Less is More: Video Trimming for Action Recognition

Borislav Antić, Timo Milbich and Björn Ommer
Heidelberg Collaboratory for Image Processing
University of Heidelberg, Germany
{borislav.antic,timo.milbich,bommer}@iwr.uni-heidelberg.de

Abstract

Action recognition is an important precursor for understanding human activities in videos. The current paradigm of action recognition is to classify a video sequence as a whole. However, actions usually occur only in part of a video sequence, rendering the rest of the video irrelevant for action recognition. In this paper, we propose a method for learning a subsequence classifier which can detect and classify part of a video that corresponds to the action. The subsequence classifier is trained from weakly labeled training videos whose subsequence labels are not provided, but need to be inferred during learning. We use the framework of multiple instance learning to solve two problems jointly: i) find the action subsequences in training videos, ii) train the subsequence classifier using the inferred action subsequences. To obtain a robust solution to the MIL problem, we propose a sequential algorithm that consecutively decreases the number of inferred action subsequences per video and trims their length until only one short subsequence is used as the action representative in each video. We evaluate the combination of the automatically trained subsequence classifier and the full sequence classifier on the very challenging Hollywood2 benchmark set and observe a significant gain in the performance over the baseline full sequence classifier. Moreover, a favorable performance of the subsequence classifier for temporal localization of actions in videos is evidenced on two categories of the Hollywood2 dataset.

1. Introduction

Action recognition in video is an important research problem with applications ranging from surveillance and security to web-based video sharing (e.g. YouTube and Google videos). Action recognition is also a prerequisite for the recognition of complex activities that are defined as sequences of elementary actions [6]. Action recognition is a difficult problem because videos are usually unconstrained and actions show a large intra-class variability. Moreover, videos typically contain adverse effects, such as large camera motion, background clutter, occlusion or change of illumination. As the actions normally occur only in part of the video sequence, a large part of the video sequence contains irrelevant clutter that interferes with action classification.

Due to the laborious and costly nature of the video annotation process, video datasets are usually only weakly labeled. An action label is assigned to a video, if an instance of the corresponding action class occurs in the video. However, information about the spatial and temporal location of the action in the video is not provided. Because of that, most action recognition approaches so far have proposed the full sequence classifier as a way to assign a label to the video. Such methods are typically based on the bag-of-features (BoF) representation [14, 20], i.e. a histogram of visual words extracted from the whole video sequence is calculated and then used by a discriminative classifier to predict the action label of a video. Recent methods try to incorporate the spatiotemporal structure of the action to get better classification performance. Laptev *et al.* [15] represent a video by histograms of visual word occurrences over the spatiotemporal volume corresponding either to the entire video or multiple subsequences defined by a video grid. However, the grid is fixed and does not adapt to the action contained by the video. A human-centric approach [12] detects and localizes human actions in challenging videos using the generic human detector and tracker, and the actions are detected only within the discovered human tracks. However, the method relies on a generic human detector and tracker that is trained on an external dataset which is typically not provided.

As actions in videos typically occur only in a small part of a video, we propose a method to learn a subsequence classifier that can detect and classify part of a video that corresponds to the action. The subsequence classifier is then applied on all subsequences in a video and selects the subsequence with the highest classification score. The classification results of the subsequence classifier and the full sequence classifier are eventually combined into a final action classification result. We propose a method that learns the subsequence classifier *automatically* from the weakly labeled training videos whose subsequence labels are not provided, but have to be inferred during learning. Thus, we solve the following two problems jointly: i) find the action subsequences in training videos, and ii) train the subsequence classifier using the inferred action subsequences. We cast the two problems in the framework of multiple instance learning (MIL), so that video sequences become *bags*, and subsequences of a video become *instances* of a bag. In MIL, the subsequence classifier is updated from the subsequences inferred in the previous round of MIL, and then it is used to infer the subsequence labels during the next round of MIL. Standard MIL keeps the length of subsequences fixed and selects only one subsequence per video for training the subsequence classifier. To improve the robustness of the MIL method, we propose a sequential method for multiple instance learning of the subsequence classifier that consecutively decreases the number of inferred action subsequences per video and trims their length until only one short subsequence is used as an action representative in each video. We evaluate the combination of the automatically trained subsequence classifier and the full sequence classifier on the very challenging Hollywood2 benchmark set and observe a significant gain in the performance over the baseline full sequence classifier. Moreover, a favorable performance of the subsequence classifier for temporal localization of actions in videos is evidenced on two categories of the Hollywood2 dataset.

The paper has the following structure. In Sect. 2 we review the related work on the action recognition in videos. Our method for sequential multiple instance learning of the subsequence classifier is described in Sect. 3. Experimental results and their comparison to the state-of-the-art are given in Sect. 4. Finally, Sect. 5 concludes the paper.

2. Related Work

Action recognition in videos has attracted a great attention in the computer vision community over the last decade. A large body of literature on action recognition in videos is addressing the question of feature detection and description in videos. Laptev [14] proposed spatiotemporal interest points obtained by the Harris detector that has been extended to the spatiotemporal domain. A novel cornerness measure that combines the Gaussian filter in space and Gabor filter in time is proposed by Dollár *et al.* [8]. Willems *et al.* [22] proposed to detect spatiotemporal interest points at places where the determinant of the spatiotemporal Hessian matrix is maximal. Recently, Wang *et al.* [21] showed that dense sampling of feature points has better performance than interest point detectors on the challenging video datasets.

Best results on the challenging action recognition datasets such as Hollywood2 have been obtained by building a bag-of-features (BoF) representation for the whole video [15, 20]. The video is then classified using the kernel support vector machine (SVM) based on the L1-normalized histogram of visual codewords that is calculated from the quantized spatiotemporal interest points. Laptev *et al.* [15] proposed to concatenate BoF representations of subvolumes defined on the grid in video. Ommer *et al.* [19] develop a generic action recognition system that is combined with compositional object segmentation and tracking. Hoai and De la Torre [11] propose a max-margin framework for training the temporal event detector to recognize partial events, which enables early event detection. Their method is based on the Structured Output SVM, that they extend to accommodate sequential data. However, training of their event detector is supervised and requires an additional annotation.

Gaidon *et al.* [10] propose a model that uses a sequence of atomic action units, termed actoms, to represents the temporal structure of the action as a sequence of histograms of actom-anchored visual features. The actom model is trained from the actoms-annotated video clips. Niebles *et al.* [18] propose to model the complex activities as temporal compositions of motion segments. They train a discriminative model to find a temporal decomposition of a complex activity. Antić and Ommer [2] create a system for video parsing that recognizes normal and abnormal objects and actions in the scene. Lan *et al.* [13] develop an algorithm for action recognition and localization in videos that uses a figure-centric visual word representation. The person location is treated as a latent variable that is inferred jointly with the action recognition. Their model is learned from training videos that are annotated with action labels and bounding boxes around the people performing the action.

The MIL framework has gained in popularity in recent years because of its powerful mechanism to deal with label ambiguities that are common in weakly annotated datasets. The idea of MIL quickly spread from drug activity prediction [7] to many other disciplines such as document analysis [1] and computer vision [4, 17, 9, 5]. MIL problem is normally tackled by alternating the steps of learning the instance-level classifier and imputing the missing instance labels. Antić and Ommer [3] proposed the concept of superbags, a random ensemble of sets of bags, whose goal is to decouple inference and learning steps in MIL and thus improve its robustness.

3. Sequential Multiple Instance Learning of Subsequence Classifiers

To recognize an action in video, we first need to find a subsequence that corresponds to the potential action in video and then classify it with a subsequence classifier. However, to train the subsequence classifier we need a set of training subsequences that correspond to the action. As the subsequences in training videos are not annotated, we

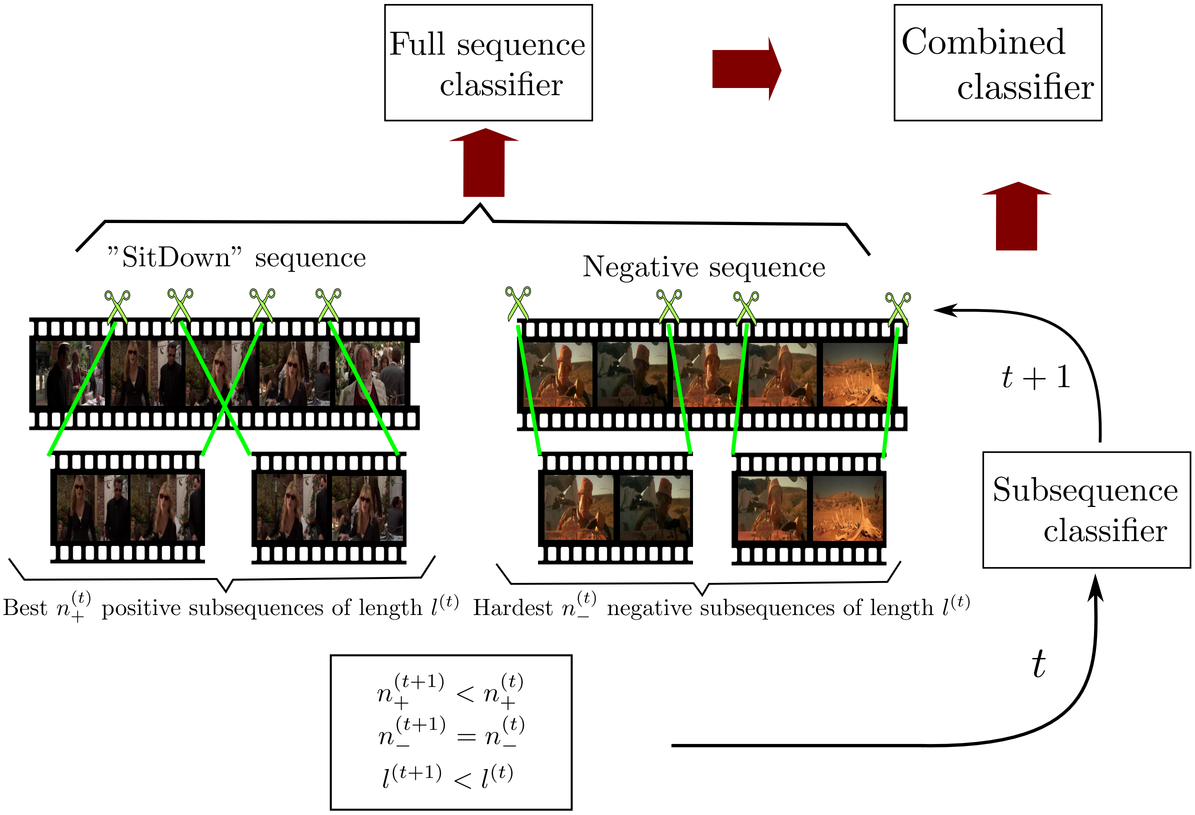
Full sequence classifier
Combined classifier
"SitDown" sequence
Negative sequence
$t+1$
Subsequence classifier
Best $n_+^{(t)}$ positive subsequences of length $l^{(t)}$
Hardest $n_-^{(t)}$ negative subsequences of length $l^{(t)}$
t
$n_+^{(t+1)} < n_+^{(t)}$
$n_-^{(t+1)} = n_-^{(t)}$
$l^{(t+1)} < l^{(t)}$

Figure 1: The sketch of our method: Subsequence classifier is updated in each iteration t using the best $n_+^{(t)}$ subsequences per positive video and the hardest $n_-^{(t)}$ subsequences per negative video. Subsequence selection and classifier training are alternated T times. The number $n_+^{(t)}$ of positive subsequences is decreased during training. The number $n_-^{(t)}$ of negative subsequences is fixed. Positive and negative subsequences have the same length $l^{(t)}$ that also decreases during MIL. The scores of the full sequence classifier and the subsequence classifier are then combined into a final classification score.

need to solve two problems jointly: i) training of the subsequence classifier, and ii) detection of subsequences from the training videos that correspond to the action. We use the multiple instance learning (MIL) to solve these two problems jointly.

In multiple instance learning, training set consists of a number of *bags* X_i, and each bag is associated with a binary label $Y_i \in \{-1,+1\}$. Bag X_i consists of a number of *instances* $X_i = \{x_j^i : 1 \leq j \leq m_i\}$ whose labels $y_j^i \in \{-1,+1\}$, $1 \leq j \leq m_i$ are not provided. Bag X_i has the positive label $Y_i = +1$ if *at least* one of its instances is positive, i.e. $\exists j : y_j^i = +1$. If all instances in a bag are negative, $\forall j : y_j^i = -1$, the bag is labeled as negative, $Y_i = -1$. The goal of MIL is then to learn an instance-level classifier from training instances whose labels are not provided and thus have to be inferred during training.

In the video subsequence classification problem, bag X_i corresponds to the full video sequence, and instances x_j^i are the subsequences of the full video sequence. For a positive video $Y_i = +1$, we are looking for a subsequence that corresponds to the action in the video, i.e. we want to find a *witness* $x_{s_i}^i$ of the positive class in each positive bag X_i. This is formulated as the multiple instance learning with instance selection (MI-SVM, Andrews *et al.* [1]) and leads to the following joint optimization problem,

$$\min_{s} \min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \tag{1}$$

$$\text{s.t. } w^\top x_{s_i}^i + b \geq 1 - \xi_i, \text{ for } Y_i = +1,$$
$$-w^\top x_j^i - b \geq 1 - \xi_i, \text{ for } Y_i = -1 \text{ and } 1 \leq j \leq m_i.$$

A standard solution to this optimization problem is to perform the following two steps iteratively: i) find the witness $x_{s_i}^i$ of a positive bag X_i which is an instance with the highest score given by the instance-level SVM classifier, and ii) re-train the instance-level SVM classifier (w,b) using the instances selected in the previous step. The iterative procedure produces only a locally optimal solution of the SVM model.

We improve the robustness of the MIL training for the task of subsequence classification in two ways. First, we select longer subsequences from positive videos in early rounds of MIL and then slowly decrease their length in later rounds. This allows us to gradually discover the action related part of the sequence when duration of the action is short. Subsequences from the negative bags have the same length as the positive subsequences, and thus they are also decreased successively during the MIL training. The length of subsequences used during MIL training is hence given as a monotonically decreasing sequence,

$$l^{(0)} > ... > l^{(t)} > ... > l^{(T)} \neq 0. \qquad (2)$$

Secondly, we increase the recall rate of the positive instances by selecting initially several instances per positive bag and then slowly reducing their number until only one instance is selected per bag. Thus, the number of positive instances that is selected by the MIL algorithm is a monotonically decreasing sequence,

$$n_+^{(0)} > ... > n_+^{(t)} > ... > n_+^{(T)} \neq 0. \qquad (3)$$

For negative videos, we can train the SVM model with all subsequences that are extracted from the video. As there is a large number of subsequences in a video, we actually work only with a small number $n_-^{(t)}$ of hard negative subsequences per bag. The hard negative subsequences are those that have the least negative score produced by the subsequence classifier, and thus they define the margin in the max-margin classification setup. The number of hard negative instances per bag is kept fixed throughout the MIL training,

$$n_-^{(t)} = const., \ 0 \leq t \leq T. \qquad (4)$$

In short, these are the steps of our algorithm for learning the subsequence classifier:

1. Sample randomly a number of long subsequences from positive and negative video sequences and train the initial subsequence classifier from them.

2. Use the subsequence classifier to select the best $n_+^{(t)}$ positive subsequences of length $l^{(t)}$ from each positive video.

3. Use the subsequence classifier to select the hardest $n_-^{(t)}$ negative subsequences of length $l^{(t)}$ from each negative video.

4. Re-train the subsequence classifier from the selected positive and negative subsequences.

5. Repeat steps 2. - 4. T times while reducing the length of all subsequences and the number of positive subsequences, and also keeping the number of negative subsequences fixed.

We use the standard bag-of-features representation (BoF) for the classification of videos and their subsequences. The features are obtained as dense trajectories using the method of Wang *et al.* [20]. Each feature is encoded using the trajectory, HoG, HoF and MBH descriptors. The trajectory descriptor represents the shape of motion trajectory. HoG and HoF encode the static appearance and the optical flow field in a local spatiotemporal volume around the trajectory. MBH is the robust representation of the motion discontinuities in video, and it is invariant to the camera motion. Features are quantized using the vocabulary learned by the k-means clustering method from randomly sampled training features. Each feature is quantized into the nearest codeword, and the histogram of all codewords is computed to produce the BoF representation of the full sequence and the subsequences. The non-linear support vector machine is used to predict the action classes based on the BoF representation. We use the histogram intersection kernel with non-linear SVM classifier because of good classification performance and the existence of fast computation method for this kernel (Maji *et al.* [16]).

To classify a novel video, we combine the scores of the full sequence and the subsequence classifiers. The full sequence classification score is computed from the BoF representation of the whole video. The classification score of the subsequence classifier is obtained by applying the subsequence classifier on all subsequences of the same length as the final-round subsequences of the MIL training. We take the maximal subsequence classification score and combine it with the full sequence classification score to yield the final score for the video. The combination is performed by a nonlinear support vector machine that uses the RBF kernel whose bandwidth is optimized on the validation set. The sketch of the proposed approach is shown in Figure 1.

4. Experimental Results

We evaluate our approach on the state-of-the-art Hollywood2 benchmark set for the action recognition. The Hollywood2 dataset consists of 12 action categories: answering the phone, driving a car, eating, fighting, getting out of a car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up. The dataset contains 1707 videos that are divided into the training set (823 video clips) and the test set (884 video clips). All video clips are taken from 69 Hollywood movies and they contain realistic, unconstrained human actions with large amount of camera motion. Video sequences vary in length from $5-25$ seconds, and most of the actions are only $1-2$ seconds long. Therefore, the length of the action subsequence is several times shorter than the length of the whole video. This motivates our subsequence classification approach which automatically finds and classifies the part of the video that corresponds to the action. Video sequences in train and test

	Full seq. classifier	Comb. classifier (full + subseq.)
AnswerPhone	28.76%	31.40%
DriveCar	89.13%	89.15%
Eat	61.78%	66.42%
FightPerson	80.47%	82.89%
GetOutCar	49.72%	53.91%
HandShake	30.65%	30.75%
HugPerson	51.25%	55.90%
Kiss	64.04%	65.72%
Run	82.30%	82.71%
SitDown	63.20%	66.25%
SitUp	20.93%	21.46%
StandUp	66.82%	71.09%
mAP	57.42%	59.80%

Table 1: The comparison of the full sequence classifier (baseline) and the combination of the full sequence and subsequence classifiers on all action classes of the Hollywood2 dataset.

Hollywood2	
Wang *et al.* (2009)	47.7%
Taylor *et al.* (2010)	46.6%
Ullah *et al.* (2010)	53.2%
Gilbert *et al.* (2011)	50.9%
Le *et al.* (2011)	53.3%
Wang *et al.* (2011)	58.2%
Wang *et al.* (2013)	59.9%
Combined classifier	59.8%

Table 2: Comparison of our combined classifier (full sequence + subsequence classifier) to the state-of-the-art methods on the Hollywood2 dataset.

set are taken from different movies. We follow the evaluation protocol of Laptev *et al.* [15] and train the one-against-all classifier for each action category. Precision-recall (PR) curves are calculated from the classification scores, and per-class performance is based on the average precision (AP) values computed from the corresponding PR plots. Overall performance is reported as the mean average precision over all classes (mAP) and it is used for the comparison with the state-of-the-art.

The full sequence classifier is used as the baseline in our comparison in Table 1. We use the same parameters as in the paper of Wang *et al.* [20], except for the choice of the kernel function. We use the intersection kernel because there exists a fast method for its computation. Wang *et al.* use χ^2 kernel function that yields 0.8% higher performance, but at the price of slower computation of the kernel matrix (about 3 times longer training and test time).

Category	Overlap
SitDown	48.6%
GetOutOfCar	34.4%

Table 3: The detection performance of the subsequence classifier on two Hollywood2 categories (SitDown and GetOutOfCar) for which we created the ground truth. The detection performance expressed as the PASCAL's overlap (intersection over union) score.

Table 1 shows the comparison of the full sequence classifier (baseline) and the combined classifier that aggregates the scores of the full sequence and subsequence classifiers. Finding a subsequence in a video that represents the action and classifying it with the subsequence classifier yields an improvement over the baseline of $1 - 5\%$ in average precision for most of the Hollywood2 classes. Largest improvements are for the classes StandUp, HugPerson and Eat, where the gain is almost 5%. On the other hand, classes such as DriveCar or Run give only a marginal improvement over the baseline. After averaging all the classes, the combination of the subsequence and full sequence classifiers yields 2.5% better performance than the baseline (59.8% vs. 57.4%).

Table 2 shows the comparison of our method and the state-of-the-art methods on Hollywood2 dataset. We see that the combination of subsequence and full sequence classifier outperforms almost all of the state-of-the-art results, in spite of using slightly inferior kernel function than other methods (intersection kernel vs. χ^2). The most recent result on the Hollywood2 dataset (Wang et al. 2013) combines the BoF representation with spatiotemporal pyramids and yields 59.9% mAP. The performance of our method is only 0.1% weaker than this, although our method does not use the stronger, but computationally more expensive, pyramidal features.

In Table 3 we show the results of detection performance for our subsequence classifier. We evaluate this performance using the PASCAL overlap (intersection over union) criterion on two categories of the Hollywood2 dataset for which we have manually labeled the ground truth subsequences. The overlap of detected subsequence and the ground truth for SitDown class is 48.6%, whereas for GetOutOfCar the overlap is 34.4%. We consider these as very good results, because the training of our subsequence detector is automatic.

Finally in Figure 2 to 4 we give some qualitative results of subsequence detection for two action classes of Hollywood2. Figure 2 and 3 illustrate the successful subsequence detection. In Figure 4 detection failed, most probably because of the large camera zoom on the main actor's face, making his body motion hardly visible while he is performing the "SitDown" action.

5. Conclusion

Actions in video typically occupy only a small part of a full video. In this paper we propose a method to automatically learn a subsequence classifier that can detect part of the video that corresponds to the action. As the subsequences of training videos are not annotated, our method jointly trains the subsequence classifier and labels the action subsequences in training videos. The framework of multiple instance learning is used to find subsequences that correspond to the action which then allow to train the subsequence classifier. To obtain a robust solution to the MIL problem, we propose a sequential algorithm that consecutively decreases the number of inferred action subsequences per video and trims their length until only one short subsequence remains as the action representative in each video. The combination of the automatically trained subsequence classifier and the full sequence classifier is evaluated on the challenging Hollywood2 dataset where it yields a significant improvement over the baseline performance on many action classes. We also examine the temporal localization performance of the subsequence classifier on two categories of Hollywood2 dataset and achieve promising results.

Acknowledgment

This work has been supported by the German Research Foundation (DFG) within the program "Spatio-/Temporal Graphical Models and Applications in Image Analysis", grant GRK 1653, and by the Excellence Initiative of the German Federal Government, DFG project number ZUK 49/1.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, pages 561–568. MIT Press, 2003.

[2] B. Antic and B. Ommer. Video parsing for abnormality detection. In *ICCV*, pages 2415–2422, 2011.

[3] B. Antic and B. Ommer. Robust multiple-instance learning with superbags. In *ACCV (2)*, pages 242–255, 2012.

[4] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, August 2011.

[5] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1931–1947, 2006.

[6] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, 2012.

[7] T. G. Dietterich and R. H. Lathrop. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1997.

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

[9] A. Eigenstetter and B. Ommer. Visual recognition using embedded feature selection for curvature self-similarity. In *NIPS*, pages 386–394, 2012.

[10] A. Gaidon, Z. Harchaoui, and C. Schmid. Actom Sequence Models for Efficient Action Detection. In *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, États-Unis, June 2011. MSR-INRIA.

[11] M. Hoai and F. D. la Torre. Max-margin early event detectors. In *CVPR*, pages 2863–2870, 2012.

[12] A. Kläser, M. Marszalek, C. Schmid, and A. Zisserman. Human focused action localization in video. In K. N. Kutulakos, editor, *ECCV Workshops (1)*, volume 6553 of *Lecture Notes in Computer Science*, pages 219–233. Springer, 2010.

[13] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *International Conference on Computer Vision (ICCV)*, 2011.

[14] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, Sept. 2005.

[15] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*, jun 2008.

[16] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 0:1–8, 2008.

[17] A. Monroy and B. Ommer. Beyond bounding-boxes: Learning object shape by model-driven grouping. In *ECCV (3)*, pages 580–593, 2012.

[18] J. C. Niebles, C. wei Chen, and L. Fei-fei. Modeling temporal structure of decomposable motion segments for activity classification. In *in Proc. 11th European Conf. Comput. Vision, 2010*, pages 392–405.

[19] B. Ommer, T. Mader, and J. M. Buhmann. Seeing the objects behind the dots: Recognition in videos from a moving camera. *International Journal of Computer Vision*, 83(1):57–71, 2009.

[20] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[21] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, page 127, Sept. 2009.

[22] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 650–663, Berlin, Heidelberg, 2008. Springer-Verlag.

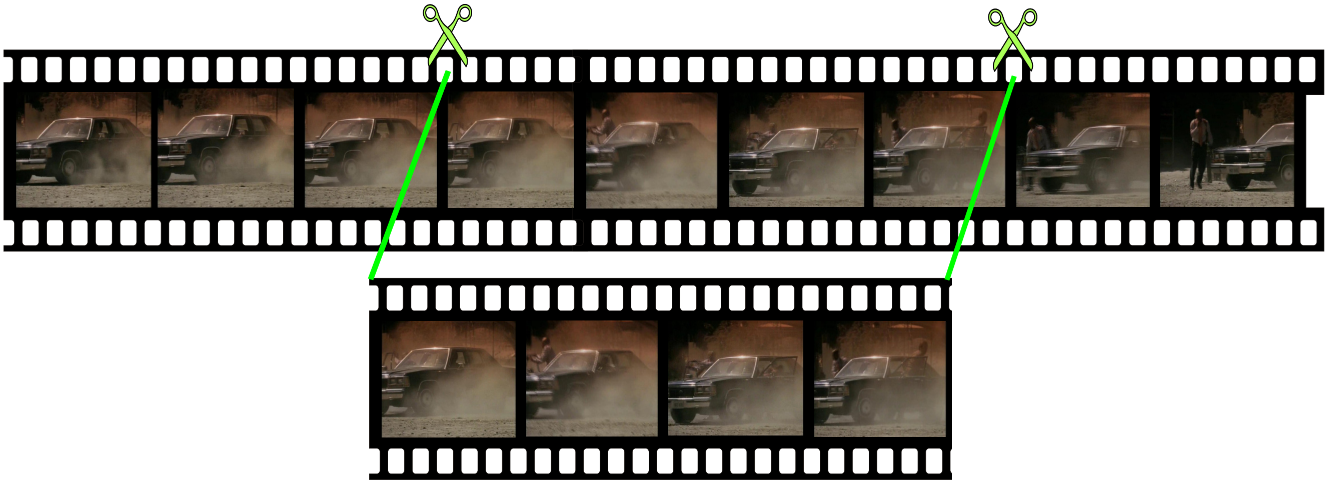

Figure 2: Detection of the ”GetOutOfCar” subsequence for the test video #234. Overlap of detected subsequence and the ground truth is 96.7%. The length of the sequence is 134 frames, and the subsequence is detected from 21th to 80th frame.

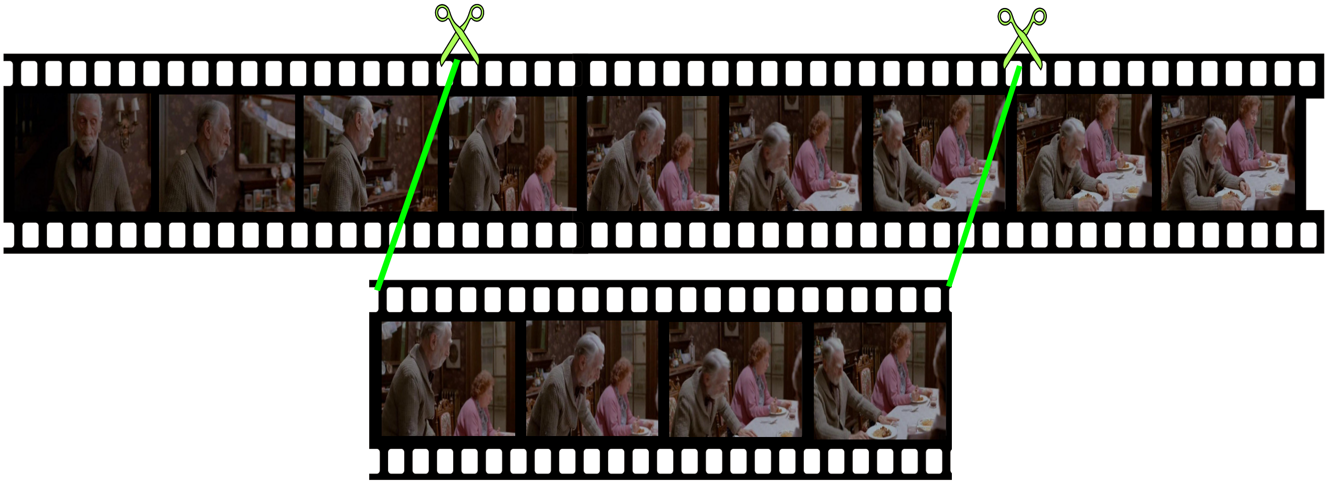

Figure 3: Detection of the ”SitDown” subsequence for the test video #682. Overlap of detected subsequence and the ground truth is to 94.9%. The length of the sequence is 169 frames, and the ”GetOutOfCar” subsequence is detected from 81th to 120th frame.

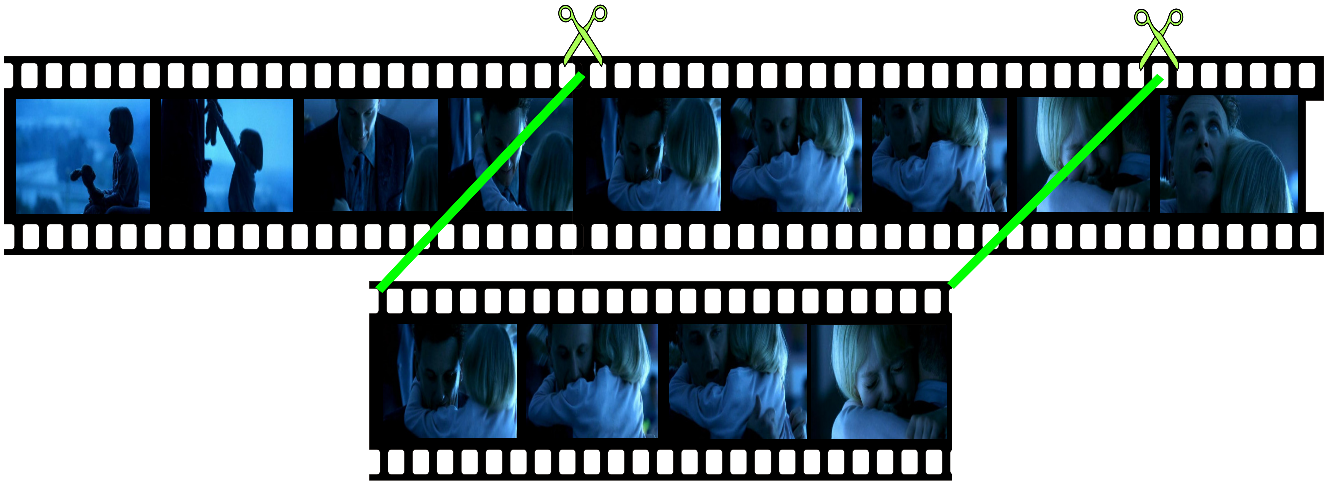

Figure 4: Wrong detection of the ”SitDown” subsequence in the test video #352. The length of the sequence is 438 frames, and the ground truth covers from 90th to 110th frame.