Strong Appearance and Expressive Spatial Models for Human Pose Estimation

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

[1]Max Planck Institute for Informatics, Saarbrücken, Germany

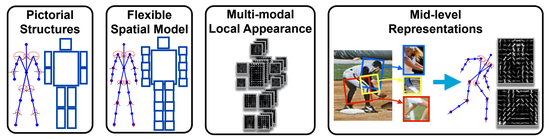
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

Goal

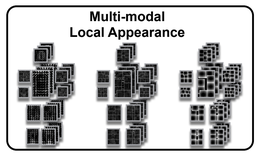
- **Analyze recent key ideas** in 2D human pose estimation
- **Push state of the art** by leveraging most powerful components

Contributions

- Show **complementarity** of recent key ideas

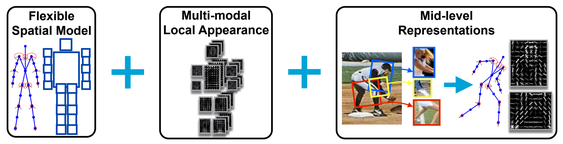


- Propose **strong local appearance** models



⇒ **state of the art results even with basic tree model**

- Propose **powerful model** leveraging complementarity



⇒ **best result to date: 69.2% PCP (+4.9%) on LSP**

Code available!
www.d2.mpi-inf.mpg.de/poselet-conditioned-ps

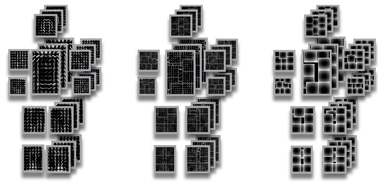
Related Work

Method	PS	Appear. mix	Spatial Model flex	Spatial Model img cond
Andriluka et al., CVPR'09	√	×	×	×
Johnson&Evering., BMCV'10	√	√	×	×
Yang&Ramanan, CVPR'11	√	√	√	×
Dantone et al., CVPR'13	√	√	√	×
Sapp&Taskar, CVPR'13	√	√	√	×
Pishchulin et al., CVPR'13	√	×	×	√
our method	√	√	√	√

⇒ **our method leverages most powerful recent ideas**

Strong Appearance Models

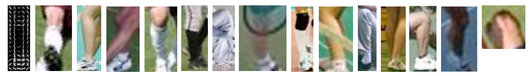
Strong local appearance via DPM detectors [3]



I. Local appearance, single component

- rotation-dependent (*rot-dep single*)

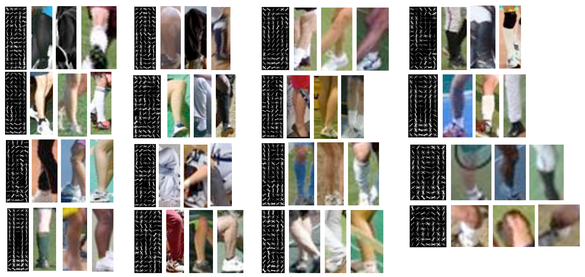
- rotation-invariant (*rot-inv single*)



II. Local appearance mixtures

- rotation-dependent mixtures (*rot-dep mix*)

- pose-dependent mixtures (*pose-dep mix*)



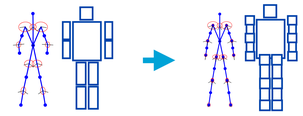
III. Specialised torso and head detectors

- *spec-torso*: regress torso from entire body detection
- *spec-head*: mixture of viewpoint-specific components

Expressive Spatial Models

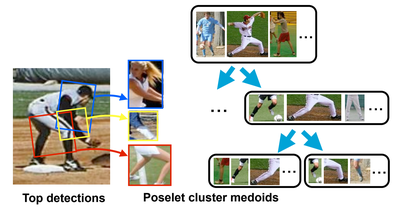
I. Flexible spatial model

- extend PS [1] by extra parts on body joints (*PS-flex*)



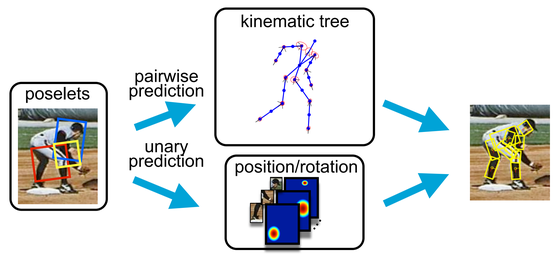
II. Mid-level representations

- use poselets to detect joint part configurations



√ capture **non-adjacent** part dependencies

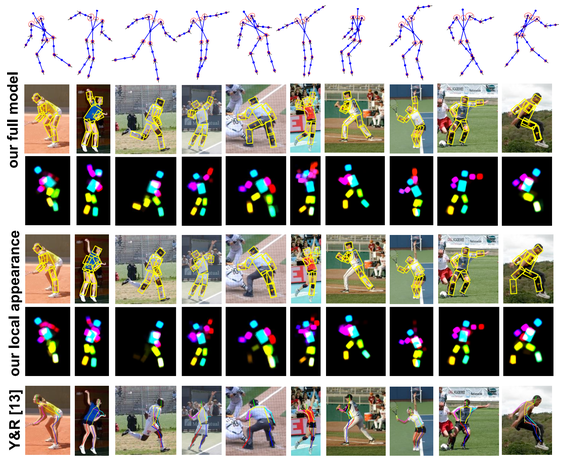
- condition spatial and appearance terms on poselets [5]



- prediction by multi-class classifier before inference (*mid-level appearance, mid-level p/wise*)

√ **exact and efficient** inference

Qualitative Results



Quantitative Results

Leeds Sports Poses (LSP) [4]

- 1,000 train, 1,000 test images
- *observer-centric* annotations for testing [2]
- *Percentage Correct Parts (PCP)* criterion

Method	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
PS [1]	80.9	67.1	60.7	46.5	26.4	74.9	55.7
PS-flex	80.5	70.2	66.5	46.7	32.0	70.2	58.1
+ rot-dep single	82.2	72.5	67.9	51.6	31.6	78.3	60.8
+ rot-inv single	83.6	73.6	69.8	52.4	39.4	78.1	63.2
+ rot-dep mix	87.2	76.0	72.2	55.9	40.5	83.3	66.0
+ pose-dep mix	84.5	75.4	70.3	53.4	40.5	78.0	64.2
+ spec head/torso	89.2	76.7	72.8	56.9	41.2	84.7	66.9
+ mid-level appearance	**89.4**	78.7	**74.0**	59.7	43.9	**86.0**	68.8
+ mid-level p/wise	88.7	**78.8**	**73.4**	**61.5**	**44.9**	85.6	**69.2**
Yang&Ramanan, CVPR'11	84.1	69.5	65.6	52.5	35.9	77.1	60.8
Pishchulin et al., CVPR'13	87.5	75.7	68.0	54.2	33.9	78.1	62.9
Eichner&Ferrari, ACCV'12	86.2	74.3	69.3	56.5	37.4	80.1	64.3

Image Parse (IP) [6]

- 100 train, 205 test images

Method	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
Our full model	**93.2**	77.1	68.0	63.4	**48.8**	86.3	**69.4**
Andriluka et al., IJCV'11	86.3	66.3	60.0	54.6	35.6	72.7	59.2
Yang&Ramanan, CVPR'11	82.9	69.0	63.9	55.1	35.4	77.6	60.7
Duan et al., BMVC'12	85.6	71.7	65.6	57.1	36.6	80.4	62.8
Pishchulin et al., CVPR'13	92.2	74.6	63.7	54.9	39.8	70.7	62.9
Yang&Ramanan, PAMI'12	85.9	74.9	**68.3**	63.4	42.7	**86.8**	67.1
Johnson&Everingham, CVPR'11	87.6	74.7	67.1	**67.3**	45.8	76.8	67.4

Limitations

Conclusion

- local and mid-level representations are complementary
- strong local appearance model already outperforms state of the art when using basic tree connectivity
- best result to date by leveraging complementarity

⇒ code available!

References

[1] M. Andriluka, S. Roth, and B. Schiele. Discriminative appearance models for pictorial structures. *IJCV'11*.

[2] M. Eichner and V. Ferrari. Appearance sharing for collective human pose estimation. In *ACCV'12*.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI'10*.

[4] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC'10*.

[5] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR'13*.

[6] D. Ramanan. Learning to parse images of articulated objects. In *NIPS'06*.