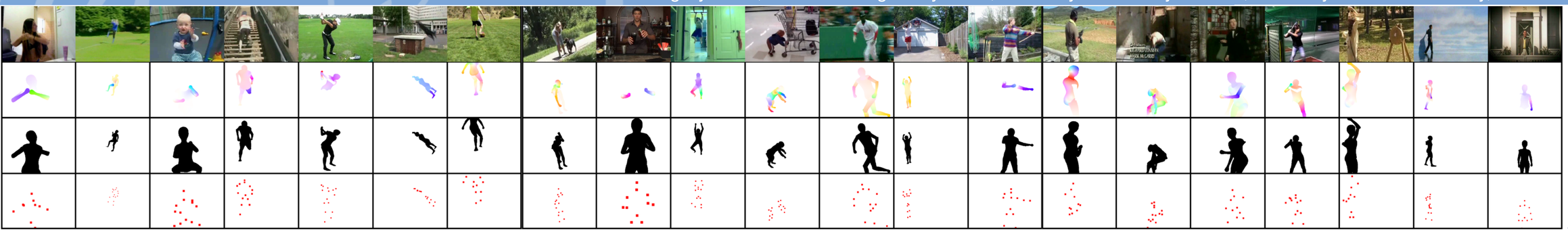
Towards Understanding Action Recognition

Hueihan Jhuang[1], Juergen Gall[2], Silvia Zuffi[1,3], Cordelia Schmid[4] and Michael J. Black[1]

[1]Perceiving Systems, MPI for Intelligent Systems, Germany [2] University of Bonn, Germany [3] Brown University, USA [4] LEAR, INRIA, France

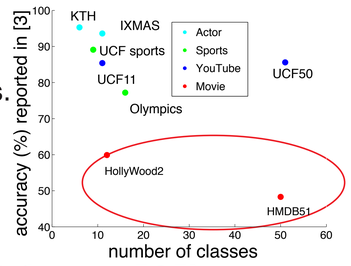


see the corrected paper and annotations at http://jhmdb.is.tue.mpg.de

Problem

Recognizing actions in movies is hard. Why?

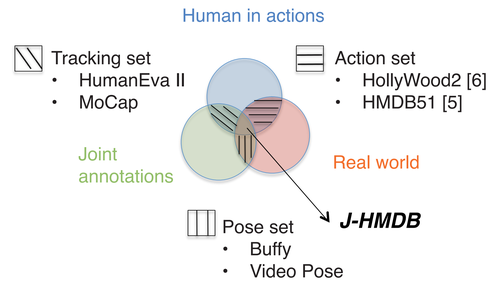
- We don't have enough annotations.
- We don't know what are important algorithm properties.
- We don't know what are important features.



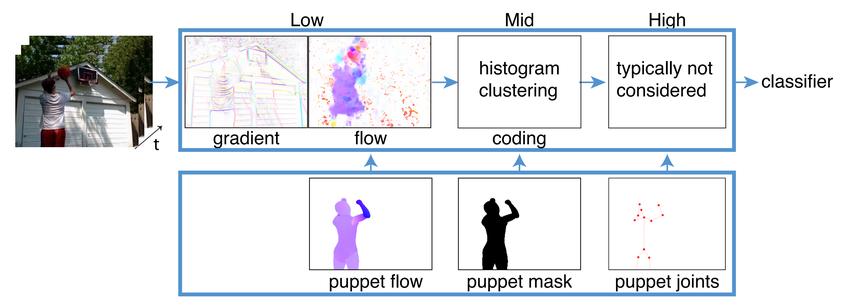
Dataset J-HMDB (videos from [1])

Videos

- movies, YouTube
- 21 classes
- one main actor
- 928 clips
- 15+ frames / clip
- 31,838 frames
- 240 x 320 pixels



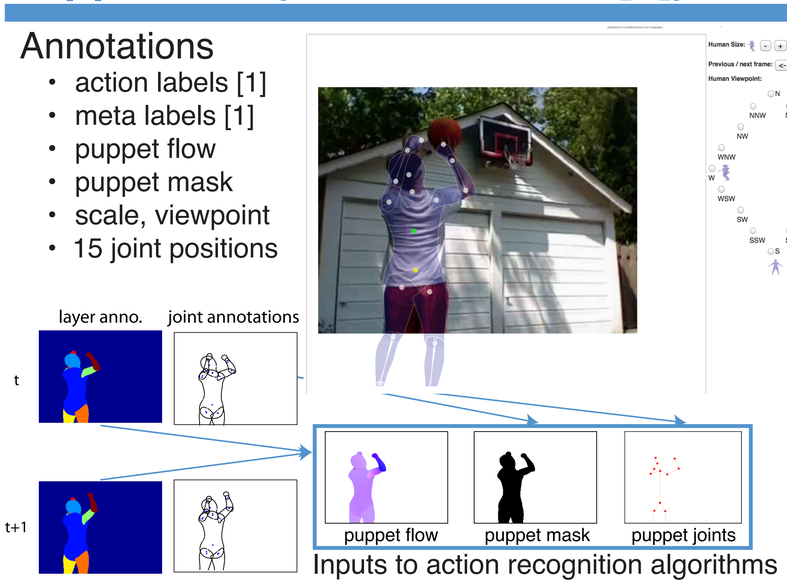
Method



Puppet tool (extended from [2])

Annotations

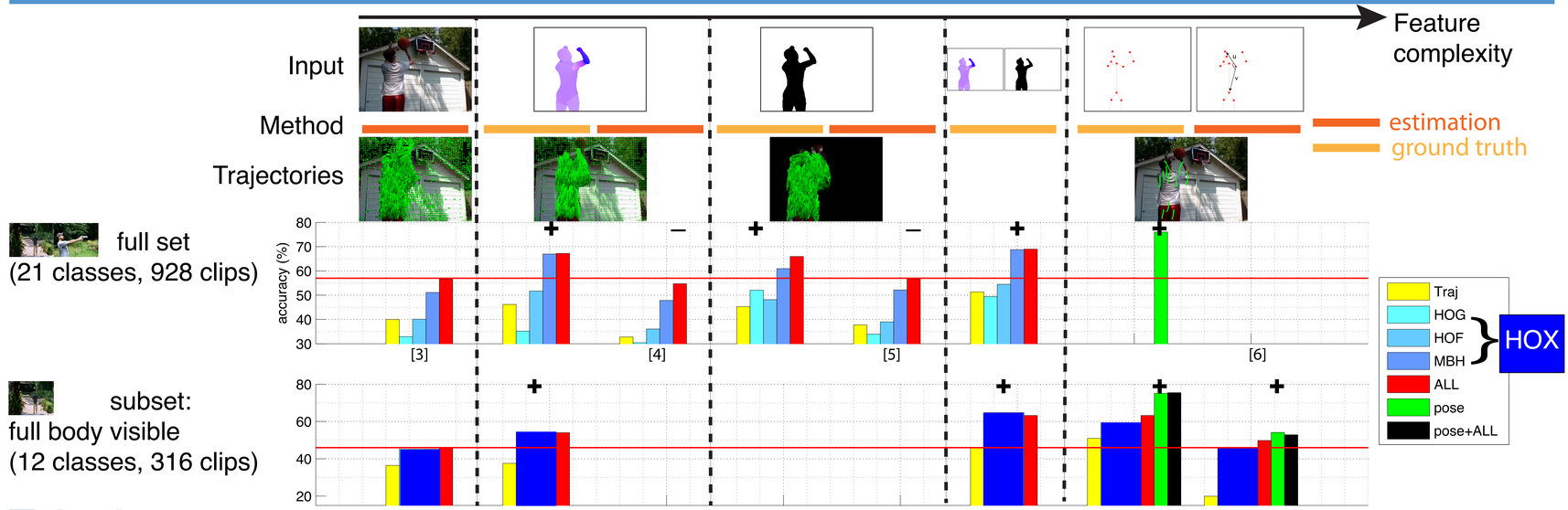
- action labels [1]
- meta labels [1]
- puppet flow
- puppet mask
- scale, viewpoint
- 15 joint positions



Inputs to action recognition algorithms

Results (on Dense Trajectories [3])

Systematically replace stages of [3] with ground truth data.



Take home messages

If you use optical flow for action recognition:
1. GT flow leads to ~11% gain. GT masks lead to ~9 % gain.
2. Better flow on standard benchmarks doesn't mean better flow for action recognition.

If you do action recognition at all:
1. GT pose-based features lead to ~20% gain.
2. Estimated pose-based already outperforms flow-based features for visible full body.

References

[1] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *ICCV*, 2011.
[2] S. Zuffi, O. Freifeld, and M. J. Black.From pictorial structures to deformable structures. In *CVPR*, 2012.
[3] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. *IJCV*, 2013.
[4] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. IJCV, to appear, 2013.
[5] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. ECCV, 2010.
[6] Y. Yang and D. Ramanan.Articulated pose estimation using flexible mixtures of parts. *PAMI*, 2013.

Acknowledgements

JG was supported in part by the DFG Emmy Noether program (GA 1927/1-1) and CS by the ERC advanced grant Allegro.