and Cigarettes".*

## 1. Introduction

Human actions are frequent and essential events within the content of feature films, documentaries, commercials, personal videos, and so forth. "Did Frodo throw the ring into the volcano?" "Did Trinity kiss Neo?" The answers to these and many other questions are hidden *exclusively* in the visual representation of human actions. Automatic recognition of human actions, hence, is crucial for video search applications and is particularly urged by the rapidly growing amounts of professional and personal video data (BBC Motion Gallery, YouTube, Video Google).

Interpretation of human actions is a well recognized problem in computer vision [2, 3, 1, 6, 9, 10, 16, 17, 19, 20, 22, 25, 26, 27]. It is a difficult problem due to the individual variations of people in expression, posture, motion and clothing; perspective effects and camera motions; illumination variations; occlusions and disocclusions; and the distracting effect of the scene surroundings. Figure 1 illustrates some of these difficulties on examples of drinking and smoking actions from the movie "Coffee and Cigarettes". To delimit the problem, previous work used a number of simplifying assumptions for example (a) restricted camera motion; (b) specific scene context, e.g. in field sports or surveillance scenes; (c) reliable spatial segmentation; and (d) restricted variation of view points. Notably, action recognition has not yet been addressed in unrestricted scenarios such as in feature films.

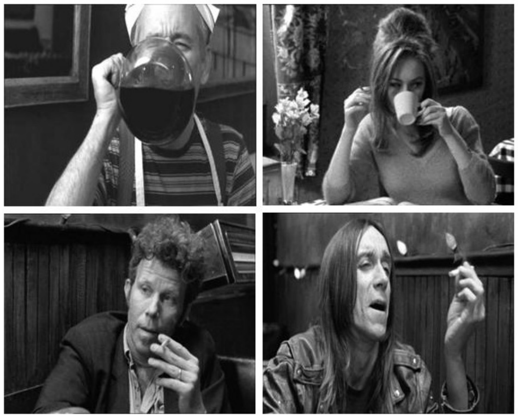


Figure 1. Examples of two action classes (drinking and smoking) from the movie "Coffee and Cigarettes". Note the high within-class variability of actions in terms of object appearance (top) and human motion (bottom). Note also the similarity of both action classes in the gross motion and the posture of people.

To deploy action recognition in video indexing applications, one needs to make a step from restricted scenarios and laboratory setups towards action recognition in generic videos. The main contribution of this paper is to undertake such a step and to address action class recognition and localization in movies with realistic variations of actions due to the changes of subjects, scenes, viewing angles, scales and other factors. To somewhat restrict the problem, however, we will focus on the "atomic" actions or *events* with a reasonably well-defined structure in time. Examples of such actions include "entering a room", "answering a phone", "shaking hands" and "drinking from a glass". In this paper we particularly study the detection of "drinking" actions and use movies "Coffee and Cigarettes" and "Sea of Love" for training and testing.

To assess the difficulty of the problem and the relative performance of different methods, we first study the task of *recognizing* pre-segmented samples of drinking actions among other human actions and motion patterns. As a promising method for action detection, we select a boosted space-time window classifier similar to the one proposed by Ke et al. [10]. We investigate the combination of shape and motion information and extend [10] to a joint shape-motion classifier. We also test two alternative methods either in terms of space-time interest points [16] or by recognizing actions from their static keyframes.

We next turn to the problem of action *detection*. We apply action classifiers to test videos in all space-time windows with variable spatial size and temporal extent. While such an exhaustive approach is shown to have a limited performance, we introduce a technique called Keyframe Priming that combines a space-time classifier with the static keyframe detection. Keyframe Priming is shown to significantly improve action detection and constitutes the second contribution of the paper. All detection results are reported for two test episodes of the movie "Coffee and Cigarettes" with 36,000 frames in total.

The rest of the the paper is organized as follows. In the next section we review related work. Section 2 presents the annotated action dataset. In Section 3 we describe action representations and methods for action classification. We study classification performance in Section 4 and introduce Keyframe Priming in Section 5. Results of action detection are reported in Section 6. Section 7 concludes the paper.

### 1.1. Related Work

A substantial body of work has been devoted to action recognition in the past. Due to the difficulty of the general problem, a number of simplifying assumptions are usually applied. 3D human tracking and 3D action recognition [22, 26] usually rely on the available tracks of body parts. An accurate segmentation of people from backgrounds is assumed in several methods, e.g. [2, 1] analyzing actions by the evolutions of silhouettes in time. Such methods assume in addition that actions can be recognized from the outlines of the body which is not always the case for many actions such as drinking actions studied in this paper. A number of methods exploit motion descriptors in the regions of actions [25, 27, 6, 16, 3, 17]. These methods often rely on external localization of people prior to action recognition with exception to [16, 3, 17]. All these methods are restricted to limited view variations while most of them have not been tested on datasets with large within-class variations of actions.

Learning discriminative action models appears to be a promising approach to handle the variation within action classes. To this end the methods [10, 19] present interesting alternatives to action detection and recognition and are closer to our approach in this paper. The method in [19], however, provides action localization in time only while in [10] action detection was studied for simple scenes with restricted view variations. The idea of Keyframe priming in Section 5 is related to [20] where human actions are recognized and tracked using a set of key postures.

## 2. Dataset and annotation

The importance of representative and annotated datasets for visual learning and recognition has been recently emphasized in the context of object class recognition [15]. Whereas for object recognition there now exist comprehensive datasets with thousands of realistic images for tens or hundreds object classes (Caltech 256, PASCAL VOC2006), the situation is different in action recognition where the existing datasets [16, 24] provide only a few action classes recorded in controlled and simplified settings with simple backgrounds, single action per scene, static camera, etc. This stands in sharp contrast to the fact that the within-class variability is likely to be higher for actions than for objects due to the additional time dimension in the data and to the involvement of multiple objects (e.g., hand, face and container in drinking).

### 2.1. Training and test sets

One reason for the lack of comprehensive and realistic action datasets is the difficulty of collecting and annotating videos of real human actions. To overcome this problem in this paper we make use of human actions in movies both for the training and testing purposes. In particular, we exploit the movie "Coffee and Cigarettes" (2003) providing an excellent pool of natural samples for action classes "drinking" (105 samples) and "smoking" (141 samples). These actions appear in different scenes, they are performed by different people while being recorded from different view points.

For the training of class "drinking" we use 41 drinking samples from 6 episodes of the movie "Coffee and Cigarettes"[1]. In addition we use 32 drinking examples from the movie "Sea of Love" and 33 drinking samples recorded in our lab. For the test we use the episodes "Cousins?" and "Delirium" from "Coffee and Cigarettes" containing 38 drinking actions and the total of 36,000 frames. *The training and the test sets have no overlap in subjects or scenes.* Figure 2 illustrates the large *within-class variability* of our drinking samples as well as the difference between the training and the test subsets. A few examples of scenes in Figure 6 illustrate the large variability of *scales, locations* and *view-points* as well as the typical "background clutter" with surrounding people acting in various ways.

[1] The training episodes from "Coffee and Cigarettes" are entitled "Strange to meet you", "Those things'll kill you", "No problem", "Jack shows Meg his tesla coil", "Cousins" and "Champagne".

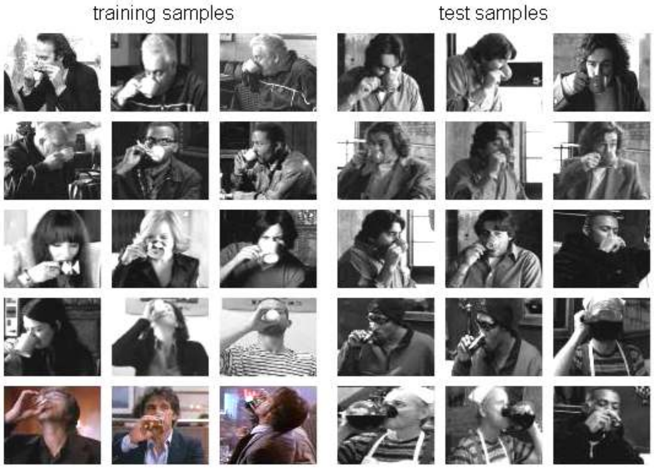


Figure 2. Selected examples of annotated actions of class "drinking" from the training episodes (left) and test episodes (right).

## 2.2. Annotation

We use annotation of drinking actions both for the alignment of training samples as well as for the evaluation of detection performance on the test set. Each drinking action is associated with a space-time volume defined by a cuboid $R = (p, \Delta p)$ with location $p = (X, Y, T)^\top$ and the spatio-temporal extent $\Delta p = (\Delta X, \Delta Y, \Delta T)^\top$ as illustrated in Figure 3. The spatial parameters of the cuboid are inferred by scaling and translating the manually annotated rectangle of the head such that $\Delta X = 1.6w$ for head width $w$ and $\Delta Y = 1.3h$ for head height $h$. To delimit actions in time, we manually select the frames corresponding to the start/end motion of a hand to/from the face. We also define the *keyframe* of an action by the time when the hand reaches the mouth. Temporal extents of drinking actions in our training set vary between 30 and 200 frames with the mean length of 70 frames. For each training action we generate several slightly randomized temporal annotations and in this way make the final classifier robust to the uncertainty of temporal extents of the action. Our action annotation is publicly available online[2].

# 3. Modeling

Treatment of generic video scenes imposes constraints on the choice of suitable methods for action interpretation. For example, we should not commit to the pre-defined locations and scales of actions in the scene, neither should we rely on static backgrounds nor on the presence of only one action at any time. Given the current progress in the respective domains, it may also not be appropriate to rely on the segmentation of human silhouettes nor on the precise localization of body parts. On the other hand, the chosen methods should be sufficiently efficient to enable the processing of several hours of video.

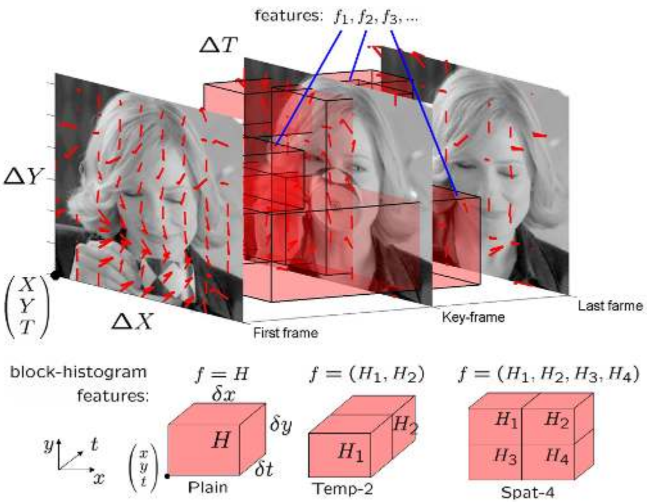


Figure 3. (Top): action volume in space-time represented by three frames of a drinking action. Arrows on the frames correspond to the computed optic flow vectors. Transparent blocks in red demonstrate some of the space-time features of a boosted space-time classifier. (Bottom): Three types of features with different arrangement of histogram blocks. Histograms for composed blocks (Temp-2, Spat-4) are concatenated into a single feature vector.

## 3.1. Boosted action classifier

In this paper we wish to exploit the consistent structure of "atomic" human actions in space-time. We build upon the intuition that such events in video can be treated similarly to the objects in images (see [21] for the related discussion in psychology). By taking this approach, we can benefit from the recent progress achieved in object class detection and recognition [4, 8, 11, 23]. In particular, we use discrete AdaBoost [7, 23] to learn a cascade of boosted action classifiers $C$

$$C(z) = \text{sgn}(\sum_{i=1}^{m} \alpha_i h_i(f_i(z))),$$

with $C$ making use of a linear combination of $m$ weak learners $h_i(f_i)$ defined for action features $f_i(z)$ of video $z$. Similar to [11] we use Fisher discriminant for the weak learners in combination with histogram features that represent actions as described below.

## 3.2. Motion and shape features

We wish to exploit appearance and motion of actions for action recognition. Shape representations in terms of histograms of image gradients have shown excellent performance on object recognition problems [4, 11, 13]. In this paper we use histograms of spatial gradient discretized in

[2] http://www.irisa.fr/vista/Equipe/People/Laptev/actiondetection.html

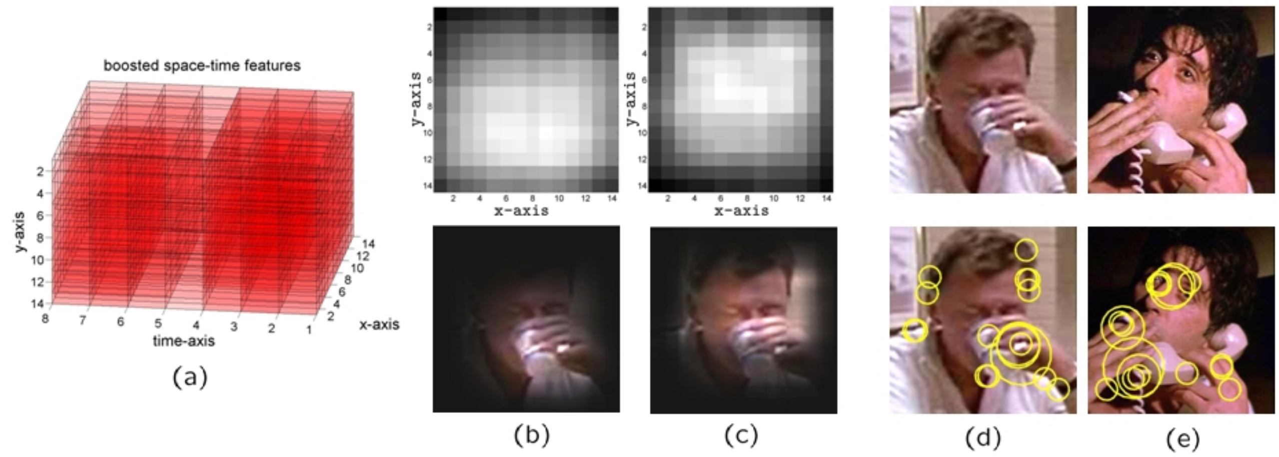


Figure 4. Video features. (a): All space-time features of OF5 classifier overlayed. High intensity values indicate high density of features; (b): Feature density for OF5 classifier projected on spatial coordinates (top) superimposed on the sample frame of a drinking action in (d) (bottom); (c): Spatial feature density for Keyframe classifier (top); superimposed on the keyframe of a drinking action in (d) (bottom); (d)-(e): Examples of drinking and smoking actions from the movie "Sea of Love" and corresponding detections of STIP features [12].

four orientations bins denoted by *Grad4*. To represent motion, we use 5-bin histograms of optical flow [14, 5] denoted by *OF5* with four bins corresponding to four discrete motion directions and the last bin corresponding to no motion.

Our action features $f$ are OF5 or Grad4 histograms accumulated in space-time blocks of the normalized action cuboid as illustrated in Figure 3. We assume that action annotation (Section 2.2) brings normalized action samples into a rough alignment in space and time and provides the correspondence for action features. Each feature $f_\theta(\cdot)$, $\theta = (x, y, t, \delta x, \delta y, \delta t, \beta, \psi)$ is defined by the space-time location $(x, y, t)$ and the space-time extents $(\delta x, \delta y, \delta t)$ of the histogram block, by the type of the block $\beta \in \{\text{Plain}, \text{Temp-2}, \text{Spat-4}\}$ (see Figure 3) and by the type of the histogram $\psi \in \{\text{OF5}, \text{Grad4}\}$.

The normalized action cuboid in our experiments has the size of $14 \times 14 \times 8$ histogram units with each unit corresponding to $5 \times 5 \times 5$ pixels. The number of all features $f_\theta(\cdot)$ defined on such a grid is $> 10^6$. To enable efficient learning, at each boosting round we select $10^3$ features with random positions and space-time extents. Our experience indicates that the random pre-selection of features does not decrease the performance of the final classifier which is consistent with similar conclusions made in [18].

To investigate the influence of shape information on action recognition we learn two classifiers, one with optic flow features only (*OF5* classifier) and another one with shape and motion features (*OFGrad9* classifier). OF5 classifier is closely related to the method by Ke et al. [10]. To efficiently compute feature vectors we use *integral video histograms* represented by the integral videos [10] for each histogram bin. To account for *scale variations* in the data we use a pyramid of integral video histograms with several spatial and temporal resolutions.

### 3.3. Keyframe classifier

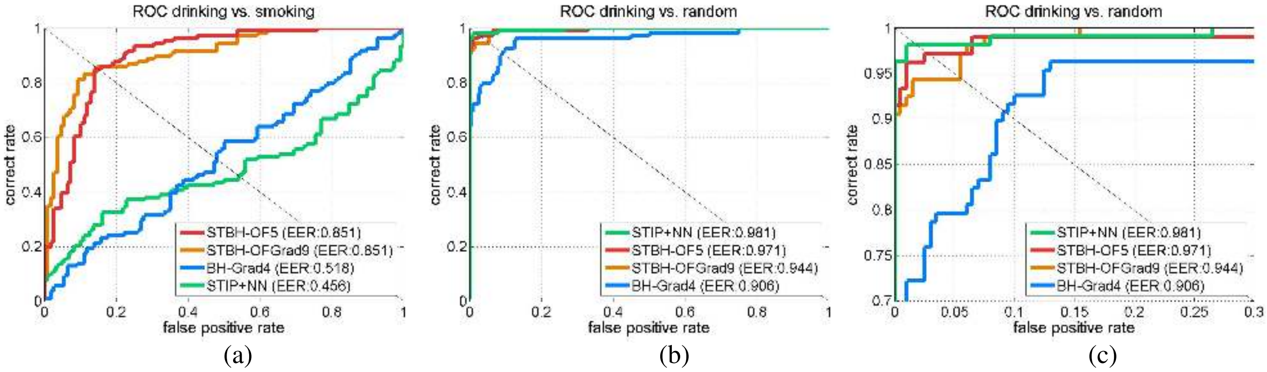
It is reasonable to ask the question if an action such as drinking can be recognized from a single frame using the state-of-the-art methods of object recognition. To investigate this possibility, we use a boosted histogram classifier [11] and train it on the *keyframes* of drinking actions while scanning random video frames excluding drinking to collect training examples for the background class (see Figure 3 and Section 2.2 for the definition of keyframes). The rectangular histogram features of the classifier are defined on keyframes represented by $14 \times 14$ grids of gradient orientation histogram blocks. The keyframe cascade classifier is trained until the false positive rate on the training set drops below $5 \cdot 10^{-5}$.

### 3.4. STIP-NN classifier

Space-time interest points (STIP) have been recently introduced [12] and applied to action recognition in [16]. This type of local motion descriptors does not rely on motion segmentation or other preprocessing steps and can be applied in complex scenes. We consider STIP in combination with the Nearest Neighbour (NN) classifier (see [16] for more details) as an alternative method for action recognition in this paper. STIP features for drinking and smoking actions are illustrated in Figure 4(d)-(e).

## 4. Classification

In this section we study the relative performance of classification methods introduced in Section 3. To assess the difficulty of the problem we in particular analyze the classification of drinking actions among either random motion patterns or among similar actions of the class "smoking".



(a) (b) (c)

Figure 5. Classification of drinking actions vs. (a): smoking actions and (b): random motion patterns. (c) is a magnified part of (b). ROC curves are obtained by thresholding on the confidence values of test samples. For boosted classifiers the confidence is defined by the number of passed cascade stages. For STIP-NN classifier the confidence is the normalized distance to the closest negative training sample.

### 4.1. Classifier properties

To better understand the performance of different methods we first study their properties. For space-time action classifiers (Section 3.1) the regions of features selected by boosting indicate the parts of the action that receive most of attention in classification. In Figure 4(a) we show all 559 features selected for the optic flow based classifier (OF5) in transparent color. From the density of color over time it is evident that this classifier has low activity at the keyframe while being primarily concerned with the motion at the start and the end of the action. When the selected space-time features are projected onto the $x$-$y$-plane in Figure 4(b,top) we observe the high concentration of features at the lower part of the action image. The most "active" parts of the classifier are, hence, associated with the regions of hand motion as illustrated in Figure 4(b,bottom).

A similar analysis for the keyframe classifier (Section 3.3) shows most of its selected (spatial) features being located at the upper part of the keyframe as illustrated in Figure 4(c,top). This classifier is, hence, mostly concerned with the head and the face regions on the keyframe as evident from the Figure 4(c,bottom). Interestingly, the learned keyframe classifier appears to be complementary to the space-time classifier by the location of selected features both in space and time. We use this property to combine classifiers in Section 5.

We also illustrate space-time interest points (STIP) used with the Nearest Neighbour classifier in Section 3.4. STIP features are detected at regions with high spatio-temporal variation of image values. In Figure 4(d)-(e) the detected STIP features of drinking and smoking actions often correspond to the moment of contact between the hand and the face. Such features, hence, may provide additional information for event classification.

### 4.2. Classification results

To challenge the recognition methods, we tested classification performance for two similar action classes "smoking" and "drinking". Test samples for both action classes were obtained from the test episodes of "Coffee and Cigarettes". The annotation of smoking samples followed the procedure described in Section 2.2. All test samples were cropped and normalized to the common rectangular cuboid in space-time. All classifiers were trained using the same set of (positive) drinking samples. The negative training samples for boosted cascade classifiers were obtained by collecting false positives from the training video episodes (see Section 2.1). For the STIP-NN classifier we explicitly provided negative training samples of smoking actions.

The classification results are illustrated in Figure 5(a) in terms of ROC curves and equal error rate (EER) values. The best results are obtained for the boosted OF5 and OFGrad9 space-time classifiers. The STIP-NN and the keyframe classifier failed this test given the close to chance performance for both methods. The performance of the keyframe classifier did not improve significantly even after explicitly retraining it on the negative action samples of smoking.

In the second test we classified drinking actions among other random motion samples obtained from the test episodes. As illustrated in Figures 5(b)-(c), the performance in this simpler test has been improved for all methods with the particularly high improvement for the STIP-NN and the keyframe classifier.

In conclusion we make two observations. First, the extension of OF5 classifier by shape information in OFGrad9 classifier did not improve classification performance in both tests despite our initial expectations. Second, the relatively high performance of all methods in the second test suggests that an improved performance can be achieved by a combination of complementary classifiers.

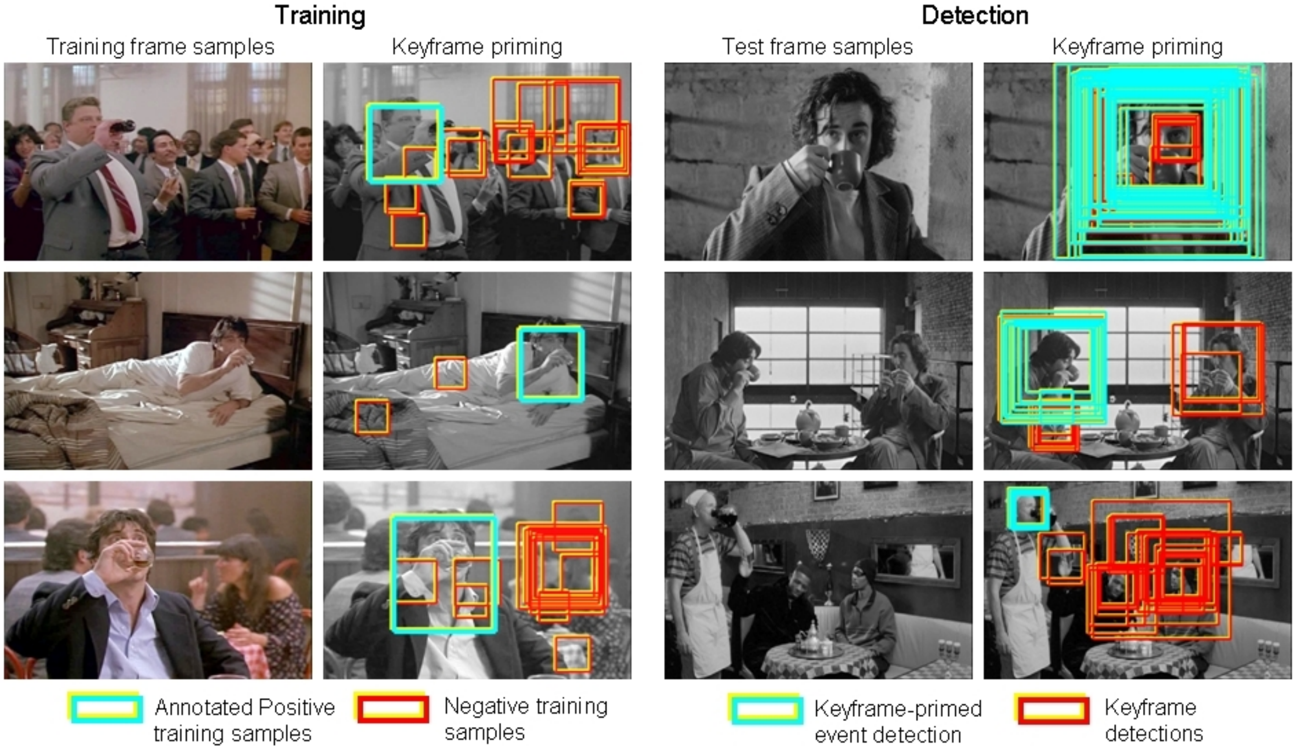


Figure 6. Keyframe priming. (Left): Examples of the training scenes from the movie "Sea of Love". False positive detections of the keyframe classifier (red) are used as negative examples during training of action classifiers. Keyframe detections are not shown for positive training samples (cyan). (Right): Examples of test scenes from the movie "Coffee and Cigarettes". The Keyframe primed action classifier correctly classifies drinking actions (cyan) among the detected keyframes (red).

## 5. Keyframe priming

This section describes a combination of the space-time action classifier and the keyframe classifier defined in Sections 3.1 and 3.3 respectively. The combination is motivated by the complementary properties of both classifiers observed in Section 4.1. More specifically our aim is to combine the discriminative model for the shape of people in drinking postures provided by the keyframe classifier with the discriminative model for the motion of people in action provided by the space-time classifier.

The combination in this paper is achieved by *Keyframe priming* and proceeds as follows. We bootstrap the space-time classifier and apply it to keyframes detected by the keyframe detector. The keyframe detector is first applied to all positions, scales and all frames of the video while being set to the rather high false positive (FP) rate $10^{-3}$ to ensure detection of all true positives (priming procedure). We then generate action hypotheses in terms of space-time blocks aligned with the detected keyframes and with different temporal extents to account for actions with different length in time. Each such block is classified according to the space-time classifier. Keyframe priming is used both to collect false positive samples for the negative training set during training as well as the priming stage of the final action detector. Keyframe-primed event detection is illustrated in Figure 6.

An acceptable FP rate for an action classifier in video is lower than for an object classifier in still images due to the temporal correlation of adjacent video frames. Reaching low values of FP rate, however, is computationally hard. With our current implementation we are able to learn a space-time classifier with FP rate $r^{st} \approx 5 \cdot 10^{-4}$. When using keyframe priming with the efficient implementation of keyframe detection adjusted to the FP rate $r^{kf} = 10^{-3}$, the FP rate of the combined action classifier becomes $r^{st} r^{kf} \approx 5 \cdot 10^{-7}$ which is a significant improvement. Hence, besides the combination of complementary models, Keyframe priming provides an additional and crucial benefit for the training of the action classifier. It also speeds up the detection at the test phase.

## 6. Detection

We next evaluate the performance of action detection obtained by space-time classifiers with and without Keyframe priming on the test set. For OF5 and OFGrad9 classifiers without Keyframe priming we generate and evaluate an exhaustive set of action hypotheses for a discrete set of spatio-temporal locations and plausible spatio-temporal scales in the test video. For action classifiers with Keyframe priming we follow the detection procedure described in the previous section. For each method we cluster multiple action detections with similar positions and sizes in space-time and use the size of the cluster as the detection confidence.

The performance for four tested detection methods is illustrated in Figure 7 in terms of precision-recall curves and average precision (AP) values. From this plot we observe a rather low performance of the two methods without Keyframe priming. It is worth noting that OFGrad9 method outperforms OF5 method due to the additional shape information.

Keyframe priming results in the significant improvement of the detection performance. The combination of shape and motion information, hence, indeed appears to be crucial for action recognition. The experiments indicate an advantage of learning shape and motion models separately (as in OF5 and keyframe classifiers) rather than jointly (as in OFGrad9 classifier).

To compare these results with previous methods, we note that Keyframe priming achieves a particularly strong improvement compared to the OF5 method without Keyframe priming. As mentioned earlier, OF5 classifier is closely related to [10]. Hence, we expect Keyframe priming would improve the performance of [10] significantly.

Twenty strongest detections obtained with the Keyframe primed OF5 classifier and sorted in the decreasing confidence order are illustrated in Figure 8. Most of the detections correspond to correctly retrieved actions with a substantial variation in subject appearance, motion, surrounding scenes, view points and scales in the video. We also observed many similar non-drinking actions (e.g. answering a phone) being correctly rejected by the detector.[3]

The detection speed for the Keyframe primed methods is currently around 3 seconds per frame using our implementation partly running in Matlab. We believe the detection speed can be improved with the faster implementation at least by the factor of ten.

## 7. Conclusion

We addressed recognition and localization of human actions in realistic scenarios with substantial variation in subject appearance, motion, surrounding scenes and view points. The variation in the data was handled either implicitly through learning or explicitly by searching actions at different space-time video resolutions. We in particular investigated the combination of shape and motion cues for action recognition and demonstrated improvement by combining discriminative models for human appearance and motion in action. We also introduced a new dataset with the annotation of human actions in movies.

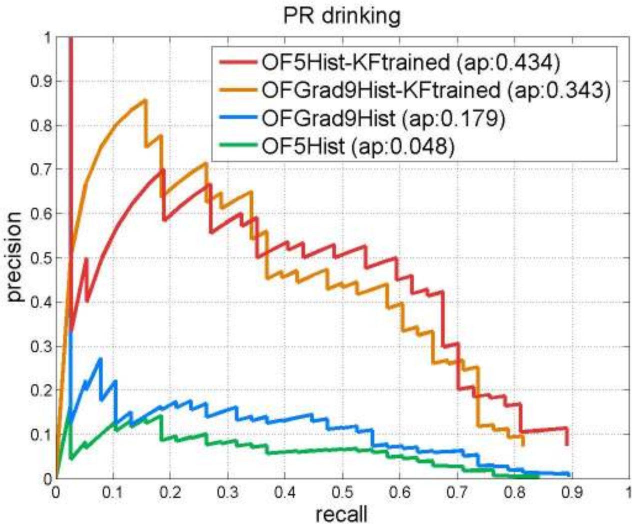


Figure 7. Precision-recall curves illustrating the performance of drinking action detection achieved by the four tested methods.

While this work examined one class of actions (drinking), we expect the presented analysis to generalize to other "atomic" actions and interactions such as "shaking hands", "kissing", etc. One of the current obstacles for action recognition is the lack of annotated data for natural human actions. This problem should be considered in the future. Concerning Keyframe priming, we believe that this technique could be extended by automatic methods for selecting possibly several keyframes for an action in order to improve detection performance and to reduce annotation efforts. The integration of other methods such as STIP [16] and behavior based similarity measure [17] may further improve detection performance.

## References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proc. ICCV*, pages II: 1395–1402, 2005.

[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE-PAMI*, 23(3):257–267, March 2001.

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *Proc. ICCV*, pages I:462–469, 2005.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, pages I:886–893, 2005.

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. ECCV*, pages II: 428–441, 2006.

[6] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. ICCV*, pages 726–733, 2003.

[7] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. of Comp. and Sys. Sc.*, 55(1):119–139, 1997.

[3] See a video with detection results at
http://www.irisa.fr/vista/Equipe/People/Laptev/actiondetection.html

[8] M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminative models for object category detection. In *Proc. ICCV*, pages II:1363–1370, 2005.

[9] D. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, 1999.

[10] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proc. ICCV*, pages I:166–173, 2005.

[11] I. Laptev. Improvements of object detection using boosted histograms. In *Proc. BMVC*, pages III:949–958, 2006.

[12] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. ICCV*, pages 432–439, 2003.

[13] D. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, 1999.

[14] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *DARPA Image Understanding Workshop*, pages 121–130, 1981.

[15] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 29–48. Springer Verlag, 2006.

[16] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proc. ICPR*, pages III:32–36, 2004.

[17] E. Shechtman and M. Irani. Space-time behavior based correlation. In *Proc. CVPR*, pages I:405–412, 2005.

[18] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc. ECCV*, pages I:1–15, 2006.

[19] P. Smith, N. da Vitoria Lobo, and M. Shah. Temporalboost for event recognition. In *Proc. ICCV*, pages I:733–740, 2005.

[20] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *Proc. ECCV*, volume 2350, page I:629 ff., 2002.

[21] B. Tversky, J. Morrison, and J. Zacks. On bodies and events. In A. Meltzoff and W. Prinz, editors, *The Imitative Mind*. Cambridge University Press, Cambridge, 2002.

[22] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *Proc. ICCV*, pages I: 403–410, 2005.

[23] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[24] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 103(2-3):249–257, November 2006.

[25] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. In *Proc. ICCV*, pages 120–127, 1998.

[26] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *Proc. ICCV*, pages I: 150–157, 2005.

[27] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proc. CVPR*, pages II:123–130, 2001.

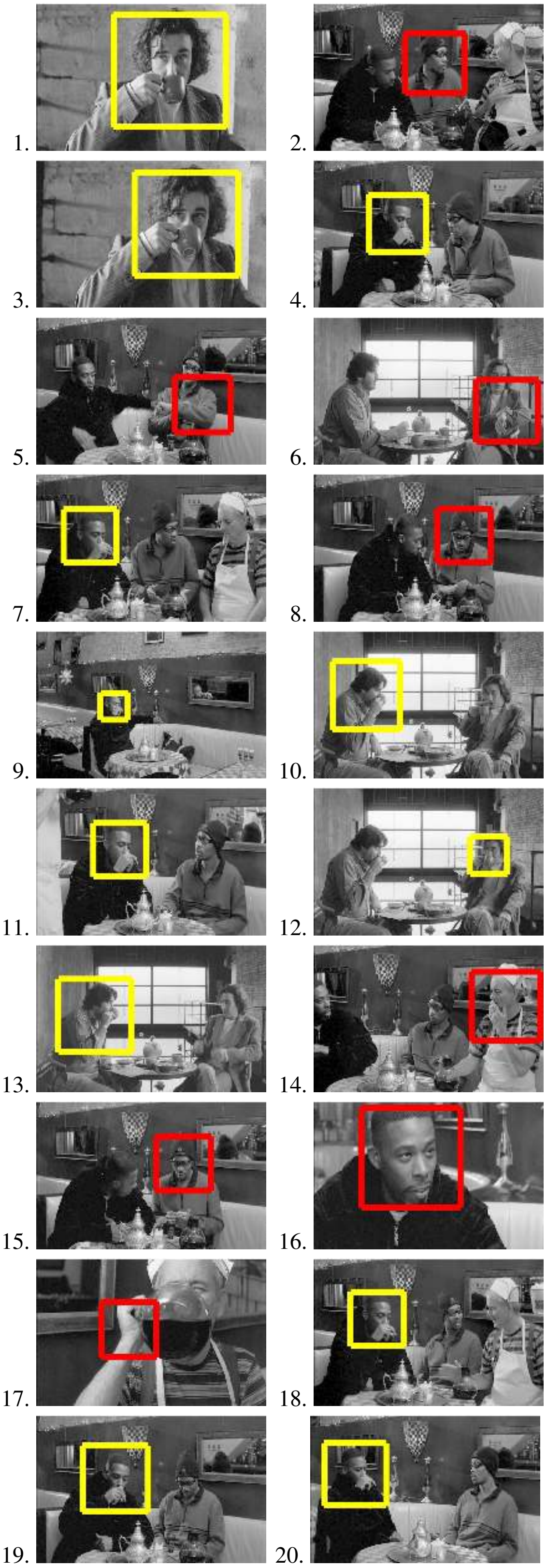


Figure 8. Detections of drinking actions (yellow: true positives, red: false positives) sorted in the decreasing confidence order and obtained with the OF5 Keyframe primed event detector.