Will person detection help bag-of-features action recognition?

Alexander Klaser, Marcin Marszalek, Ivan Laptev, Cordelia Schmid

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Will person detection help bag-of-features action recognition?

Alexander Kläser — Marcin Marszałek — Ivan Laptev — Cordelia Schmid

N° 7373

Septembre 2010

Vision, Perception and Multimedia Understanding

ISSN 0249-6399 ISRN INRIA/RR--7373--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

Will person detection help bag-of-features action recognition?

Alexander Kläser*, Marcin Marszałek†, Ivan Laptev‡, Cordelia Schmid*

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

Rapport de recherche n° 7373 — Septembre 2010 — 17 pages

Abstract: Bag-of-feature (BoF) models currently achieve state-of-the-art performance for action recognition. While such models do not explicitly account for people in video, person localization combined with BoF is expected to give further improvement for action recognition. The purpose of this paper is to validate this assumption and to quantify the improvements in action recognition expected from current and future person detectors. Given locations of people in video, we find that—somewhat surprisingly—background suppression leads only to a limited gain in performance. This holds for actions in both simple and complex scenes. On the other hand, we show how spatial locations of people enable to incorporate strong geometrical constraints in BoF models and in this way to improve the accuracy of action recognition in some cases. Our conclusions are validated with extensive experiments on three datasets with varying complexity, basic KTH, realistic UCF Sports and challenging Hollywood.

Key-words: computer vision, action recognition, video, bag-of-features, human detection, tracking, classification, local descriptors

* INRIA Grenoble, LEAR, LJK – {klaser,schmid}@inrialpes.fr
† Visual Geometry Group, University of Oxford, – marcin@robots.ox.ac.uk
‡ INRIA / Ecole Normale Supérieure, Paris – ivan.laptev@inria.fr

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

La détection de personnes, peut-elle aider la reconnaissance d'actions

Résumé : Pas de résumé

Mots-clés : vision par ordinateur, reconnaissance d'actions, vidéo, sac-de-mots, détection de personnes, tracking, classification, descripteurs locals

Contents

1 Introduction

Human action recognition is an important component of video analysis with potential applications, for example, in video indexing, surveillance, gesture recognition, and analysis of sport events. As a consequence, many research efforts have been devoted to action recognition in the past.

Recently, bag-of-features representations [25] have demonstrated excellent performance for action recognition [11, 20, 23]. They allow to recognize a rich set of actions ranging from simple periodic motion (running, waving) to interactions (shaking hands, kissing), even in difficult, realistic conditions [9, 15, 19, 31]. Bag-of-features, however, have no explicit notion of objects or actors and provide a global video representation which is inherently sensitive to background clutter [32]. Furthermore, the lack of explicit object knowledge prevents geometric modeling which has been shown to increase performance [3, 17]

The purpose of this paper is to investigate whether and how the aforementioned deficiencies of the bag-of-features representation can be addressed by the tracking of human actors, cf. figure 1. Our contributions are two-fold. Firstly, we evaluate an "external" use of human tracks by treating the tracks as an approximate actor-background segmentation to suppress clutter (cf. middle part of figure 1). Secondly, we investigate an "internal" use of tracks by learning action models with stronger geometry (cf. right part of figure 1).

Precisely, the first contribution of this paper is the evaluation of the track based background suppression. Intuitively, narrowing down the attention to actors should benefit action recognition accuracy. The result is nevertheless worth quantifying, since in natural settings context might play an important role in recognition. If the reduction of background clutter gets outweighted by the loss of important context, the outcome might be counter-intuitive.

The second contribution of this work is the evaluation of the geometry in the action model. We frame a hypothesis which states that narrowing down

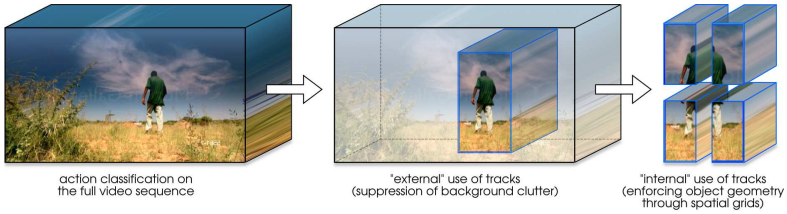

Figure 1: This paper analyzes the importance of human-centered attention in the context of bag-of-features based action recognition. Given a video sequence (left) we use tracks to suppress background (middle) and improve spatial modeling of human actions (right).

the attention to the actor will allow to enforce stronger geometry on the model, which in turn should result in better accuracy for action recognition. We propose to control the geometry level by varying the depth of spatial pyramids [17] and verify our hypothesis experimentally.

To obtain human tracks for the experiments mentioned above, we use off-the-shelf pedestrian and upper body detectors [3, 8] and combine detection into tracks according to [6]. We also use "ground-truth" tracks emulating an "ideal" detector. This allows us to make conclusions regarding desirable system designs, which might concern both current and future systems. Furthermore, we run our experiments on three datasets of varying complexity—basic KTH, realistic UCF and challenging Hollywood—in order to investigate how our conclusions might depend on the task.

The paper is structured as follows. Section 2 gives an overview of the related work. We present our recognition framework in detail in section 3. The experimental setup and results are discussed in section 5. We conclude the paper in section 6.

2 Related work

Several bag-of-features based methods for action recognition have been proposed in the past. Among the first, Schüldt et al. [23] proposed a method in terms of space-time interest points [14] and associated jet descriptors. Dollar et al. [4] proposed an alternative spatio-temporal interest point detector based on 1D Gabor filters. Willems et al. [30] proposed a detector based on the determinant of the space-time Hessian matrix. Junejo et al. [12] computed local features from temporal self-similarity matrices of actions. Kläser et al. [13] proposed an alternative local space-time descriptor based on space-time gradients. Other local approaches to action recognition include [9, 11, 24, 28, 31]. Whereas most of the published methods consider supervised classification, Niebles et al. [20] used bag-of-features for unsupervised learning of human actions.

Person-centered approaches to action recognition have been explored in the past. Temporal evolution of person silhouettes has been considered in [1, 2]. Efros et al. [5] used a person tracker to compute an actor-centric motion descriptor using smoothed optical flow. Similarly, Schindler and van Gool [22] used optical flow and Gabor filter responses computed at person locations. Fathi and

Mori [7] proposed a two layer AdaBoost scheme for learning actions using optical flow features inside person tracks. Recently, Wang and Mori [28] proposed a discriminative model with latent variables representing actions as constellations of parts. Although these papers are related to ours, they do not exploit bag-of-features models nor consider action recognition in realistic settings as we do in this paper.

Recent papers report high recognition accuracies for actions in controlled settings such as in Weizman [1] and KTH [23] datasets. At the same time, action recognition remains a very challenging problem in realistic settings of TV broadcasts, movies or surveillance videos as demonstrated in [15, 26]. Several recent works address action recognition in realistic settings. Laptev and Pérez [16] addressed action recognition and localization in movies with an AdaBoost classifier. Rodriguez et al. [21] used Maximum Average Correlation Height (MACH) filters to recognize actions in movies as well as in TV sports videos. Laptev et al. [15] used bag-of-features models extended with space-time grids for action recognition in movies. Marszalek et al. [19] demonstrated the advantage of contextual cues for action recognition in realistic settings.

Similar to these works, we address action recognition in realistic settings. For this, we extend the bag-of-features approach to action recognition and integrate the location of people into the action model. We evaluate our model in extensive experiments and draw conclusions from three datasets with an increasing degree of difficulty.

3 Recognition framework

The bag-of-features approach represents video sequences as occurrence histograms of local features on which classification is carried out. In the following, we give details of our implementation in this work. The sampling procedure and the employed spatio-temporal descriptor are presented in subsections 3.1 and 3.2, respectively. Details on how we gradually incorporate geometry in the bag-of-features representation are given in subsection 3.3. Finally, subsection 3.4 discusses the classification part of our framework.

3.1 Feature sampling

For our representation, we use dense local features obtained by sampling the video into space-time patches along image dimensions (x, y), along time (t) and for different spatio-temporal scale values (σ, τ). We employ a spatial stride of 12×12 (for UCF and Hollywood) as well as 6×6 pixels (for KTH due to its smaller resolution) and a temporal stride of 3 frames throughout all our experiments. Features are computed such that half of their volume overlaps with a neighboring feature. For spatial and temporal scales, stride widths are scaled accordingly. Neighboring scale levels differ by a factor of $\sqrt{2}$. Sampling is performed for all combinations of different spatial and temporal scale levels.

We construct tracks —spatio-temporal "corridors" that encompass actors in the video volume—from a set of bounding-boxes connected in time. In this work, bounding boxes are obtained either automatically using off-the-shelf pedestrian and upper-body detectors [3, 8] or they are provided as ground-truth. In order to obtain features on the foreground (actors and their closest vicinity), we reuse

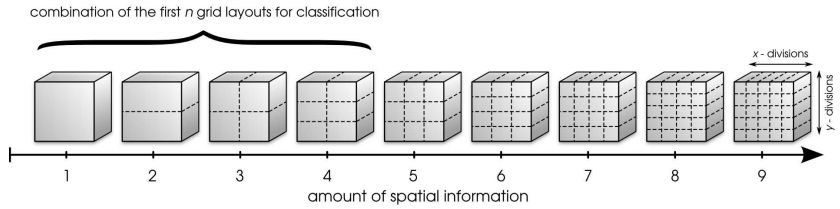

Figure 2: Geometric information is encoded through spatial grids. We use a sequence of grids of increasing density to control the amount of spatial information (geometric constraints). We combine the first n grid layouts, in this example $n = 4$.

features from the full videos and keep only those that fall into the bounding box of a human track. This allows a fair comparison since the same features are used and no information is added.

3.2 Feature descriptor

For our experiments, we employ the spatio-temporal HoG descriptor proposed by Kläser et al. [13]. This descriptor is an extension of the popular SIFT descriptor [18] to video sequences. It is based on histograms of 3D gradient orientations. Regular polyhedrons are used to uniformly quantize the orientation of spatio-temporal gradients. In this way, the descriptor combines shape and motion data at the same time.

3.3 Spatial bags-of-features

Vocabularies are constructed by randomly sampling 4000 training features. In our experiments, this codebook size has shown empirically good results. Its performance was comparable to the one of codebooks constructed using k-means. However, the computation of random codebooks is much faster. All experiments are repeated three times, each time with a different codebook. This allows us to estimate mean and standard deviation. Features are assigned to their closest vocabulary word using Euclidean distance.

To encode spatial information within the bag-of-features representation, we use spatial grids [15, 17], see figure 2. The video sequence is split into (spatial) subsequences, and a histogram is computed for each subsequence. The final histogram is obtained by concatenating histograms of all cells in the grid. We compute grids over the whole video as well as within tracks. In the latter case the grid position of a feature is defined relatively to the position of the track's bounding box at the corresponding time instant. For multiple tracks that are overlapping, features can vote multiple times into the final histogram, i.e., once for each track.

In our experiments we wish to quantify the "amount" of geometric information we use for action recognition. For this, the first n of the following grid layouts are combined: 1x1, 2x1, 2x2, 3x2, 3x3, 4x3, 4x4, 5x4, 5x5 (cf. figure 2). A larger n will consequently enforce more geometry.

3.4 Classification

For classification, we use non-linear support vector machines with a multi-channel Gaussian kernel [32]

$$K(H_i, H_j) = \exp\left(-\sum_{c \in \mathcal{C}} \frac{1}{A_c} D_c\left(H_i, H_j\right) \right) \quad , \tag{1}$$

where $H_i = \{h_{ik}\}$ and $H_j = \{h_{jk}\}$ are the histograms for channel c and $D_c(H_i, H_j)$ is the χ^2-distance given as

$$D_c(H_i, H_j) = \frac{1}{2} \sum_{k=1}^{V} \frac{(h_{ik} - h_{jk})^2}{h_{ik} + h_{jk}} \quad . \tag{2}$$

V is the vocabulary size and A_c is the average distance between all training samples for channel c. For multi-class classification, we use the *one-against-rest* approach.

4 Datasets

We carry out experiments on three different action datasets. The **KTH actions** dataset [23] consists of six human action classes: walking, jogging, running, boxing, waving, and clapping (cf. figure 3, top). Each action class is performed several times by 25 subjects. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The background is homogeneous and static in most sequences. In total the data consists of 2391 video samples. We follow the original experimental setup of the authors, i.e., divide the samples into training/validation set (8+8 people) and test set (9 people). Evaluation on this dataset is done via multi-class classification. We report the performance as average accuracy over all classes as proposed by the authors of the dataset. Since only one person is visible per sequence, we obtain tracks by detecting upright humans [3] in all frames and by applying a simple outlier removal strategy along with temporal smoothing. Results are shown in the top row of figure 4.

The **UCF sport actions** dataset [21] contains ten different types of human actions: swinging (on the pommel horse and on the floor), diving, kicking (a ball), weight-lifting, horse-riding, running, skateboarding, swinging (at the high bar), golf swinging and walking (cf. figure 3, middle). The dataset consists of 150 video samples which show a large intra-class variability. To increase the amount of data samples, we extend the dataset by adding a horizontally flipped version of each sequence to the dataset. As proposed by the authors, we use a leave-one-out setup (each sequence and its flipped version are left out) and report the average accuracy (over all samples) for a multi-class classifier. Ground truth tracks of the person performing an action are provided with the dataset (cf. figure 5, top). Unlike KTH, the UCF dataset often involves several people in the scene. In order to obtain automatic tracks, we run the same pedestrian detector [3] as for KTH and link detections into tracks using agglomerative clustering as in [6]. We exploit temporal consistency to improve detection results by (i) removing short tracks (ii) filling in missing detections within tracks and (iii) applying

Figure 3: Sample frames from video sequences of KTH (top), UCF Sports (middle), and Hollywood (bottom) human action datasets.

temporal smoothing of detections. UCF sequences contain high variation of the background and highly articulated human poses, which results in a decreased precision and recall of human detection. Example detections are shown in the middle row of figure 4.

The **Hollywood actions** dataset [15] consists of 32 different Hollywood movies; we use the standard setup available online. Action classes contain: answering the phone, getting out of the car, hand shaking, hugging, kissing, sitting down, sitting up, and standing up (see figure 3, bottom). In total, there are 430 action samples divided into a training set (219 sequences) and a test set (211 sequences). Train and test sequences are obtained from different movies. In our experiments, we only make use of the clean training dataset. The performance is evaluated as in the standard setup by the average precision (AP) for each of the action classes. In this type of video data, humans are in general visible only with their upper body. Therefore, we manually annotate as ground truth upper bodies tracks (cf. figure 5, bottom). Training is performed using all tracks with humans performing a given action, and for testing sequences of all visible humans are annotated, mimicking a perfect human detector. Automatic tracks (cf. figure 4, bottom) are obtained with the same detector [3] as for KTH and UCF, but trained for upper bodies as in [8]. We also use the same temporal association [6] as for UCF.

5 Experimental results

Our goal is to quantify the improvement in human action recognition when extending the bag-of-features representation with knowledge about actor localization. For this, we compare the performance of our baseline bag-of-features system (cf. section 3) with the same system, but with background features removed based on human tracks. The recognition accuracy is given as a function of the geometry level. We compare results for the bag-of-features baseline (red

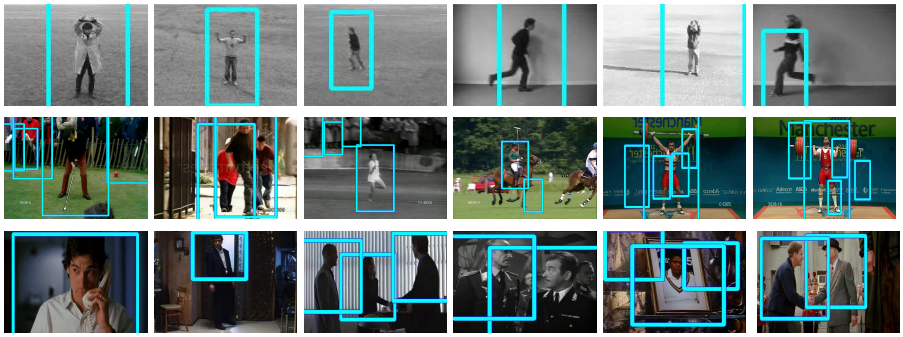

Figure 4: Examples of automatic tracks on KTH (top) UCF Sports (middle) and Hollywood (bottom) action datasets. Note how complex is the UCF and Hollywood video material and how challanging it is to robustly detect actors is such video.

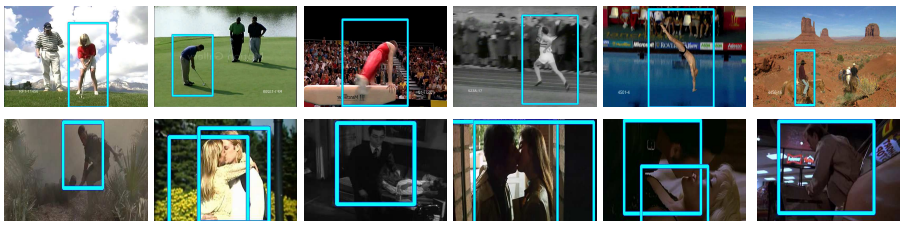

Figure 5: Ground-truth tracks for UCF (top) and Hollywood (bottom) datasets.

squares) and each of the track types used to select features (blue triangles for tracks automatically obtained from person detections [3, 8]; green circles for ground-truth tracks)—see figures 6–8. For each of those figures, we draw two types of observations. First, we evaluate the gain due to background suppression by comparing the performance of the orderless representation (only one "grid" level—leftmost measurement on each plot, highlighted). Second, we assess the gain due to stronger geometry. Results are reported separately for each dataset: KTH actions (subsection 5.1), UCF sports (subsection 5.2), and Hollywood1 (subsection 5.3).

5.1 KTH actions

Results for the KTH dataset are plotted in figure 6. Comparing the values for orderless BoF (highlighted measurements in the leftmost column of the plot) allows to estimate the gain in recognition accuracy due to background suppression. For the KTH dataset the reduction of background clutter using automatically detected human tracks leads to a small accuracy gain of about 0.5%.

A more significant improvement of over 2% is possible by increasing the number of grids and encoding more layout information. Note, however, that this only holds for the features obtained using tracks, not for the full video where results degrade; the difference between the tracks and the baseline reaches

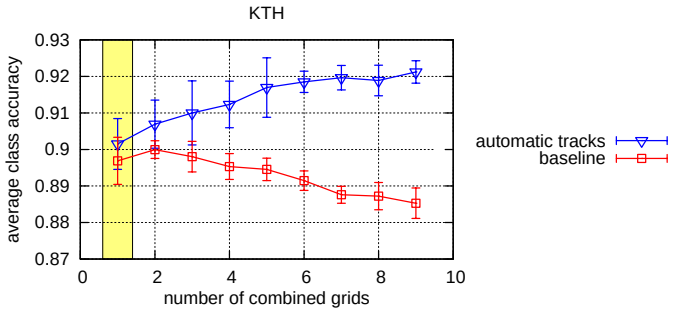

Figure 6: Performance plots for the KTH actions dataset. Bars indicate standard deviation from the mean.

		Predicted class					
		boxing	clapping	waving	walking	jogging	running
True class	boxing	98.4	0.0	0.0	1.6	0.0	0.0
	clapping	3.9	96.1	0.0	0.0	0.0	0.0
	waving	0.2	4.4	95.4	0.0	0.0	0.0
	walking	0.0	0.0	0.0	96.5	3.5	0.0
	jogging	0.0	0.0	0.0	3.0	93.3	3.7
	running	0.0	0.0	0.0	0.0	18.5	81.5

Table 1: Confusion matrix for the KTH dataset. Classification was performed using our full system, i.e., features from detected actors and combining all 9 grid layouts. Note the confusion between running and jogging.

almost 4% for the full combination. This demonstrates that layout information can help to learn a better action model if tracks are used.

The confusion matrix in table 1 shows that the main source of confusion is an inherent overlap between jogging and running. Looking at examples of these classes, we have observed that there is no visual difference between some sequences of the two classes.

The currently best result on this dataset has been reported for the hierarchical data mining approach by Gilbert et a [9] which achieved 94.5%. Han et al. [10] obtained 94.1% accuracy with a multi-kernel classifier. Among the results that have been reported with a pure BoF representation, the combination of Harris3D interest points together with HOF (92.1%) as well as HOG-HOF (91.8%) gave highest results [27] in the literature.

Our average accuracy over three runs (for our full method, i.e., using automatic detections to suppress background and combining all 9 grid layouts) is 92.1%. In general, our results are situated among the state-of-the-art results. However, our method is not optimized for high performance, yet rather for a fair comparison with the baseline. We showed that performance on KTH can be improved significantly using layout information on the tracks. Therefore our approach shows the potential to improve the performance of other methods, as well.

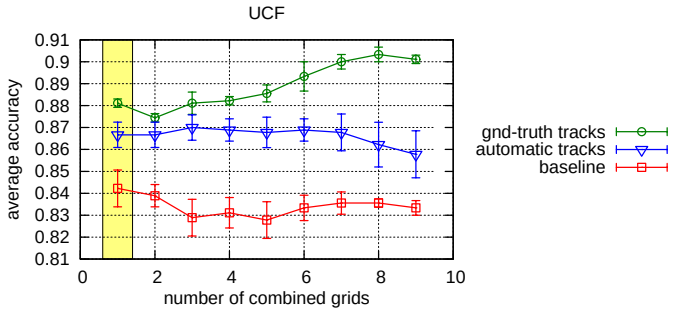

Figure 7: Performance plots for the UCF sport actions dataset. Bars indicate standard deviation from the mean.

5.2 UCF Sports

Experimental results for the UCF dataset are presented in figure 7. If we compare the results for orderless BoF (highlighted measurements on the left of the plot), we clearly see a gain due to suppressing background features and narrowing down attention. The recognition accuracy improves significantly by 4% with "ideal" tracks provided as ground-truth. The off-the-shelf pedestrian detector is also able to out-perform the baseline by over 2%.

Further interesting conclusions can be drawn from the evaluation of layout information. Enforcing stronger layout models can degrade the performance of the baseline and also of automatic tracks. For the baseline, the degradation of its results is permanent, while for the automatic case we can observe only a minor improvement up to three grid combinations. An ideal detector and tracker, however, allows to significantly and consistently improve the recognition accuracy when more layout information is included. This shows the importance of a good human tracker in order to fully exploit the knowledge about actor localization.

It is also interesting to look at the confusion matrices for this dataset. Table 2 compares the matrices obtained for the baseline with an orderless bag model (top) and by using the ground truth actor annotations and enforcing a stronger layout model (bottom). In the first case, note the general confusion for actions such as riding and weight lifting with other classes. This confusion is significantly reduced in the second case for most classes. Nevertheless, some confusion remains using tracks—the accuracy for running even dropped. This is presumably due to the reduced amount of context information, such as strong camera ego-motion during running. Other actions that remain confused are skateboarding and walking. This is explainable given their visual similarity.

Works that published results on the UCF sports dataset are Rodriguez et al. [21] who also published the dataset and Wang et al. [27]. Rodriguez et al. reported an accuracy of 69.2% with a template matching approach, and Wang et al. obtained 85.6% in a BoF setup close to ours. In an "ideal" setup (i.e., with ground truth tracks), our system achieves 90.1% average accuracy (combining all 9 grid layouts) which is significantly higher than the current state-of-the-art. For the automatic case with human detections, we obtain with our features 86.7% by only considering foreground.

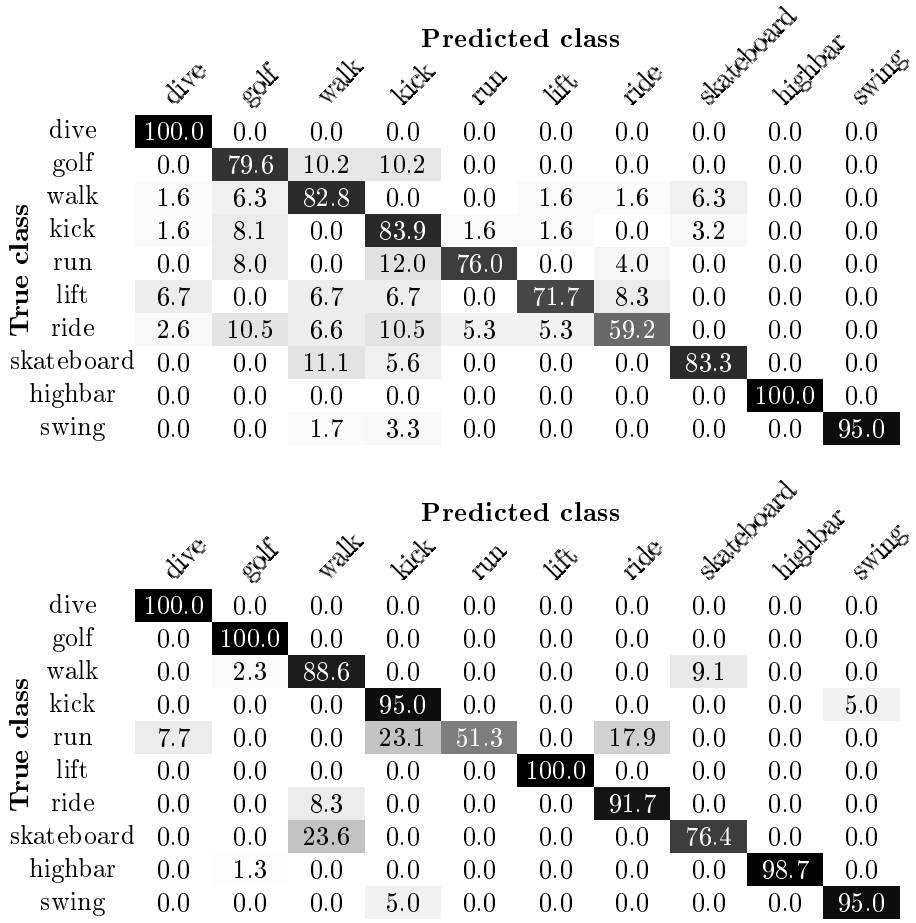

Table 2: Confusion matrices for (top) the UCF sports dataset using orderless features on the full video and (bottom) using (ground truth) actor annotation and spatial grids (all combinations). Note how the stronger layout model pruned the worst confusions.

5.3 Hollywood actions

Experimental results for the Hollywood dataset, the most challenging dataset in our setup, are given in figures 8 and 9. Since the classification task for this dataset consists of multiple binary tasks, we also show results for each class individually (figure 8). One immediately notices that (unlike for the previous datasets) the results degrade significantly when using automatic tracks. This is largely due to dynamic camera, clutter and occlusion, which make human detection in Hollywood videos difficult. For instance, people getting out of car are typically not visible at the beginning of the action and are often occluded by the door of the car throughout the action. Additional occlusion and non-upright poses render the detection of people difficult, as well, cf. figure 5. Furthermore, even a perfect detector is not guaranteed to improve recognition accuracy. This is most likely due to the fact that this dataset better reflects natural conditions where context can play an important role for action recognition, e.g., for actions such as getting out of a car or kissing. Hollywood actions include interactions between different humans and interactions with objects that might also be harder to interpret without context information [19]. Overall, one can observe a significant gain for the classes HugPerson, StandUp and SitUp. For the classes AnswerPhone and SitDown we can note a slight improvement. However, the performace decreases for Kiss and GetOutCar, most likely due to the context information playing an important role for these action classes.

Since track information is not useful for all types of actions, we combine both representations—baseline and track-based. We employ a simple selector choosing the best representation for a particular action in an automatic manner. During training, the representation that performs best on the training set (evaluated via cross-validation) is selected. Figure 9 shows the average AP gain in such setup. The result is consistent with those for other datasets: the improvement due to background suppression is relatively small, while enforcing stronger layout information is beneficial.

For the Hollywood dataset, our baseline (a single orderless channel) obtains 31.3% mean AP and outperforms the corresponding orderless HoG (27.0%) and HoF (21.5%) channels of [15]. It is also close to the performance of their best channel (32.2%). With an "ideal" detector in combination with the BoF on the full video, we improve up to 36.4% with a single feature type. Laptev et al. proposed a method to learn combinations of different features which they showed to lead to a higher average precision of up to 38.4% on this dataset. However, combining different feature types is beyond the scope of this work.

Similar to KTH, Gilbert et al. [9] (53.5%) and Han et al. [10] (47.5%) have reported overall the highest results in the literature. Note that Han et al. obtained with their best channel alone 33.3% which is comparable to our results. Compared to existing, standard BoF approaches, best results have been reported by Willems et al. [29] (29.6%) by using a Hessian feature detector along with a variant of HOG3D.

Our results compare favorably to the state-of-the-art with only single feature types. As stated before, employing human localization offers cues for action recognition that are complementary to existing approaches, e.g., feature combination [19, 10]. In a combined setup, it can therefore further improve existing state-of-the-art methods.

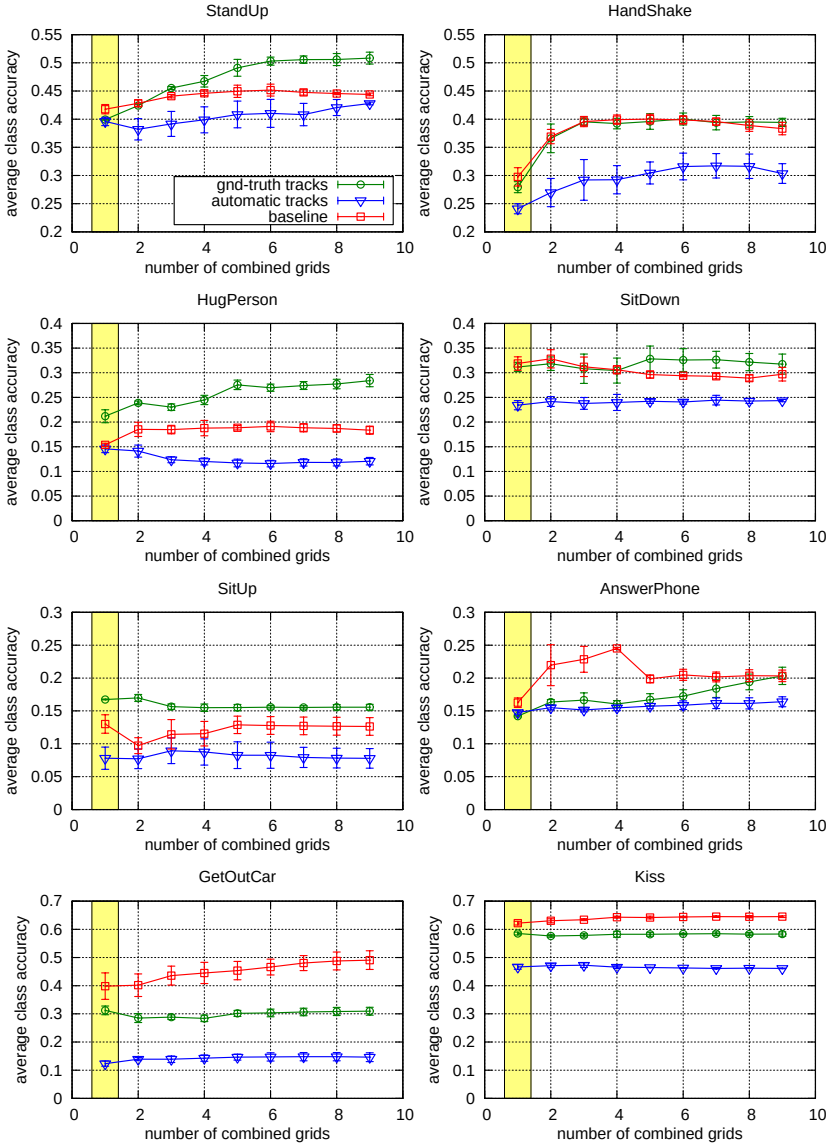

Figure 8: Per class results on Hollywood. Note that a performance improvement using human tracks is dependent on the action class. A significant gain can be observed for the classes HugPerson, StandUp and SitUp. The performance decreases for Kiss and GetOutCar, most likely due to the context information playing an important role for these action classes.

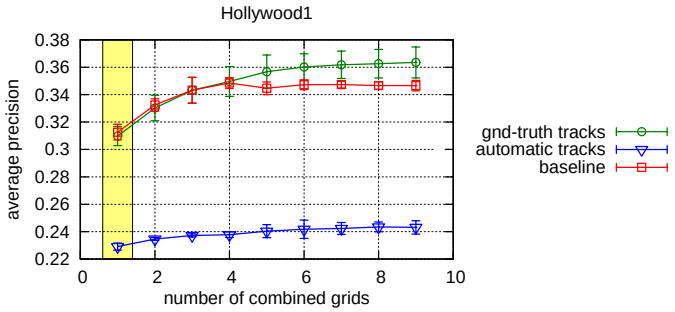

Figure 9: Performance plots for the Hollywood actions dataset. Performance for ground-truth tracks is a learned combination of ground-truth tracks and the BoF baseline. Bars indicate standard deviation from the mean.

6 Summary

In this chapter, we have shown that action recognition can benefit from human localizations in videos. Quite surprisingly, it turns out that this gain is not due to suppressing background clutter. Only in the case of simple scenarios, background suppression helps to improve classification results. However, for realistic settings, removing background can lead to removal of valuable context. Therefore background suppression resulted in general in only minor recognition accuracy improvement. In the case of a few action classes (getting out of a car, kissing) we observed even a performance degradation.

Furthermore, we have proposed to use human tracks to improve action modeling. We have redefined a popular spatial pyramid concept as a model with controlled levels of spatial constraints. We have shown that narrowing down the attention to human actors allows to incorporate more layout information into the learned model. In general, this positively benefited recognition accuracy. However, on realistic videos and for some action classes, we observed no or only minor improvement.

Finally, our work is showing the need for better human detectors and trackers for action recognition. We have improved over the state-of-the-art results by combining off-the-shelf techniques, but we have also shown an even greater potential is to be obtained with more accurate human detectors.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005.

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

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

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

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.

[6] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy – automatic naming of characters in TV video. In *BMVC*, 2006.

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

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

[9] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*, 2009.

[10] D. Han, L. Bo, and C. Sminchisescu. Selection and context for action recognition. In *ICCV*, 2009.

[11] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007.

[12] I. Junejo, E. Dexter, I. Laptev, and P. Pérez. Cross-view action recognition from temporal self-similarities. In *ECCV*, 2008.

[13] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.

[14] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

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

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

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[19] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[20] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[21] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[22] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *CVPR*, 2008.

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

[24] E. Shechtman and M. Irani. Space-time behavior based correlation. In *CVPR*, 2005.

[25] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[26] Trecvid evaluation for surveillance event detection, National Institute of Standards and Technology (NIST), 2008. http://www-nlpir.nist.gov/projects/trecvid.

[27] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[28] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *CVPR*, 2009.

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

[30] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008.

[31] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009.

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

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

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

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399