Pictorial Structures Revisited: People Detection and Articulated Pose Estimation

Mykhaylo Andriluka

Stefan Roth

Bernt Schiele

Department of Computer Science
TU Darmstadt

Generic model for human detection and pose estimation

TECHNISCHE UNIVERSITÄT DARMSTADT

Human pose estimation

[Felzenszwalb&Huttenlocher, ICCV'05], [Ren et al., ICCV'05], [Sigal&Black, CVPR'06], [Zhang et al., CVPR'06], [Jiang&Marin, CVPR'08], [Ramanan, NIPS'06], [Ferrari et al., CVPR'08], [Ferrari et al., CVPR'09]

often rather simple appearance model
focus on finding optimal assembly of parts

People Detection

[Viola et al., ICCV'03], [Dalal&Triggs, CVPR'05], [Leibe et al., CVPR'05], [Andriluka et al., CVPR'08]

complex appearance model
no pose model or limited to walking motion

Generic model for human detection and pose estimation

Human pose estimation

[Felzenszwalb&Huttenlocher, ICCV'05], [Ren et al., ICCV'05], [Sigal&Black, CVPR'06], [Zhang et al., CVPR'06], [Jiang&Marin, CVPR'08], [Ramanan, NIPS'06], [Ferrari et al., CVPR'08], [Ferrari et al., CVPR'09]

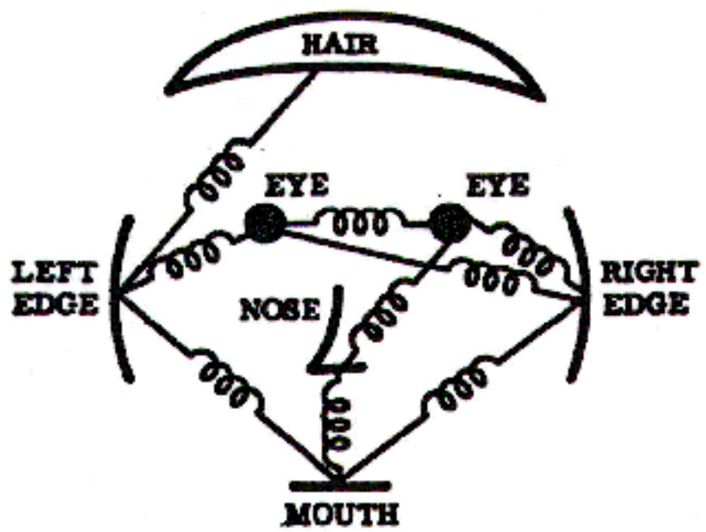
often rather simple appearance model
focus on finding optimal assembly of parts

People Detection

[Viola et al., ICCV'03], [Dalal&Triggs, CVPR'05], [Leibe et al., CVPR'05], [Andriluka et al., CVPR'08]

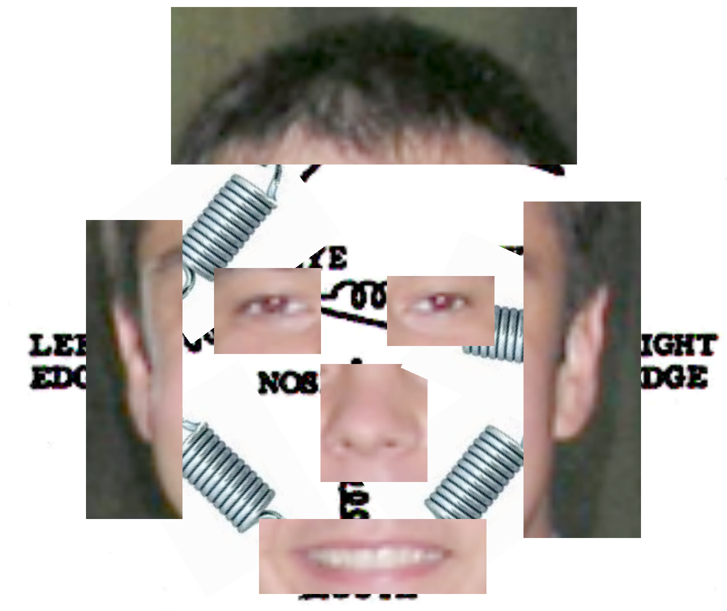
complex appearance model
no pose model or limited to walking motion

Can we make pictorial structures model effective for these tasks?



[Fischler&Elschlager, 1973]

Can we make pictorial structures model effective for these tasks?



Yes... if the model components are chosen right.

Pictorial Structures Model

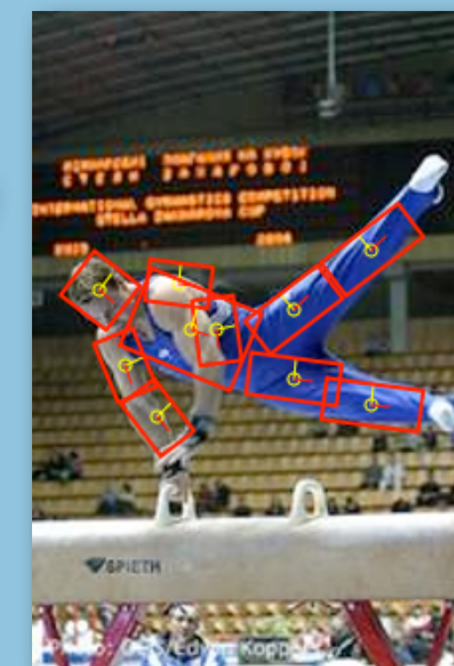
TECHNISCHE UNIVERSITÄT DARMSTADT

- Body is represented as flexible configuration of body parts

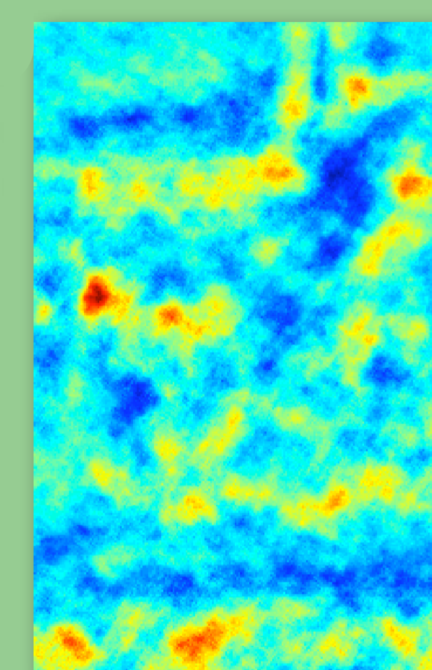
$L = \{\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_N\}$ - configuration of parts

$D = \{\mathbf{d}_0, \mathbf{d}_1, \dots, \mathbf{d}_N\}$ - part evidence

L



$\mathbf{d}_i$



posterior over body poses

$$p(L|D) \propto p(D|L)p(L)$$

likelihood of observations

prior on body poses

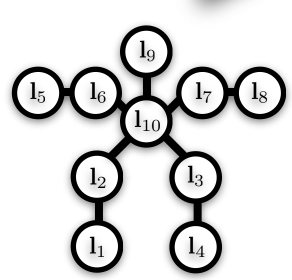
Pictorial Structures Model

Pictorial structures allow
exact and efficient inference.

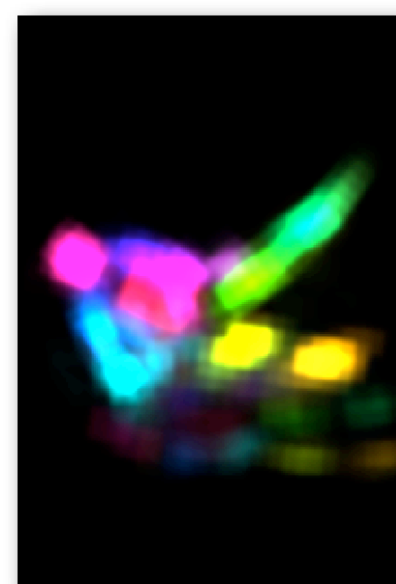
- tree-structured prior
- independent part appearance model
- discretized part locations

- Gaussian pairwise part relationships

sum-product BP

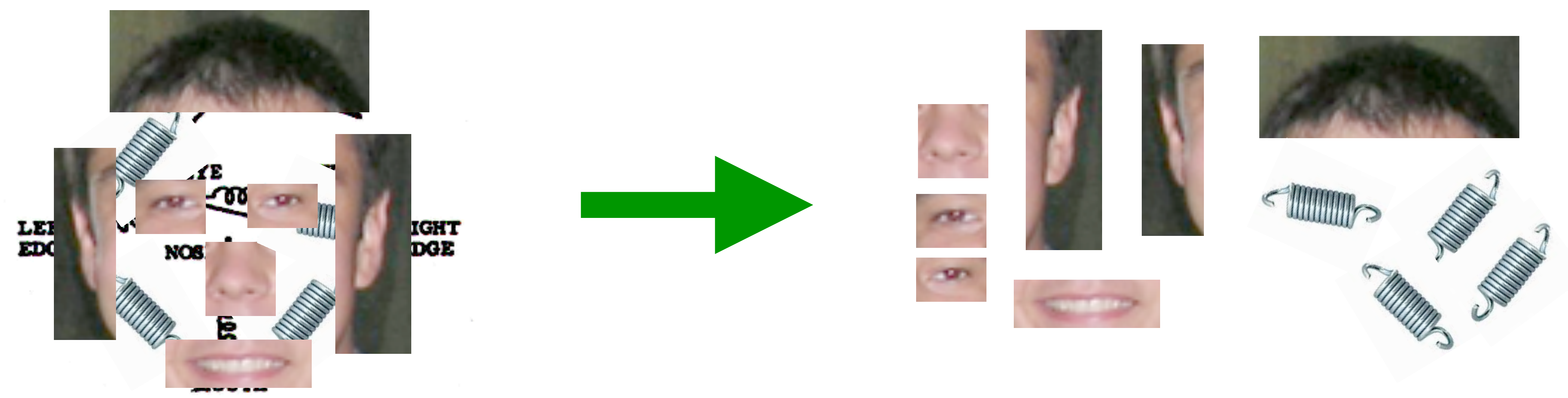


posterior marginals



$$p(\mathbf{l}_i|D) \propto \sum_{L \setminus \mathbf{l}_i} p(L|D)$$

Can we make pictorial structures model effective for these tasks?



So... what are the right components?

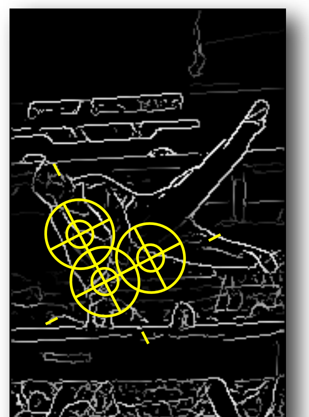
Model Components

TECHNISCHE UNIVERSITÄT DARMSTADT

Appearance Model:

Local Features

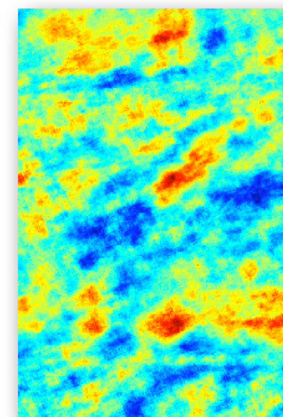
orientation 1



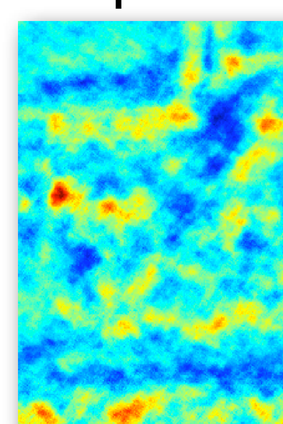
orientation K

AdaBoost

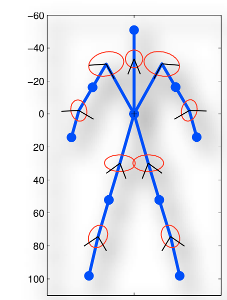
likelihood of part 1



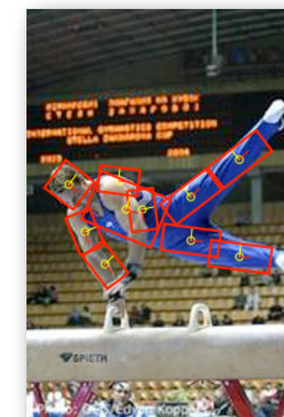
likelihood of part N



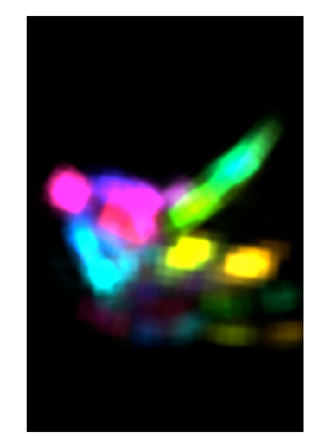
Prior and Inference:



estimated pose



part posteriors



Model Components

Appearance Model:

Local Features

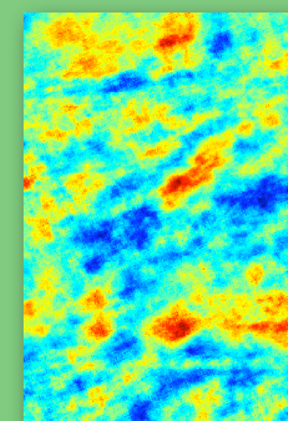
orientation 1

⋮

orientation K

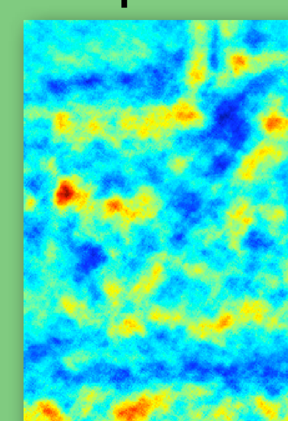
AdaBoost

likelihood of part 1

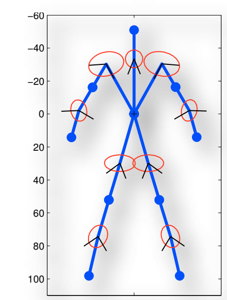


⋮

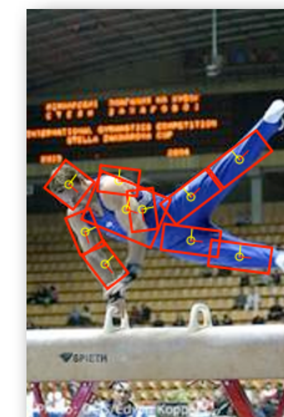
likelihood of part N



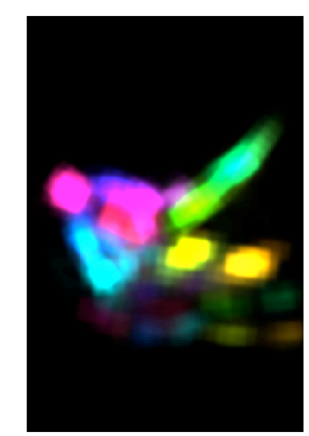
Prior and Inference:



estimated pose



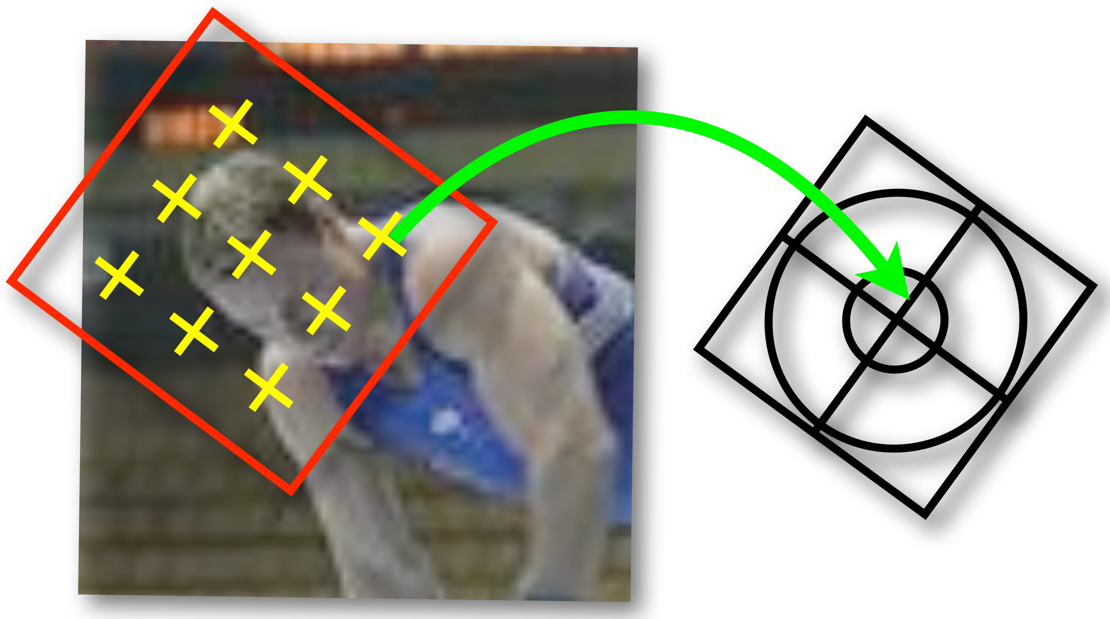
part posteriors



Likelihood Model

TECHNISCHE UNIVERSITÄT DARMSTADT

- Build on recent advances in object detection:
  - state-of-the-art image descriptor: Shape Context
    [Belongie et al., PAMI'02; Mikolajczyk&Schmid, PAMI'05]
  - dense representation
  - discriminative model: AdaBoost classifier for each body part



- Shape Context: 96 dimensions (4 angular, 3 radial, 8 gradient orientations)

- Feature Vector: concatenate the descriptors inside part bounding box
- head: 4032 dimensions
- torso: 8448 dimensions

Likelihood Model

- Part likelihood derived from the boosting score:

decision stump weight

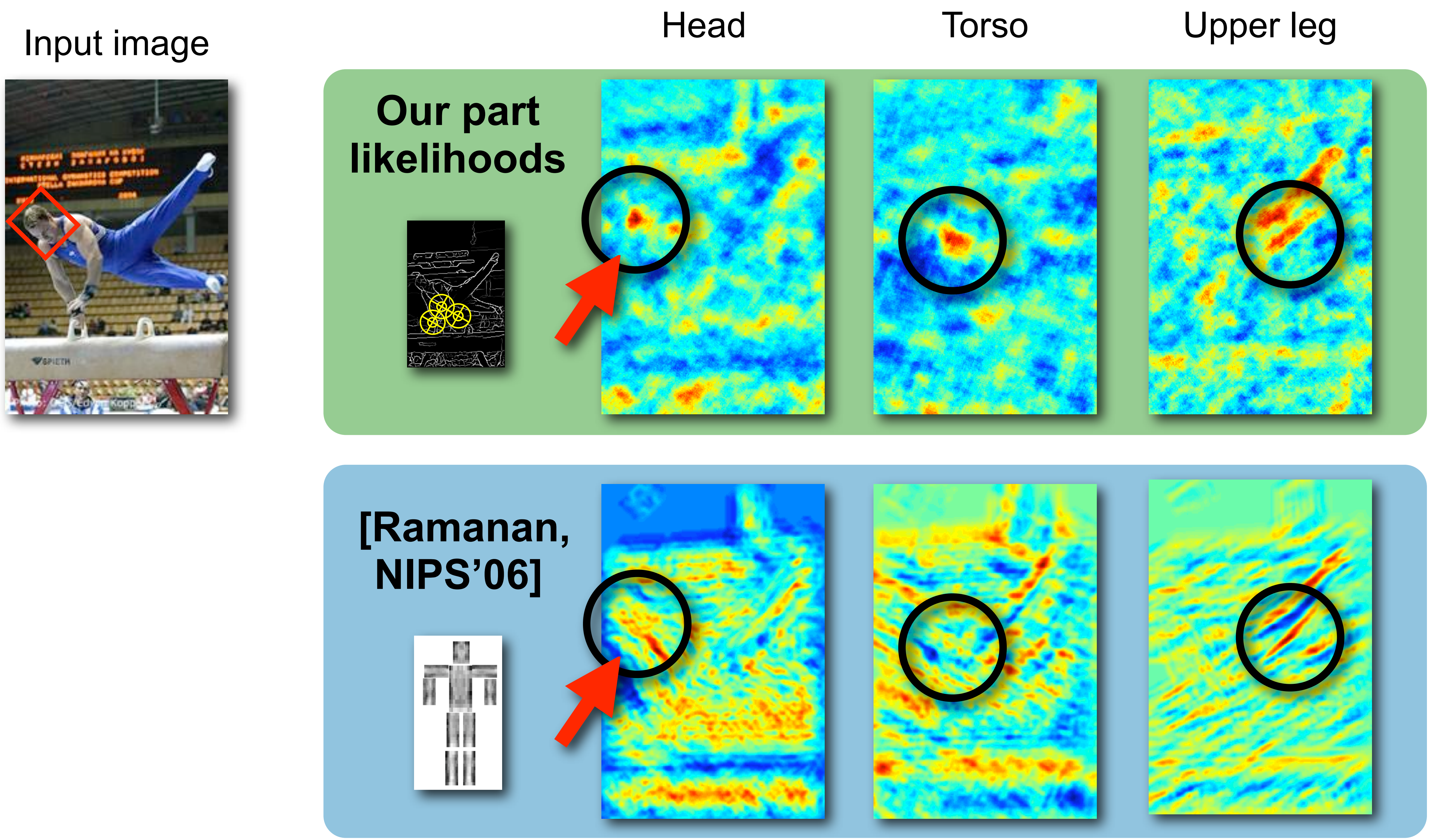
decision stump output

$$\tilde{p}(\mathbf{d}_i|\mathbf{l}_i) = \max\left(\frac{\sum_t \alpha_{i,t} h_t(\mathbf{x}(\mathbf{l}_i))}{\sum_t \alpha_{i,t}}, \varepsilon_0\right)$$

part location

small constant to deal with part occlusions

Likelihood Model



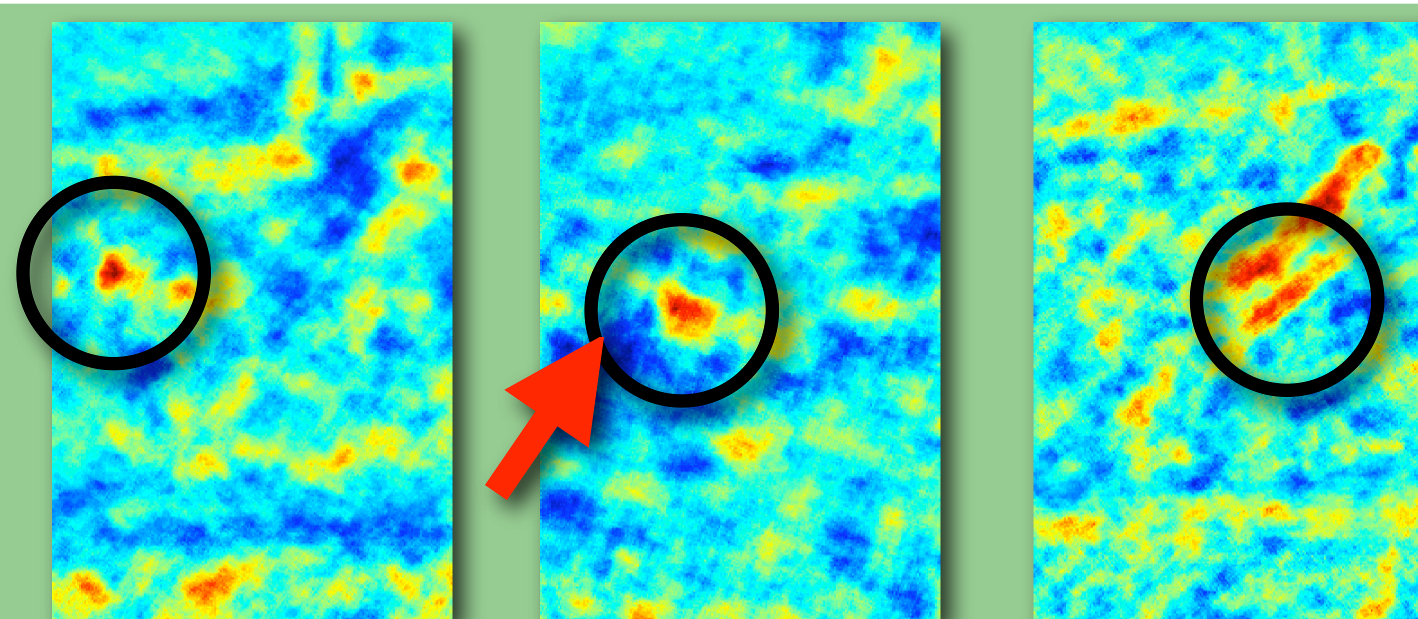
Likelihood Model

TECHNISCHE UNIVERSITÄT DARMSTADT

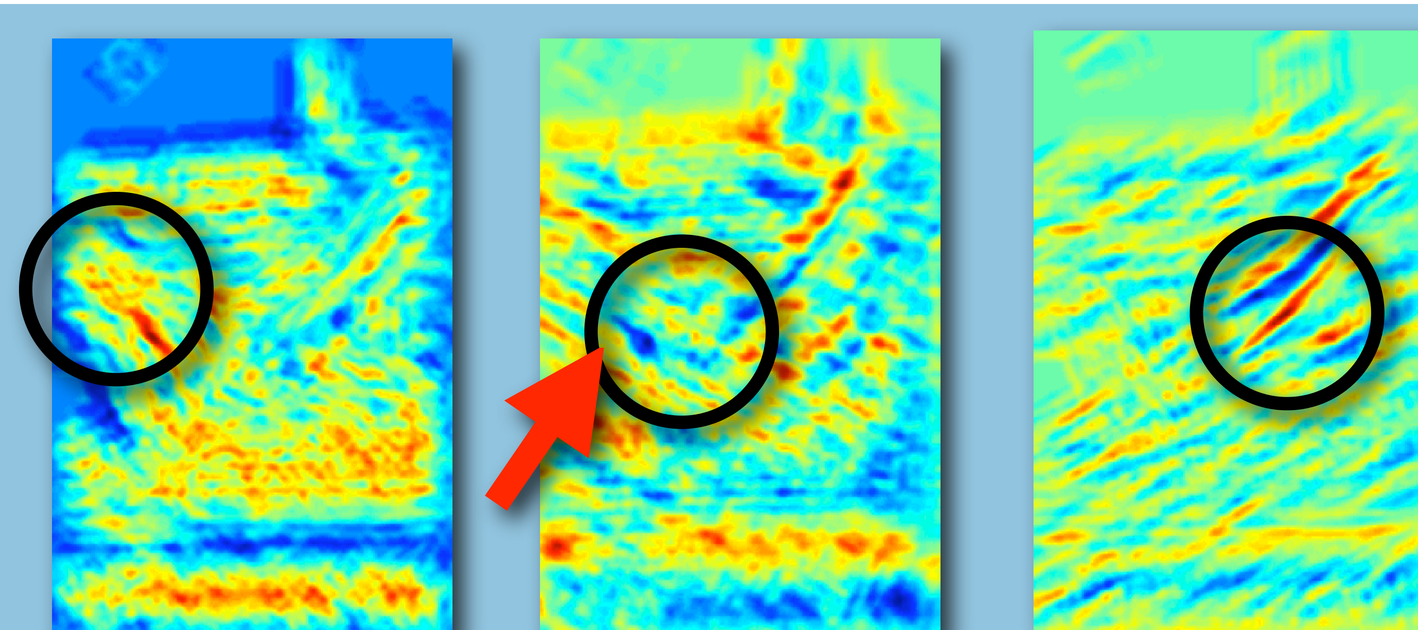
Input image

Head | Torso | Upper leg

Our part likelihoods



[Ramanan, NIPS'06]

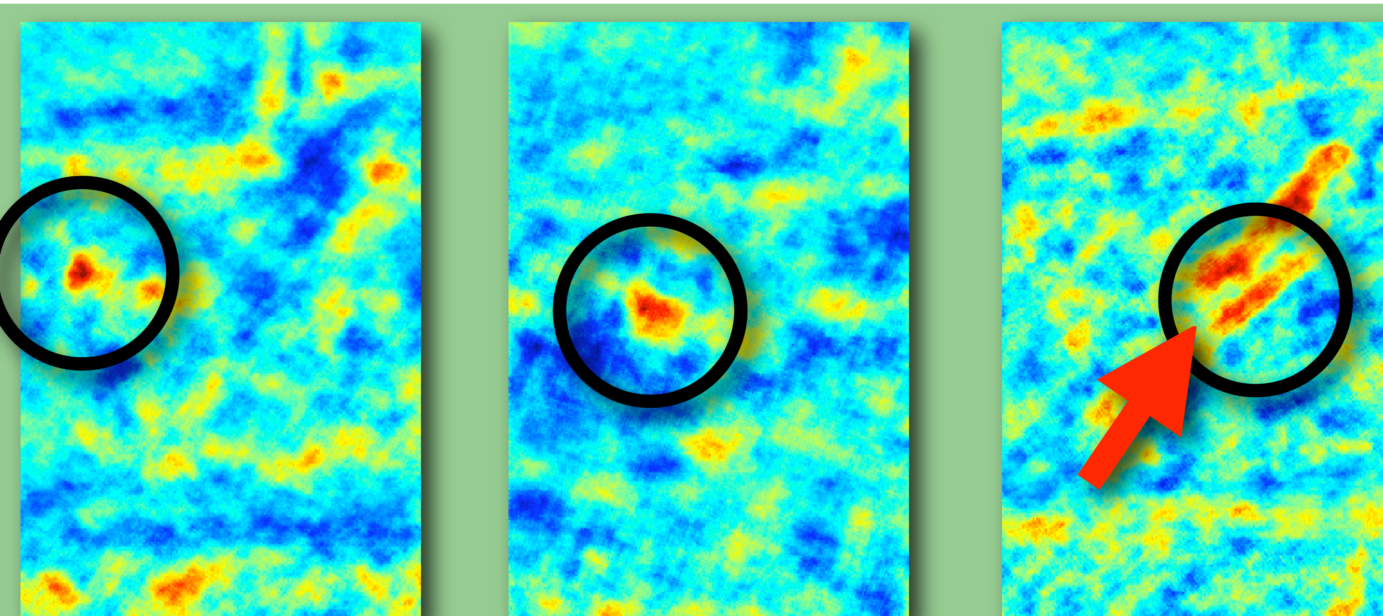


Likelihood Model

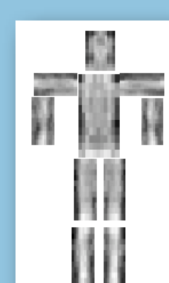
Head | Torso | Upper leg

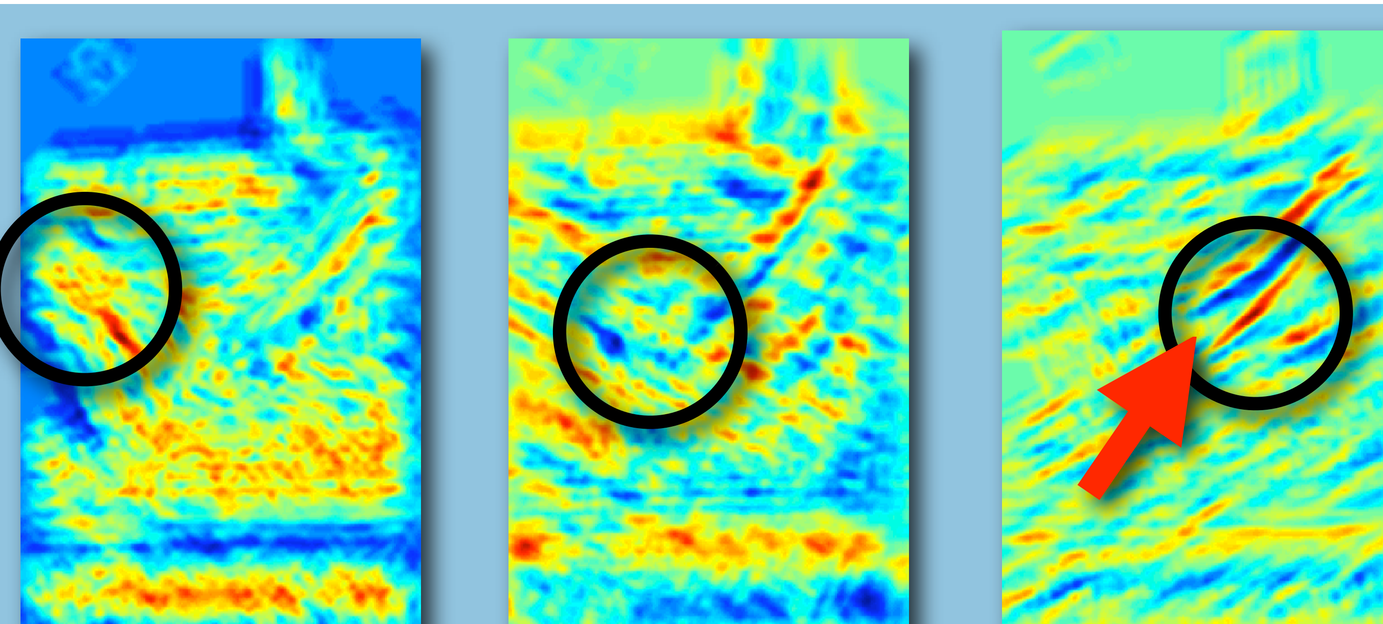
Input image

Our part likelihoods



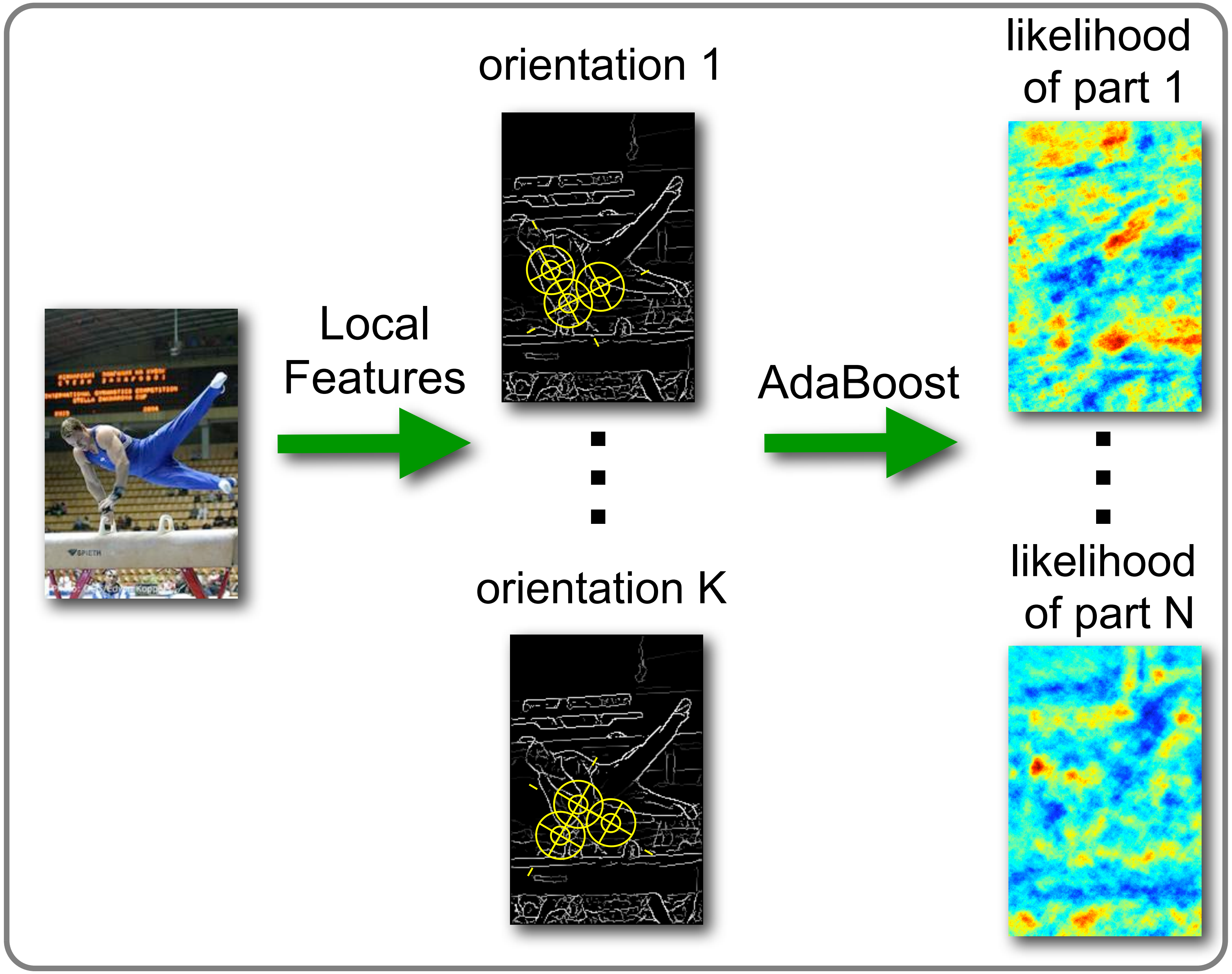
[Ramanan, NIPS'06]



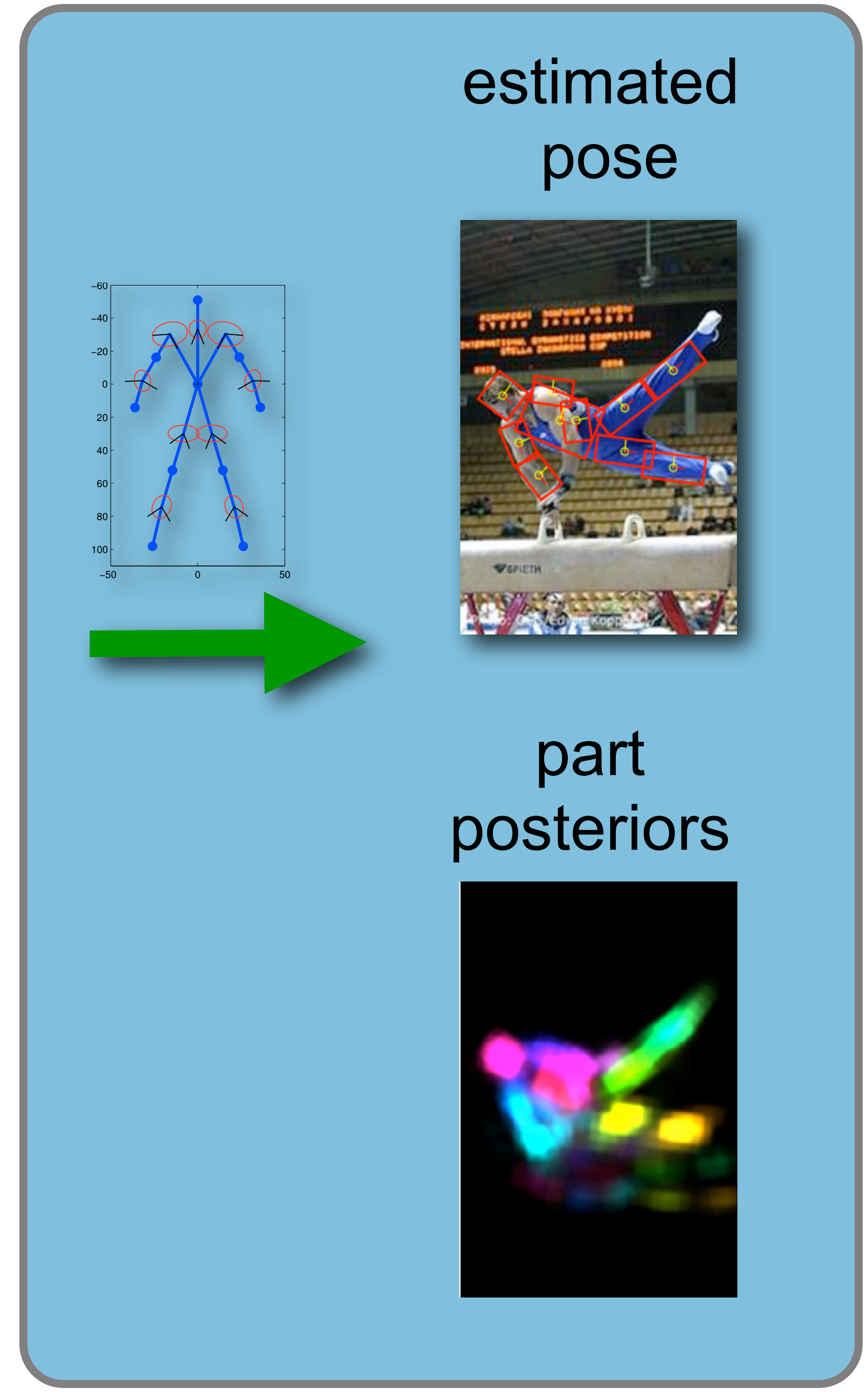


Model Components

Appearance Model:



Prior and Inference:

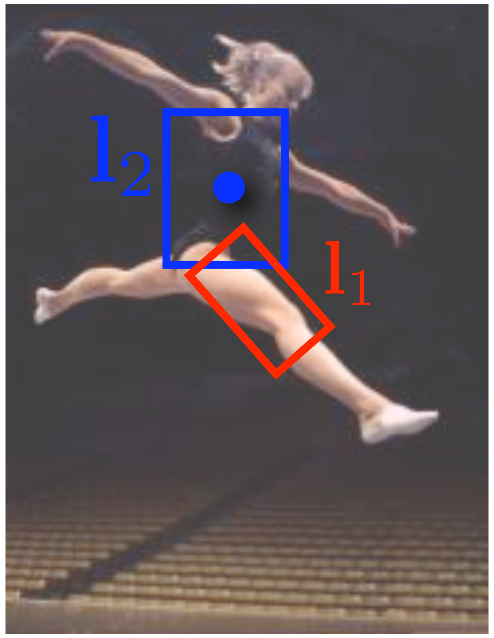


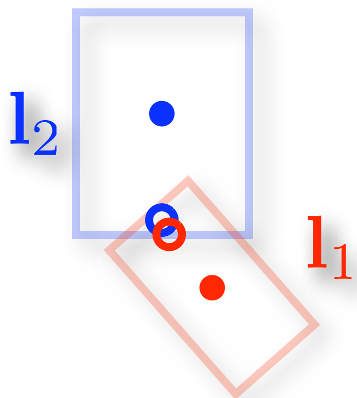
Kinematic Tree Prior

- Represent pairwise part relations
  [Felzenszwalb & Huttenlocher, IJCV'05]

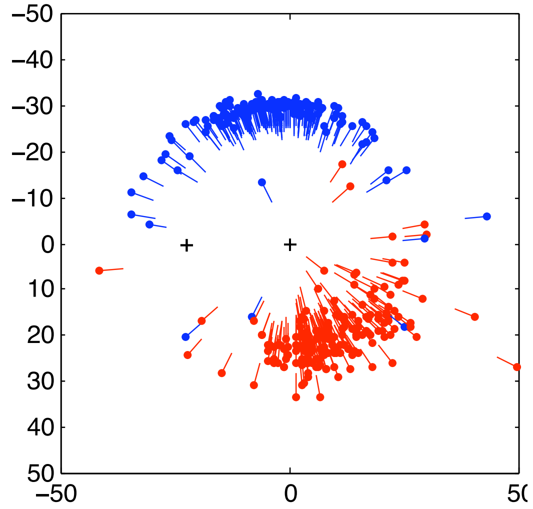
$$p(L) = p(\mathbf{l}_0) \prod_{(i,j)\in E} p(\mathbf{l}_i|\mathbf{l}_j),$$

$$p(\mathbf{l}_2|\mathbf{l}_1) = \mathcal{N}(T_{12}(\mathbf{l}_2)|T_{21}(\mathbf{l}_1), \Sigma^{12})$$

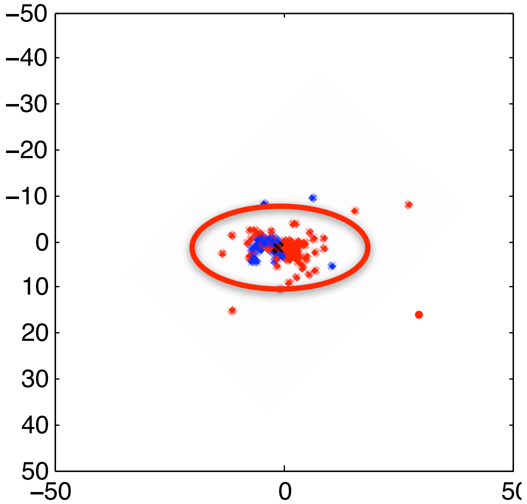




part locations relative to the joint



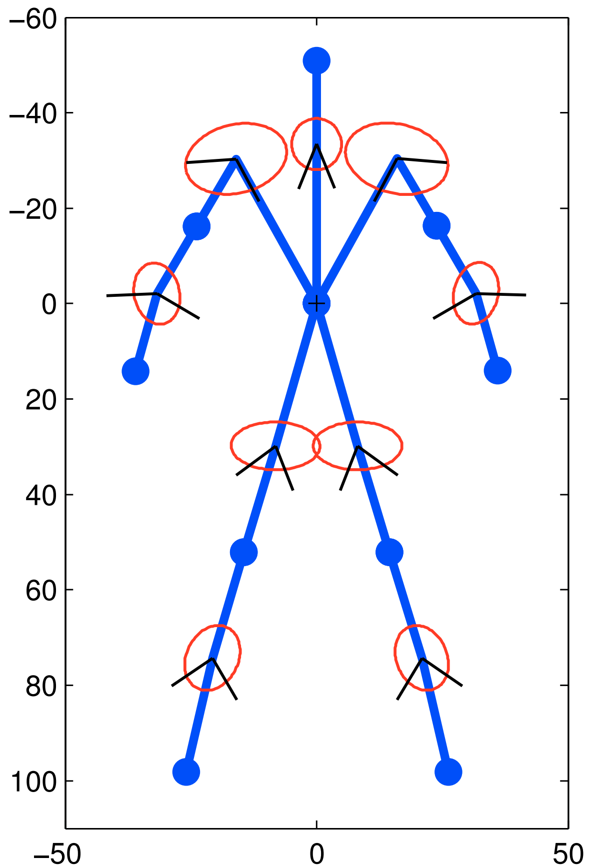
transformed part locations



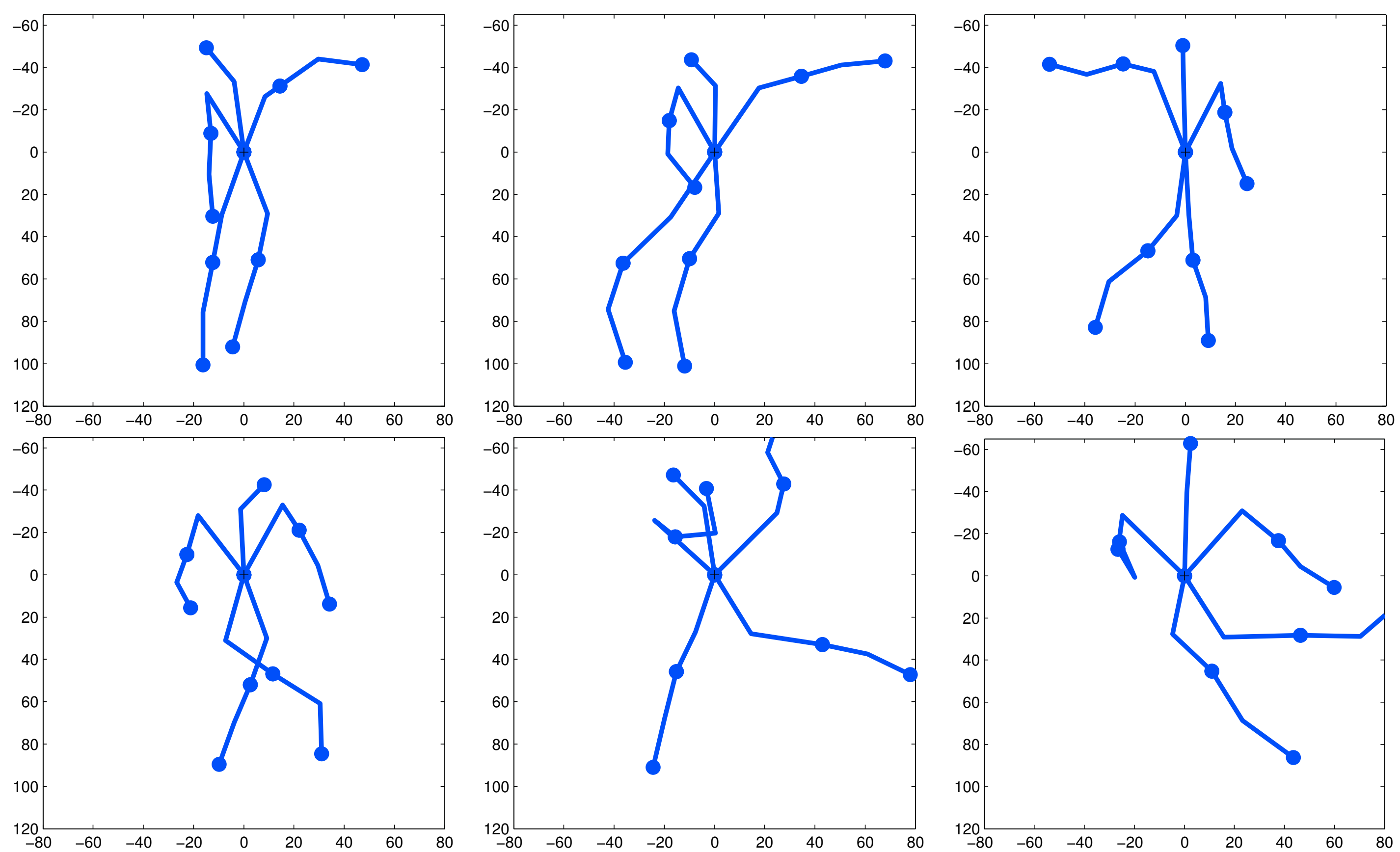
Kinematic Tree Prior

- Prior parameters: $\{T_{ij}, \Sigma^{ij}\}$
- Parameters of the prior are estimated with maximum likelihood

mean pose

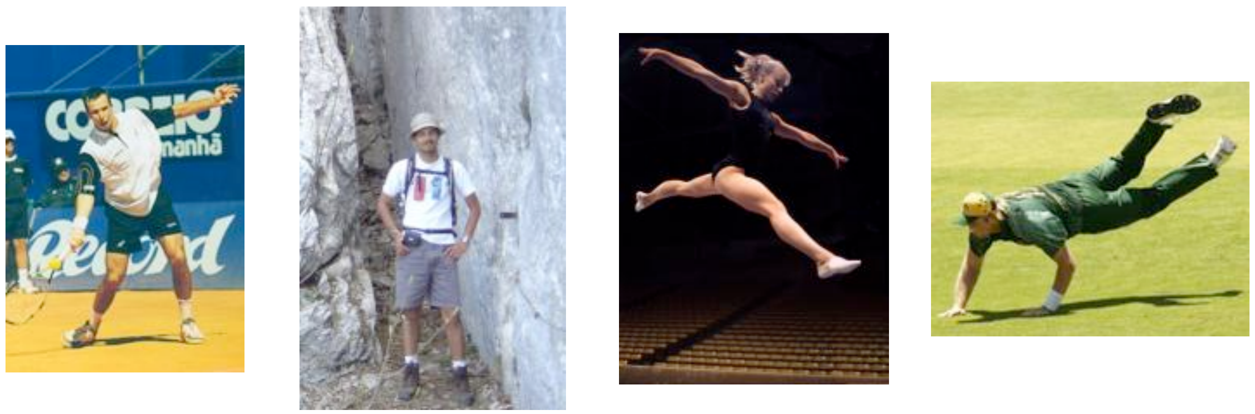


several independent samples



Evaluation Scenarios

1. Human Pose Estimation
   "People" dataset
   [Ramanan, NIPS'06]



2. Upper-body Pose Estimation
   "Buffy" dataset
   [Ferrari et al., CVPR'08]

3. Pedestrian Detection
   "TUD Pedestrians" dataset
   [Andriluka et al., CVPR'08]

Evaluation Scenarios

1. Human Pose Estimation
   "People" dataset
   [Ramanan, NIPS'06]

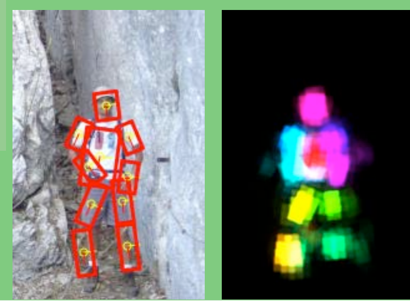
2. Upper-body Pose Estimation
   "Buffy" dataset
   [Ferrari et al., CVPR'08]

3. Pedestrian Detection
   "TUD Pedestrians" dataset
   [Andriluka et al., CVPR'08]

Scenario 1: Qualitative Results

Our model

8/10



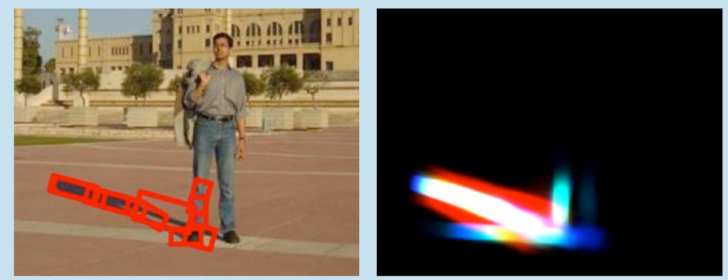
8/10

7/10

[Ramanan, NIPS'06]

7/10

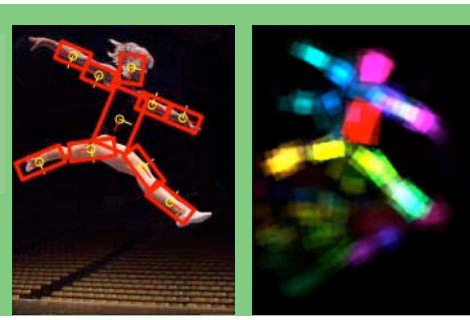
0/10



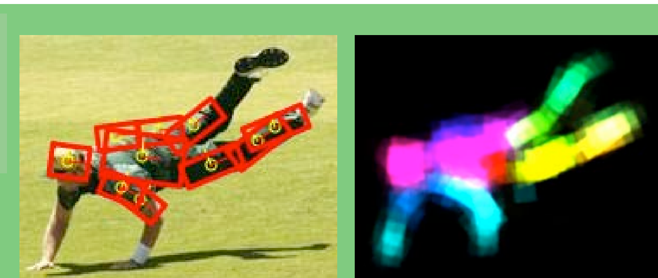
3/10

Our model

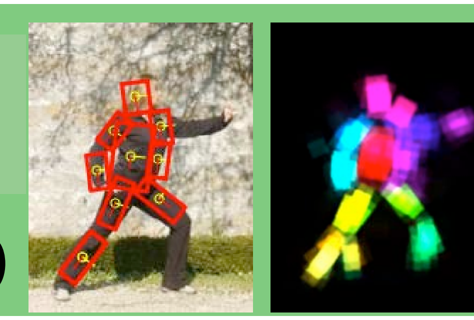
8/10



6/10

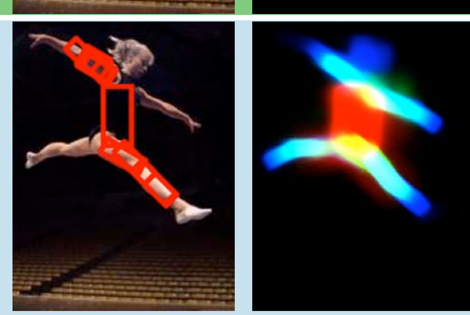


7/10

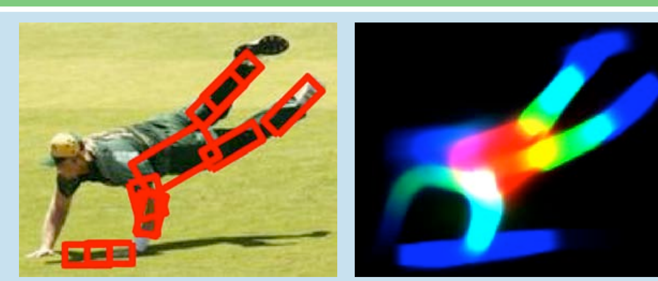


[Ramanan, NIPS'06]

4/10



3/10



3/10

Scenario 1: Quantitative Results

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	Torso	Upper legs	Lower legs	Upper arm	Forearm	Head	Total
[Ramanan, NIPS'06] 2nd parse	52	30	29	17	13	37	27
Our inference, edge features from [Ramanan, NIPS'06]	63	48	37	26	20	45	37
Our part detectors (SC)	29	12	18	3	4	40	14
Our prior, our part detectors (SC)	**81**	**63**	**55**	**47**	**31**	**75**	**55**
Our prior, our part detectors (SIFT)	78	58	54	44	31	66	52

Scenario 1: Quantitative Results

Method	Torso	Upper legs	Lower legs	Upper arm	Forearm	Head	Total
[Ramanan, NIPS'06] 2nd parse	52	30	29	17	13	37	27
Our prior, edge features from [Ramanan, NIPS'06]	63	48	37	26	20	45	37
Our part detectors (SC)	29	12	18	3	4	40	14
Our prior, our part detectors (SC)	**81**	**63**	**55**	**47**	**31**	**75**	**55**
Our prior, our part detectors (SIFT)	78	58	54	44	31	66	52

Scenario 1: Quantitative Results

Method	Torso	Upper legs	Lower legs	Upper arm	Forearm	Head	Total
[Ramanan, NIPS'06] 2nd parse	52	30	29	17	13	37	27
Our inference, edge features from [Ramanan, NIPS'06]	63	48	37	26	20	45	37
Our part detectors (SC)	29	12	18	3	4	40	14
Our prior, our part detectors (SC)	**81**	**63**	**55**	**47**	**31**	**75**	**55**
Our prior, our part detectors (SIFT)	78	58	54	44	31	66	52

SC = Shape Context

Scenario 1: Quantitative Results

Method	Torso	Upper legs	Lower legs	Upper arm	Forearm	Head	Total
[Ramanan, NIPS'06] 2nd parse	52	30	29	17	13	37	27
Our inference, edge features from [Ramanan, NIPS'06]	63	48	37	26	20	45	37
Our part detectors (SC)	29	12	18	3	4	40	14
Our prior, our part detectors (SC)	**81**	**63**	**55**	**47**	**31**	**75**	**55**
Our prior, our part detectors (SIFT)	78	58	54	44	31	66	52

SC = Shape Context

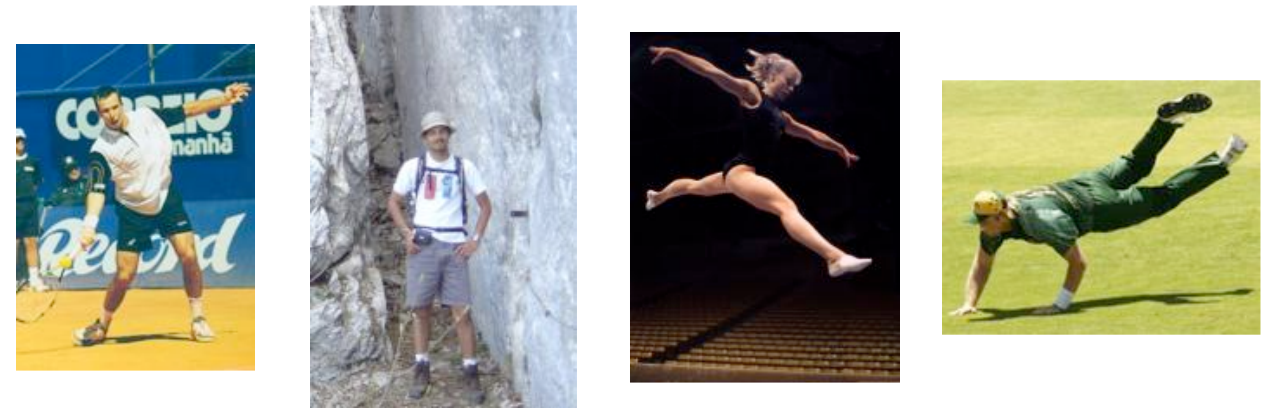
Scenario 1: Quantitative Results

Method	Torso	Upper legs	Lower legs	Upper arm	Forearm	Head	Total
[Ramanan, NIPS'06] 2nd parse	52	30	29	17	13	37	27
Our inference, edge features from [Ramanan, NIPS'06]	63	48	37	26	20	45	37
Our part detectors (SC)	29	12	18	3	4	40	14
Our prior, our part detectors (SC)	**81**	**63**	**55**	**47**	**31**	**75**	**55**
Our prior, our part detectors (SIFT)	78	58	54	44	31	66	52

SC = Shape Context

Evaluation Scenarios

1. Human Pose Estimation
   “People” dataset
   [Ramanan, NIPS’06]



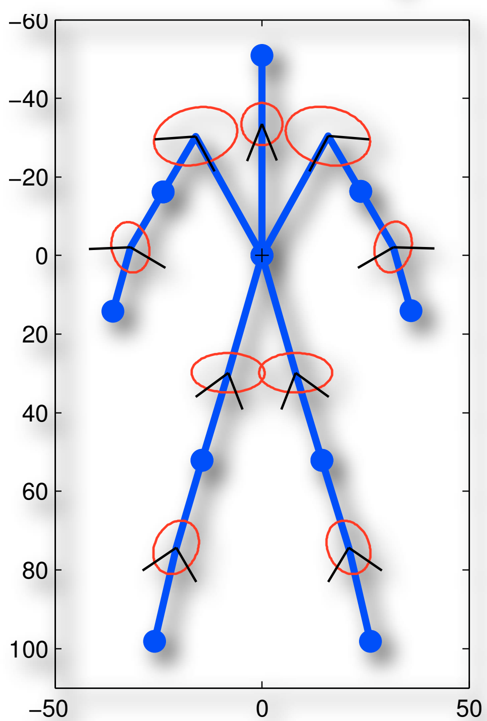
2. Upper-body Pose Estimation
   “Buffy” dataset
   [Ferrari et al., CVPR’08]

3. Pedestrian Detection
   “TUD Pedestrians” dataset
   [Andriluka et al., CVPR’08]

Estimated upper-body poses

Quantitative Results

Method	Torso	Upper arm	Lower arm	Head	Total
[Ferrari et al. CVPR'08]	-	-	-	-	57.9
detectors only	18.9	6.8	3.1	47.2	14.3
full model	**90.7**	79.3	41.2	**95.9**	71.3

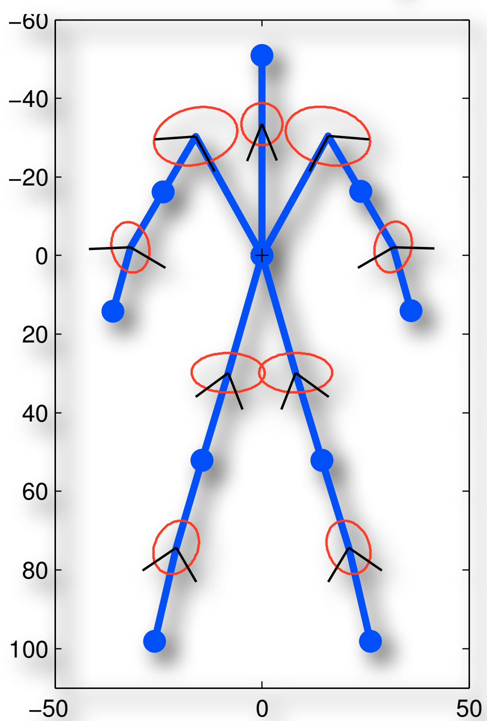


- generic model
- prior and appearance learned on the "People" dataset

Quantitative Results

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	Torso	Upper arm	Lower arm	Head	Total
[Ferrari et al. CVPR'08]	-	-	-	-	57.9
detectors only	18.9	6.8	3.1	47.2	14.3
full model	**90.7**	79.3	41.2	**95.9**	71.3
[Ferrari et al. CVPR'09]	-	-	-	-	**72.2**

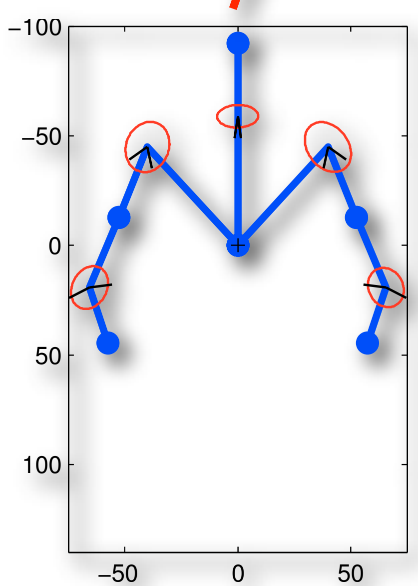


- generic model
- prior and appearance learned on the "People" dataset

Quantitative Results

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	Torso	Upper arm	Lower arm	Head	Total
[Ferrari et al. CVPR'08]	-	-	-	-	57.9
detectors only	18.9	6.8	3.1	47.2	14.3
full model	**90.7**	79.3	41.2	**95.9**	71.3
[Ferrari et al. CVPR'09]	-	-	-	-	72.2
full model, Buffy pose prior	**90.7**	**81.35**	**46.5**	95.5	**73.5**

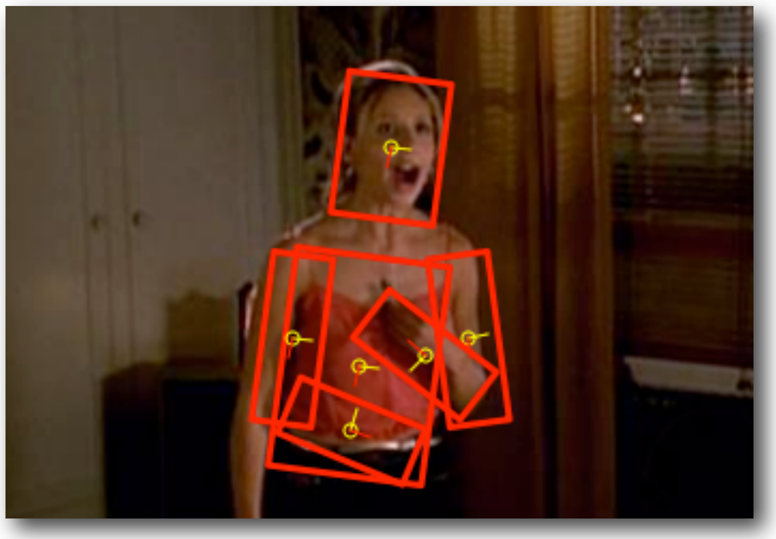


- specialized upper body prior
- appearance learned on the "People" dataset

Typical Failure Cases

Foreshortening

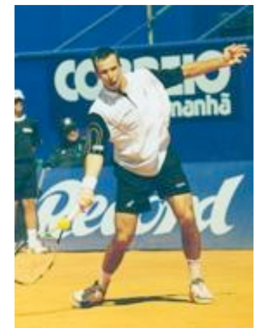
Part occlusion



Detections on other body parts

Evaluation Scenarios

1. Human Pose Estimation
   “People” dataset
   [Ramanan, NIPS’06]




2. Upper-body Pose Estimation
   “Buffy” dataset
   [Ferrari et al., CVPR’08]



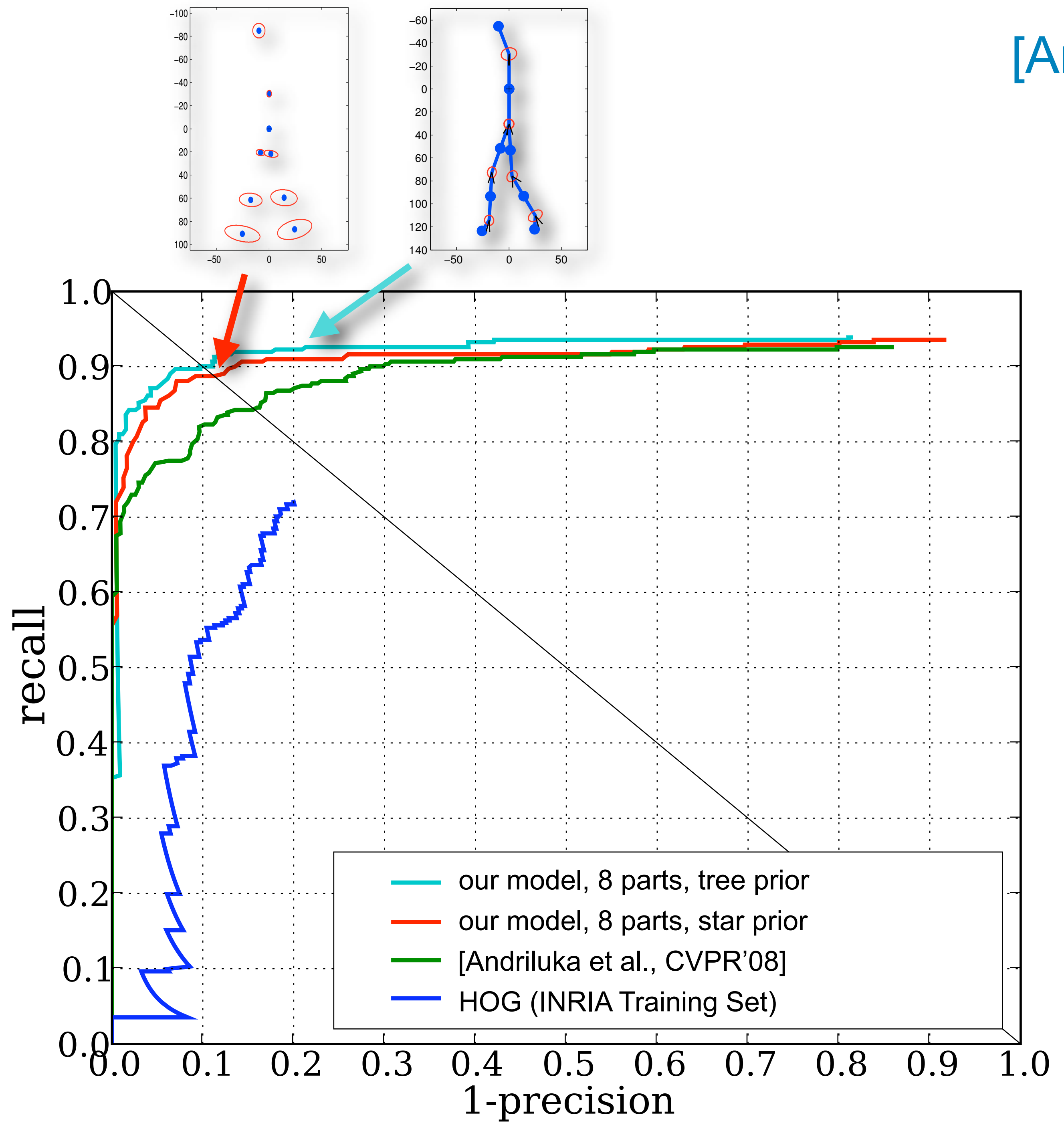
3. Pedestrian Detection
   “TUD Pedestrians” dataset
   [Andriluka et al., CVPR’08]



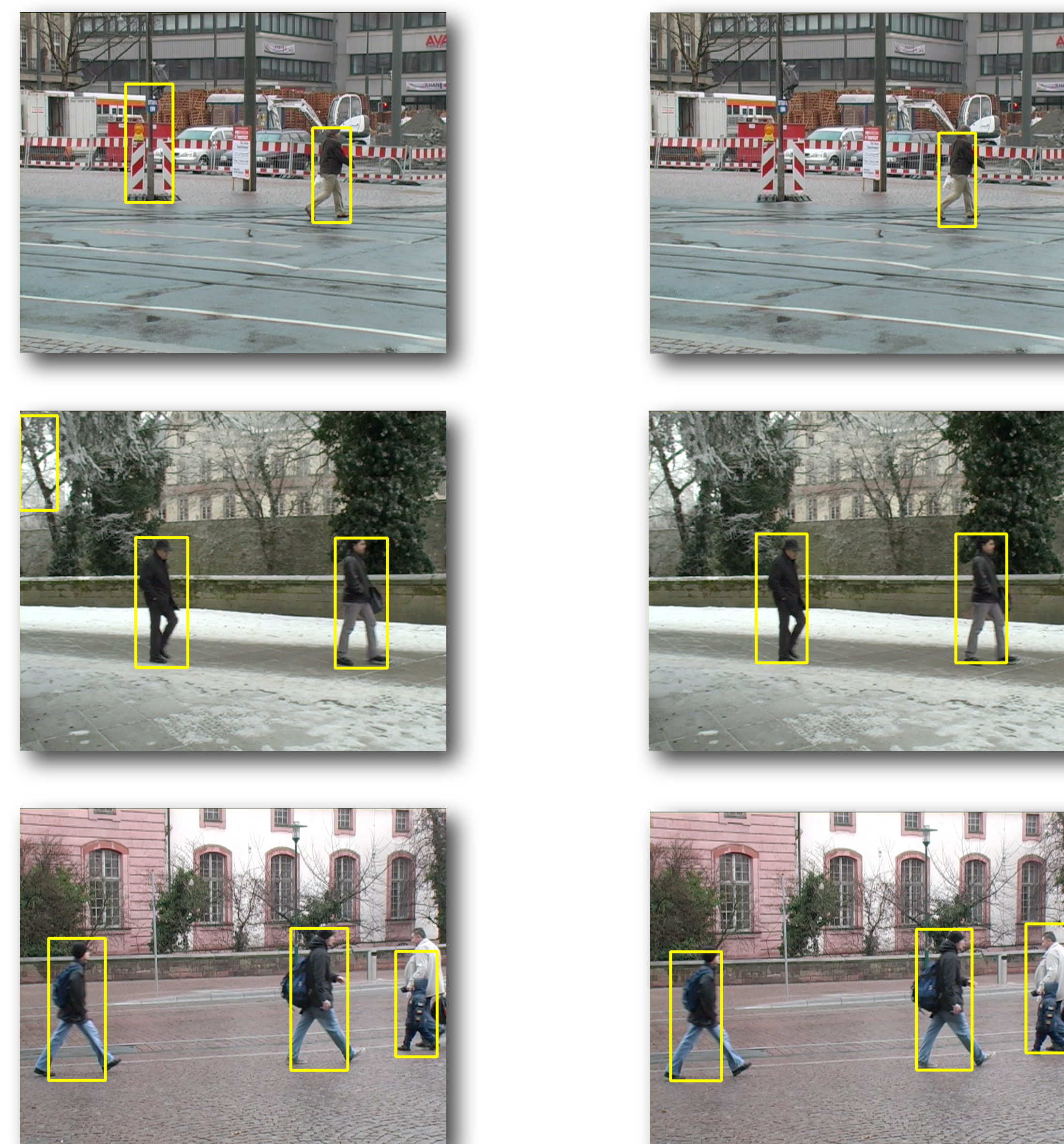
People Detection: Results

TECHNISCHE UNIVERSITÄT DARMSTADT

- Comparison with state-of-the art in people detection



[Andriluka et al., CVPR'08] This work



Conclusion & Future Work

TECHNISCHE UNIVERSITÄT DARMSTADT

- Success of pose estimation by "body part" detection
  - use well understood pose estimation framework (Pictorial Structures)
  - use appropriate representation for kinematic dependencies
  - use state of the art appearance representation (SIFT, SC) and classification (AdaBoost)
- Next steps:
  - estimate poses in 3D
  - part occlusions
  - appearance constraints between parts

Thanks!

- Acknowledgements:
  - Thanks to Krystian Mikolajczyk for image descriptors code
  - Thanks to Christian Wojek for AdaBoost code and helpful suggestion
  - Thanks to Deva Ramanan and Vittorio Ferrari for making their code and datasets publicly available
  - This work is partially funded by German Research Foundation (DFG) through GRK 1362.

- Code and pre-trained models will be available at:
  - http://www.mis.informatik.tu-darmstadt.de/code