Fine-grained Activity Recognition with Holistic and Pose based Features

Leonid Pishchulin[1], Mykhaylo Andriluka[1,2], and Bernt Schiele[1]

[1] Max Planck Institute for Informatics, Germany
[2] Stanford University, USA

Abstract. Holistic methods based on dense trajectories [29, 30] are currently the de facto standard for recognition of human activities in video. Whether holistic representations will sustain or will be superseded by higher level video encoding in terms of body pose and motion is the subject of an ongoing debate [12]. In this paper we aim to clarify the underlying factors responsible for good performance of holistic and pose-based representations. To that end we build on our recent dataset [2] leveraging the existing taxonomy of human activities. This dataset includes 24, 920 video snippets covering 410 human activities in total. Our analysis reveals that holistic and pose-based methods are highly complementary, and their performance varies significantly depending on the activity. We find that holistic methods are mostly affected by the number and speed of trajectories, whereas pose-based methods are mostly influenced by viewpoint of the person. We observe striking performance differences across activities: for certain activities results with pose-based features are more than twice as accurate compared to holistic features, and vice versa. The best performing approach in our comparison is based on the combination of holistic and pose-based approaches, which again underlines their complementarity.

1 Introduction

In this paper we consider the task of human activity recognition in realistic videos such as feature movies or videos from YouTube. We specifically focus on how to represent activities for the purpose of recognition. Various representations were proposed in the literature, ranging from low level encoding using point trajectories [29, 30] to higher level representations using body pose trajectories [12] and collection of action detectors [23]. At the high level human activities can often be accurately characterized in terms of body pose, motion, and interaction with scene objects. Representations based on such high level attributes are appealing as they allow to abstract the recognition process from nuisances such as camera viewpoint or person clothing, and facilitate sharing of training data across activities. However, articulated pose estimation is a challenging and non-trivial task that is subject of ongoing research [18, 19, 32, 8, 24]. Therefore, state-of-the-art methods in activity recognition rely on holistic representations [15, 9, 29, 30] that extract appearance and motion features from the entire video and leverage discriminative learning techniques to identify information relevant for the task.

Recent results on the JHMDB dataset [12] suggest that state-of-the-art pose estimation methods might have reached sufficient accuracy to be effective for activity recognition. Motivated by these results, we further explore holistic and pose based representations aiming for much broader scale and coverage of activity classes. To that end we employ our recent "MPI Human Pose" dataset [2]. Compared to 21 activity classes considered in [12] the "MPI Human Pose" dataset includes 410 activities and more than an order of magnitude more images ($\sim$ 32K in JHMDB vs. over 1M images in "MPI Human Pose"). "MPI Human Pose" aims to systematically cover a range of activities using an existing taxonomy [1]. This is in contrast to existing datasets [13, 26] that typically include ad-hoc selections of activity classes. Using the rich labelling of people provided with "MPI Human Pose" we evaluate the robustness of holistic and pose based representations to factors such as body pose, viewpoint, and body-part occlusion, as well as to the number and speed of dense trajectories covering the person.

This paper makes the following contributions. First, we perform a large-scale comparison of holistic and pose based features on the "MPI Human Pose" dataset. Our results complement the findings in [12], indicating that pose based features indeed outperform holistic features for certain cases. However we also find that both types of features are complementary and their combination performs best. Second, we analyze factors responsible for success and failure, including number and speed of trajectories, occlusion, viewpoint and pose complexity.

Related work. There is a large body of work on human activity recognition and its review is out of the scope of this paper. Instead we point to the respective surveys [28, 4] and concentrate on the methods most relevant for this work.

Holistic appearance based features combined with the Bag-of-Words representation [15, 9, 29, 30, 12] are considered the de facto standard for human activity recognition in video. Many methods create discriminative feature representations of a video by first detecting spatiotemporal interest points [5, 14] or sampling them densely [31] and then extracting various feature descriptors in the space-time volume. Most commonly used feature descriptors are histograms of oriented gradients (HOG) [6], histograms of flow (HOF) [7] or Harris 3D interest points [14]. In this paper we examine the recent holistic approach [29, 30] which tracks dense feature points and extracts strong appearance based features along the trajectories. This method achieves state-of-the-art results on several datasets. Other holistic approaches include template based [20] or graph based methods constructing graphs from a spatiotemporal segmentation of the video [3].

Another line of research explores ways of higher level video encoding in terms of body pose and motion [11, 25, 21, 12]. The intuition there is that many activities exhibit characteristic body motions and thus can reliably be described using human body pose based features. Pose based activity recognition was shown to work particularly well in images with little clutter and fully visible people [25]. However, in more challenging scenarios with frequent occlusions, truncations and complex poses, body features significantly under-perform holistic appearance based representations [21]. Recently, it was shown that body features extracted from detected joint positions outperform holistic methods, and their combination did not improve over using body features only [12]. However, these conclusions were made on a subset of the HMDB dataset [13], where actions with global body motion are performed by isolated and fully visible individuals – a setting that seems well suited for pose estimation methods. In this work we examine a wide range of underlying factors responsible for good performance of body based and holistic methods. In contrast to [12] our analysis on hundreds of activity classes reveals that holistic and pose based methods are highly complementary, and their performance varies significantly depending on the activity.

We build our analysis on our recent "MPI Human Pose" dataset collected by leveraging an existing taxonomy of every day human activities and thus aiming for a fair coverage. This is in contrast to existing activity recognition datasets [13, 26, 17, 16, 20] that typically include ad-hoc selections of activity classes. A large number of activity classes (410) and more than an order of magnitude more images compared to [12] facilitate less biased evaluations and conclusions.

2 Dataset

In order to analyze the challenges for fine-grained human activity recognition, we build on our recent publicly available "MPI Human Pose" dataset [2]. The dataset was collected from YouTube videos using an established two-level hierarchy of over 800 every day human activities. The activities at the first level of the hierarchy correspond to thematic categories, such as "Home repair", "Occupation", "Music playing", etc., while the activities at the second level correspond to individual activities, e.g. "Painting inside the house", "Hairstylist" and "Playing woodwind". In total the dataset contains 20 categories and 410 individual activities covering a wider variety of activities than other datasets, while its systematic data collection aims for a fair activity coverage. Overall the dataset contains $24,920$ video snippets and each snippet is at least 41 frames long. Altogether the dataset contains over a 1M frames. Each video snippet has a key frame containing at least one person with a sufficient portion of the body visible and annotated body joints. There are $40,522$ annotated people in total. In addition, for a subset of key frames richer labels are available, including full 3D torso and head orientation and occlusion labels for joints and body parts.

Static pose estimation complexity measures. In addition to the dataset, in [2] a set of quantitative complexity measures aiming to asses the difficulty of pose estimation in each particular image was proposed. These measures map body image annotations to a real value that relates the complexity of each image w.r.t. each factor. These complexity measures are listed below.

1. *Pose*: deviation from the mean pose on the entire dataset.
2. *Occlusion*: number of occluded body parts.
3. *Viewpoint*: deviation of 3D torso rotation from the frontal viewpoint.
4. *Part length*: deviation of body part lengths from the mean part lengths.
5. *Truncation*: number of truncated body parts.

Novel motion specific complexity measures. We augment the above set with the measures assessing the amount of motion present in the scene.

1. *# dense trajectories (# DT)*: total number of DT computed by [30].
2. *# dense trajectories on body (# DT body)*: number of DT on body mask.
3. *Motion speed (MS)*: mean over all trajectory displacements in the video.
4. *Motion speed on body (MS body)*: *MS* extracted on body mask.
5. *# people*: number of people in the video.

3 Methods

In order to analyze the performance on the challenging task of fine-grained human activity recognition, we explore two lines of methods that extract relevant features. The first line of methods extracts holistic appearance based features and is represented by the recent "Dense Trajectories" method [29] which achieves state-of-the-art performance on several datasets. The second line of methods computes features from locations of human body joints following the intuition that body part configurations and motion should provide strong cues for activity recognition. We now describe both types of methods and their combinations.

3.1 Dense trajectories (DT)

DT computes histograms of oriented gradients (HOG) [6], histograms of flow (HOF) [15], and motion boundary histograms (MBH) [7] around densely sampled points that are tracked for 15 frames using median filtering in a dense optical flow field. In addition, x and y displacements in a trajectory are used as a fourth feature. We use a publicly available implementation of the improved DT method [30], where additional estimation removes some of the trajectories consistent with camera motion. Following [30] we extract all features on our data and generate a codebook for each of the four features of 4K words using k-means from a million of sampled features, and stack L_2-normalized histograms for learning.

3.2 Pose-based methods

It has recently been shown that body features provide a strong signal for recognition of human activities on a rather limited set of 21 distinctive full body actions in monocular rgb video sequences [12]. We thus investigate the usefulness of body features for our task where the variability of poses and granularity of activities is much higher. We explore different ways of obtaining body joint locations and extract the body features using the code kindly provided by [12]. We use the same trajectory length of 7 frames with a step size of 3, generate a codebook of 20 words for each descriptor type and finally stack the L_2-normalized histograms for learning. We now present different ways of obtaining body joint locations.

GT single pose (GT). We directly use the ground truth locations of body joints in the key frame to compute single pose based features. As some of the body parts may be truncated, we compute features only for the present parts.

GT single pose + track (GT-T). As the ground truth information is not available for the rest of the frames in a sequence, we approximate the positions

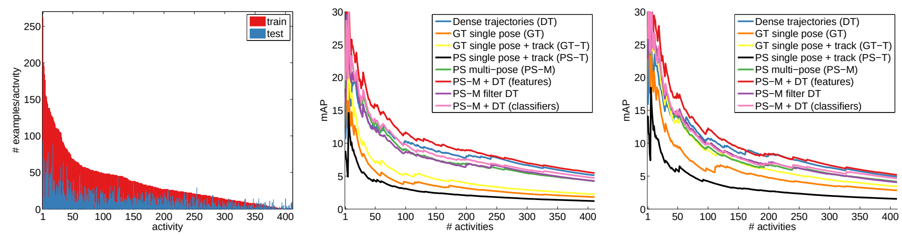
of body joints in the neighboring frames by tracking the joints using sift-based tracker [21]. The tracker is initialized with correct positions of body joints, and thus provides reliable tracks of joints in the local temporal neighborhood.
PS single pose + track (PS-T). It is not realistic to expect the ground truth information to be available at test time in real world scenarios. We thus replace the given body joint locations by automatically estimated ones using the publicly available implementation [32]. This efficient method is based on pictorial structures (PS) and obtained good performance on the "MPI Human Pose" [2].
PS multi-pose (PS-M). Using the method [32] also allows to obtain joint locations independently in each frame of a sequence without using the sift tracker. Notably, the same method was shown by [12] to outperform the holistic approach.

3.3 Combinations of holistic and body based methods

As the holistic *DT* approach does not extract any pose information, and pose based methods do not compute any appearance features, both are potentially complementary. Thus we expect that an activity recognition system will profit from their combinations. We investigate two ways of combining the methods.
PS-M + DT (features). We perform a *feature* level fusion of both *DT* and *PS-M* by matching both types of features independently to the respective codebooks and then stacking the normalized histograms into a single representation.
PS-M + DT (classifiers). We also investigate a *classifier* level fusion. To do so we first run pre-trained *DT* and *PS-M* classifiers (see Sec. 4) independently on each sequence and stack the scores together into a single feature vector.
PS-M filter DT. Another way of combining both types of methods is using estimated joint locations to filter the trajectories computed by *DT*. We first estimate poses in all video frames and generate a binary mask using the union of rectangles around detected body parts for all single top detections per frame. We then only preserve the trajectories overlapping with the mask in all frames.

4 Analysis of activity recognition performance

In this section we analyze the performance of holistic and pose based methods and their combinations on the challenging task of fine-grained human activity recognition with hundreds of activity classes.
Data splits. As main test bed for our analysis, a split with videos containing sufficiently separated individuals [2] is used. This restriction is necessarily for using the pose estimation method [32]. This *Separate people* split contains 411 activities with 15244 training and 5699 testing video snippets. Fig. 1(a) shows statistics of the training and testing videos per activity. Notably, the videos may still contain multiple people and some body parts may be truncated by a frame border. To rule out the confusion potentially caused by the presence of multiple truncated people, we define a subset of the test set from *Separate people*. This subset contains 2622 videos with exactly one fully visible person per video. This



(a) # examples/activity on *Separate people* (b) Performance (mAP) on *Separate people* (c) Performance (mAP) on *Single fully visible people*

Fig. 1. Dataset statistics and performance (mAP) as a function training set size. Shown are (a) number of training/testing examples/activity in *Separate people* subset; performance on (b) *Separate people* and (c) *Single fully visible people*. Best viewed in color.

Single fully visible people setup is inspired by [12] and is favorable for the pose estimation method [32] designed to predict body joints of fully visible people.

Training and evaluation. We train activity classifiers using feature representations described in Sec. 3 and ground truth activity labels. In particular, we train one-vs-all SVMs using mean stochastic gradient descent (SGD) [22] with a χ^2 kernel approximation [27]. At test time we perform one-vs-all prediction per each class independently and report the results using mean Average Precision (AP) [10]. When evaluating on a subset, we always report the results on the top N activity classes arranged w.r.t training set sizes.

4.1 Overall performance

We start the evaluation by analyzing the performance on all activity classes.

Separate people. It can be seen from Fig. 1(b) that performance is reasonable for a relatively small number of classes (the typical case for many activity recognition datasets), but quickly degrades for a large number of classes, clearly leaving room for improvement of activity recognition methods.

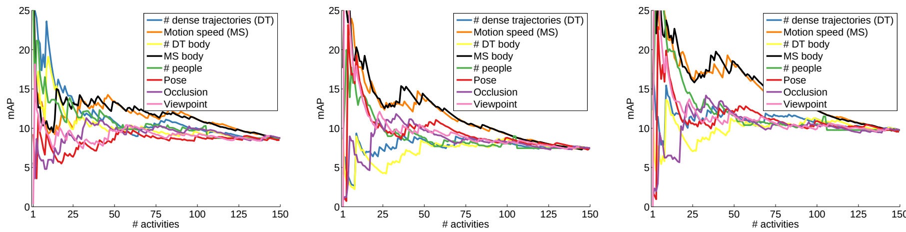
We observe that *Dense trajectories (DT)* alone outperforms all pose based methods achieving 5.1% mAP. Expectedly, *GT single pose (GT)* performs worst (1.8% mAP). Although *GT* uses ground truth joint positions to extract body features, they are computed in a single key frame, thus ignoring motion. Expectedly, adding motion via sift tracking (*GT single pose + track (GT-T)*) improves the results to 2.2% mAP. Replacing ground truth by predicted joint locations (*PS single pose + track (PS-T)*) results in a performance drop (1.2% vs 2.2% mAP) due to unreliable initialization of the tracker by imperfect pose estimation. Surprisingly, *PS multi-pose (PS-M)* significantly improves the results, achieving 4.2% mAP. It shows that performing body joint predictions in each individual frame is more reliable than simple tracking. Interestingly, the feature level fusion *PS-M + DT (features)* noticeably improves over *DT* alone and classifier level fusion *PS-M + DT (classifiers)*, achieving 5.5% mAP. This shows that both holistic *DT* and pose based *PS-M* methods are complementary. We analyze whether the complementarity of DT comes from the holistic features extracted on the person or elsewhere in the scene. By restricting the extraction to the body mask (*PS-M filter DT*), we observe a drop of performance w.r.t. DT (4.3% mAP vs. 5.1% mAP). It shows that the features extracted outside of the body mask do contain additional information which helps to better discriminate between activities in a fine-grained recognition setting. This intuition is additionally supported by the similar performance of *PS-M filter DT* w.r.t. *PS-M*. Overall, we conclude that holistic and pose based methods are complementary and should be used in a combination for better activity recognition.

Single fully visible people. We now analyze the results in Fig. 1(c). Although the absolute performances are higher, which is explained by an easier setting, the ranking is similar to Fig. 1(b). Two differences are: 1) *GT-T* achieves similar performance to *PS-M* on many activity classes, but looses in total (3.4% mAP vs. 4.2% mAP); and 2) *PS-M filter DT* is better than both DT and *PS-M* on a small set of classes, probably because the trajectory features on the background mostly contribute to confusion on this set of activities.

Differences to [12]. Our analysis in a fine-grained activity recognition setting on hundreds of classes leads to conclusions which go beyond the results of [12] obtained from much smaller number of classes from HMDB dataset [13]. First, we compare the performance of DT to a larger number of pose based methods and show the superior performance of DT, when evaluated on hundreds of activities. This is in contrast to [12] showing that the pose based *PS-M* is better. Second, we discover that holistic DT and pose based *PS-M* are complementary and their combination outperforms each of the approaches alone. This contradicts the conclusions of [12] which does not show any improvement when combining DT and *PS-M*. Finally, we showed that using the trajectories restricted to body only degrades the performance, which suggests that the context adds to the discrimination between activity classes.

4.2 Analysis of activity recognition challenges

After analyzing the overall recognition performance on all classes, we explore which factors affect the performance of best performing holistic DT, pose based *PS-M* and combination *PS-M + DT (features)* of both methods. We use the complexity measures $1-3$ specific for static pose estimation and our novel $1-5$ motion specific complexity measures described in Sec. 2. To make the evaluation consistent with the rest of the experiments, we compute the average complexities for the whole activity class and use the obtained real values to rank the classes. This is in contrast to [2] which computes the measures per single pose and thus operates on individual instance level. To visualize the performance, we sort the activities using the pose related complexity measures in *increased* complexity order, and motion related complexity measures in the *decreased* order. As performance may still be dominated by the training set size when only few examples are available, we restrict the evaluation to the 150 largest activity classes. This corresponds to a slice at 150 in Fig. 1(b). The results are shown in Fig. 2.



(a) *DT* (b) *PS multi-pose (PS-M)* (c) *PS-M + DT (features)*

Fig. 2. Performance (mAP) on a subset of 150 activities from *Separate people* as a function of the complexity measures. Best viewed in color and with additional zooming.

Dense trajectories (DT). Analyzing the results in Fig. 2(a) we observe that a high number of dense trajectories everywhere in the video (*# DT*) and on human body (*# DT body*) leads to the best performance of the *DT* method. Also, we notice that high motion speed (*MS*, *MS body*) is an indicative factor for good recognition results. Surprisingly, *DT* performs better on activities with a high number of people (*# people*). This is explained by the fact that more people potentially produce more motion, which is a positive factor for *DT*. On the other hand, being close to the average pose (*Pose*) and having little occlusion (*Occlusion*) hurts performance. The former is not very surprising, as the average pose is common to many activities, which makes it more difficult for *DT* to capture distinctive features. We discover that activities with little occlusion often contain either little motion (e.g. "sitting, talking in person") or fine-grained motion (e.g. "wash dishes") and thus are hard for *DT*.

PS multi-pose (PS-M). We now analyze Fig. 2(b) and observe several distinctive differences w.r.t. which factors mostly affect the performance of *PS-M*. It can be seen that *MS* and *MS body* have stronger effect on *PS-M* compared to *DT*, and the higher the speed, the better the performance. In order to better understand this nontrivial trend, we analyze which activities happen to produce highest *MS body*. We note that those are sports, dancing and running related activities, for which the pose estimation performance of [32] is above average (cf. Fig. 7 in [2]). Also, these activities exhibit characteristic body part motions and can successfully be encoded using body features. At the other end of the *MS body* scale are the activities with low fine-grained motion, related to home repair, self care and occupation, for which the pose estimation performance is much worse. *Pose* and *Viewpoint* strongly affect the performance as well, as frontal upright people whose pose is close to the mean pose are easier for pose estimation. This is again in contrast to *DT*, where the performance is not noticeably affected by *Viewpoint* and even drops in case of low *Pose*. Surprisingly, high *# people* positively affects *PS-M*. Looking at top ranked activities, we notice that many of them are related to active group exercises or team sports, such as "aerobic" and "frisbee", or to simple standing postures, such as "standing, talking in person". Body features can again be successfully used to encode the corresponding motions. On the other hand, we observe that the high *# DT* and *# DT body* hurts performance, which is in contrast to the *DT* method. We

	yoga, power	bicycl., mount.	skiing, downh.	cooking or food	skate-board.	rope skip.	softball, general	forestry
Dense trajectories (DT)	10.6	14.5	51.9	0.5	11.4	36.0	**12.7**	8.4
GT single pose (GT)	22.3	26.5	7.5	1.8	3.4	51.2	2.2	1.4
GT single pose + track (GT-T)	**37.0**	28.0	10.9	**2.6**	4.6	69.2	3.6	1.2
PS single pose + track (PS-T)	8.8	6.6	6.0	1.3	1.7	63.1	1.6	1.8
PS multi-pose (PS-M)	18.3	34.0	27.3	2.6	17.2	**90.5**	3.0	5.2
PS-M + DT (features)	19.6	**40.7**	32.9	2.2	**19.5**	88.7	3.9	7.2
PS-M filter DT	16.1	20.4	**52.2**	0.8	13.5	55.7	4.2	**10.6**

	carpentry, general	bicycl., racing	golf	rock climb.	ballet, modern	aerobic step	resist. train.	total
Dense trajectories (DT)	5.5	5.5	33.0	**41.5**	12.7	24.5	**16.5**	19.0
GT single pose (GT)	2.7	7.1	36.1	2.3	1.0	1.1	1.4	11.2
GT single pose + track (GT-T)	2.8	8.7	25.3	8.9	1.7	3.3	1.3	13.9
PS single pose + track (PS-T)	5.3	0.5	14.7	1.2	2.8	11.1	1.6	8.5
PS multi-pose (PS-M)	3.4	8.6	47.9	4.7	22.9	10.4	7.2	20.2
PS-M + DT (features)	5.0	12.1	**51.9**	14.4	**23.7**	17.1	14.4	**23.5**
PS-M filter DT	**6.1**	**15.5**	15.9	38.6	7.1	**25.8**	9.6	19.5

Table 1. Activity recognition results (mAP) on 15 largest classes from *Separate people*.

observe that for high *# DT body* many activities correspond to water related activities, such as "fishing in stream", "swimming, general", "canoeing, kayaking". Interestingly, the presence of water leads to high *# DT* and characteristic motions captured by *DT*. At the same time *PS-M* fails due to unreliable pose estimation caused by complex appearance and occlusions.

PS-M + DT (features). The differences for *DT* and *PS-M* methods suggest that both methods are complementary. We analyze in Fig. 2(c) which factors affect the performance of *PS-M + DT (features)*. It can be seen that positively affecting factors are either positive for both *DT* and *PS-M* (*MS*, *MS body*, *# people*), or positive for *PS-M* only (*Pose*, *Viewpoint*). In contrast to *PS-M* the high *# DT* slightly improves the performance, while high *# DT body* does not hurt as much. Expectedly, *Viewpoint* hurts performance as it does for both *DT* and *PS-M*. This shows the complementarity of both *DT* and *PS-M* and leaves room for improvement in finding better ways of combining both methods.

4.3 Detailed analysis on a subset of activities

After analyzing the factors affecting the results by different methods, we conduct a detailed analysis on a smaller set of the top 15 activities from *Separate people*.

The results are shown in Tab. 1. None of the methods outperforms all others on all activities and different approaches are better on different activities. On average methods perform well on activities with simple poses and motions e.g. "rope skipping", "skiing, downhill" and "golf" - typical cases in most of the current activity recognition benchmarks. However, the performance of all methods is low for activities with more variability in motion and poses, e.g. "cooking", "carpentry, general" and "forestry". This leaves room for improvement of current methods. Analyzing the performance on individual activities, we observe that for "yoga, power" activity *GT* outperforms holistic *DT* and *PS-M filter DT* methods (22.3% vs. 10.6% and 16.1% mAP, respectively) and is better than the pose based *PS-M* (22.3% vs. 18.3% mAP). It is interesting, as *GT* does not use

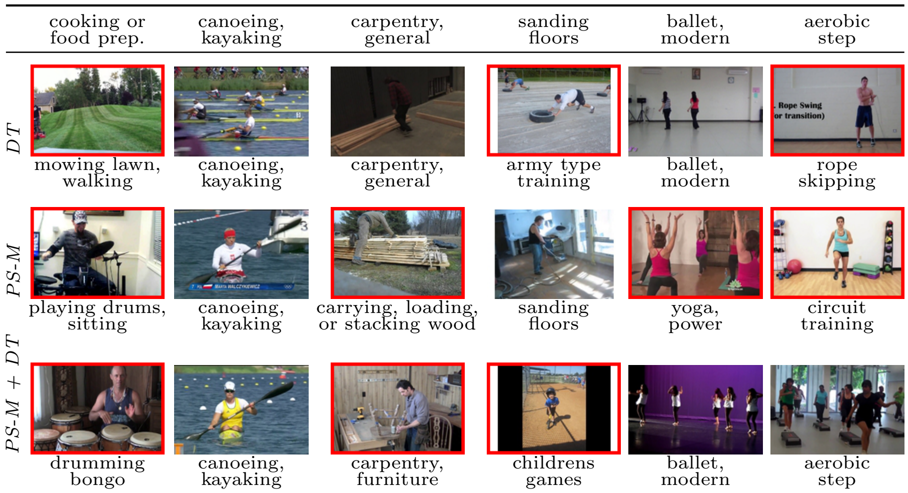


Fig. 3. Successful and failure cases on several activity classes. Shown are the most confident prediction per class. False positives are highlighted in red.

any motion and relies on static body features only. The explanation is that the "yoga, power" activity contains distinctive body poses and thus can be reliably captured by *GT*, while *PS-M* fails due to unreliable pose estimations. It can be seen that in many cases the combination *PS-M + DT (features)* noticeably outperforms both *PS-M* and *DT* alone. The differences are most pronounced for "bicycling, mountain", "bicycling, racing", "skateboarding" exhibiting characteristic motions, and "golf" activity having distinctive body motion and poses. Overall *PS-M + DT (features)* achieves the best performance of 23.5% mAP. We visualize several successful and failure cases of the methods in Fig. 3.

5 Conclusion

In this work we address the challenging task of fine-grained human activity recognition on a recent comprehensive dataset with hundreds of activity classes. We study holistic and pose based representations and analyze the factors responsible for their performance. We reveal that holistic and pose based methods are complementary, and their performance varies significantly depending on the activity. We found that both methods are strongly affected by the speed of trajectories. While the holistic method is also strongly influenced by the number of trajectories, pose based methods are strongly affected by human pose and viewpoint. We observe striking performance differences across activities and experimentally show that the combination of both methods performs best.

Acknowledgements. The authors would like to thank Marcus Rohrbach and Sikandar Amin for helpful discussions. This work has been supported by the Max Planck Center for Visual Computing & Communication.

References

1. Ainsworth, B., Haskell, W., Herrmann, S., Meckes, N., Bassett, D., Tudor-Locke, C., Greer, J., Vezina, J., Whitt-Glover, M., Leon, A.: 2011 compendium of physical activities: a second update of codes and MET values. MSSE'11
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR'14
3. Brendel, W., Todorovic, S.: Learning spatiotemporal graphs of human activities. In: ICCV'11
4. Cardinaux, F., Bhowmik, D., Abhayaratne, C., Hawley, M.S.: Video based technology for ambient assisted living: A review of the literature. J. Ambient Intell. Smart Environ.'11
5. Chakraborty, B., Holte, M.B., Moeslund, T.B., Gonzalez, J., Xavier Roca, F.: A selective spatio-temporal interest point detector for human action recognition in complex scenes. In: ICCV'11
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR'05
7. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: ECCV'06
8. Dantone, M., Gall, J., Leistner, C., Gool., L.V.: Human pose estimation using body parts dependent joint regressors. In: CVPR'13
9. Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: ICCV'09
10. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
11. Ferrari, V., Marin, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR'08
12. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV'13
13. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision (ICCV) (2011)
14. Laptev, I.: On space-time interest points. IJCV'05
15. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR'08
16. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: CVPR'09
17. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR'09
18. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: CVPR'13
19. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Strong appearance and expressive spatial models for human pose estimation. In: ICCV'13
20. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR'08
21. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. In: CVPR'12
22. Rohrbach, M., Stark, M., Schiele, B.: Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: CVPR'11

23. Sadanand, S., J., C.J.: Action bank: A high-level representation of activity in video. In: ECCV'12
24. Sapp, B., Taskar, B.: Multimodal decomposable models for human pose estimation. In: CVPR'13
25. Singh, V.K., Nevatia, R.: Action recognition in cluttered dynamic scenes using pose-specific part models. In: ICCV'11
26. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human action classes from videos in the wild. Tech. Rep. CRCV-TR-12-01, UCF (2012)
27. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. In: CVPR'10
28. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. VC'13
29. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. IJCV'13
30. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV'13
31. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC'09
32. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. PAMI'13