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

Goal

Analysis of the state of the art in the fine-grained human activity recognition on a large scale dataset

Contributions

- Analysis of holistic and pose based approaches for human activity recognition
- Large scale comparison on "MPII Human Pose" dataset
- Analysis of factors responsible for success and failure of holistic and pose based methods

"MPII Human Pose" dataset [2]

- **Systematically collected from YouTube** videos using established taxonomy [1] of everyday human activities
- Covers **410 human activities**
- Contains around **25K images**, **40K annotated poses**
- **Rich annotations** on test set: 3D torso and head orientation, body part occlusions
- **Video** snippet for each image, over **1M frames**
- Available at **human-pose.mpi-inf.mpg.de**

Methods

Holistic method

- Dense Trajectories (DT) [5]

Pose based methods

- Ground truth (GT) single pose
- GT single pose + track (GT-T)
- Pictorial Structures (PS) single pose + track (PS-T) [4]
- PS multi-pose (PS-M) [3]

Holistic + pose based methods

- PS-M + DT (features): feature level fusion
- PS-M + DT (classifiers): decision level fusion
- PS-M filter DT: filter using body part masks

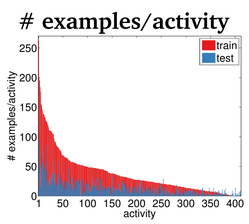
Experimental Setup

Data

- Sufficiently separated people
- 15,2K videos train / 5,7K test

Training and evaluation

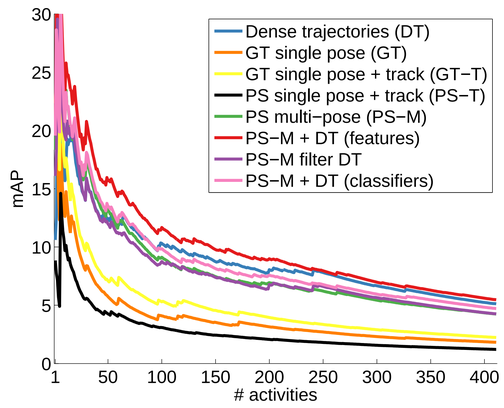
- Bag-of-Words representation
- One-vs-all SVMs using SGD and χ^2 kernel
- Evaluation using mean Average Precision (mAP)



Randomly chosen activities and images from 18 top level categories of our "MPII Human Pose" dataset. One image per activity is shown. The full dataset is available at **human-pose.mpi-inf.mpg.de**.

Overall Activity Recognition Performance

- Order activities based on training set size



⇒ Performance quickly drops for large number of classes

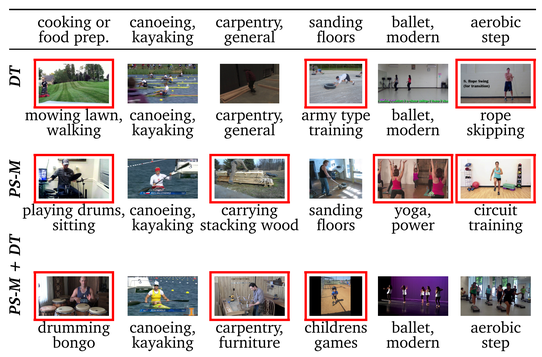
⇒ Holistic *DT* outperforms all pose based methods

⇒ *PS-M* performs best among pose based approaches

⇒ Combination *PS-M + DT (features)* outperforms both *PS-M* and *DT*

⇒ **Holistic and pose based methods complementary**

Successful and failure cases



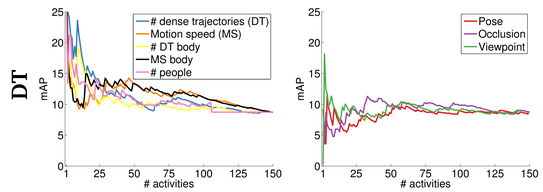
Analysis of Activity Recognition Challenges

Motion specific challenges

- *# DT*: number of dense trajectories
- *MS*: motion speed of dense trajectories
- *# DT body*, *MS body*: trajectories on body mask only
- *# people*: number of people

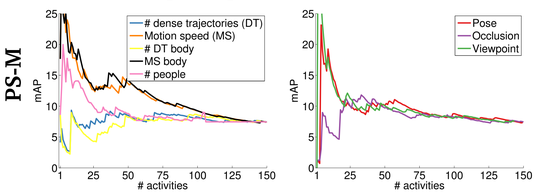
Human pose specific challenges

- *Pose*: deviation from the mean pose
- *Occlusion*: number of occluded body parts
- *Viewpoint*: deviation of 3D torso rotation from frontal



✓ Much motion indicative for good performance

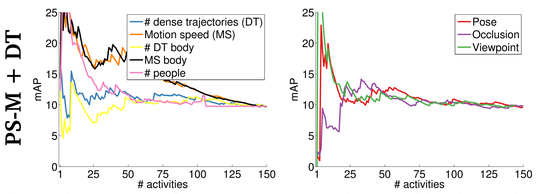
× Close to mean poses produce non-discriminative DT



✓ High *MS* is indicative - "easy" sport related poses

✓ Close to mean poses and frontal views - easy poses

× High *# DT*: water related activities - hard poses



✓ Combination inherit positive qualities of both methods

⇒ **Holistic and pose based methods complementary**

Detailed Analysis on a Subset of Activities

	yoga, power	bicycl., mount.	skiing, downh.	cooking or food	skate-board.	rope skip.	softball, general	forestry
DT	10.6	14.5	51.9	0.5	11.4	36.0	**12.7**	8.4
GT	22.3	26.5	7.5	1.8	3.4	51.2	2.2	1.4
GT-T	**37.0**	28.0	10.9	**2.6**	4.6	69.2	3.6	1.2
PS-T	8.8	6.6	6.0	1.3	1.7	63.1	1.6	1.8
PS-M	18.3	34.0	27.3	2.6	17.2	**90.5**	3.0	5.2
PS-M + DT (feat.)	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
DT	5.5	5.5	33.0	**41.5**	12.7	24.5	**16.5**	19.0
GT	2.7	7.1	36.1	2.3	1.0	1.1	1.4	11.2
GT-T	2.8	8.7	25.3	8.9	1.7	3.3	1.3	13.9
PS-T	5.3	0.5	14.7	1.2	2.8	11.1	1.6	8.5
PS-M	3.4	8.6	47.9	4.7	22.9	10.4	7.2	20.2
PS-M + DT (feat.)	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

⇒ Each method performs best on few activities only

⇒ Good performance on "golf" and "rope skipping": simple poses and motions

⇒ Poor performance on "cooking" and "forestry": high variability in motion and poses

⇒ Combination *PS-M + DT (features)* is best on average

⇒ **Holistic and pose based methods complementary**

Conclusion

- Striking performance differences across activities
- Holistic method influenced by high degree of motion
- Pose methods affected by human pose and viewpoint
- Combination holistic + pose method performs best

References

[1] B. Ainsworth, W. Haskell, S. Herrmann, N. Meckes, D. Bassett, C. Tudor-Locke, J. Greer, J. Vezina, M. Whitt-Glover, and A. Leon. 2011 compendium of physical activities: a second update of codes and MET values. *MSSE'11*.

[2] M. Andriluka, L. Pishchulin, P Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR'14*.

[3] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV'13*.

[4] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR'12*.

[5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV'13*.

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