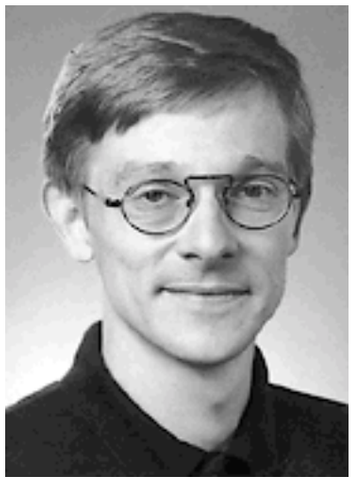
Monocular 3D Pose Estimation and Tracking by Detection



Mykhaylo Andriluka [1]

Stefan Roth [1]

Bernt Schiele [1,2]

[1] Department of Computer Science
TU Darmstadt

[2] MPI Informatics
Saarbrücken

Motivation

- Goal: find people and estimate their poses in 3D from monocular potentially moving camera

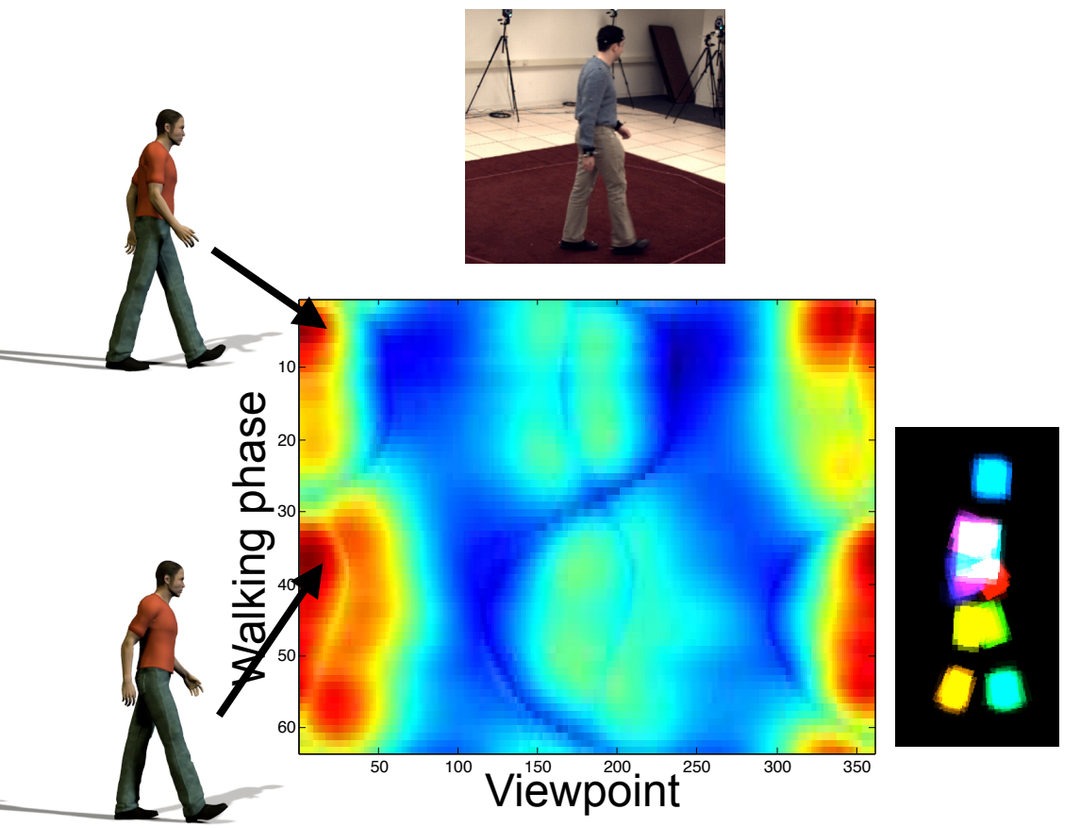
- Challenges:
  - dynamic background
  - multiple people
  - ambiguity in 2D to 3D lifting

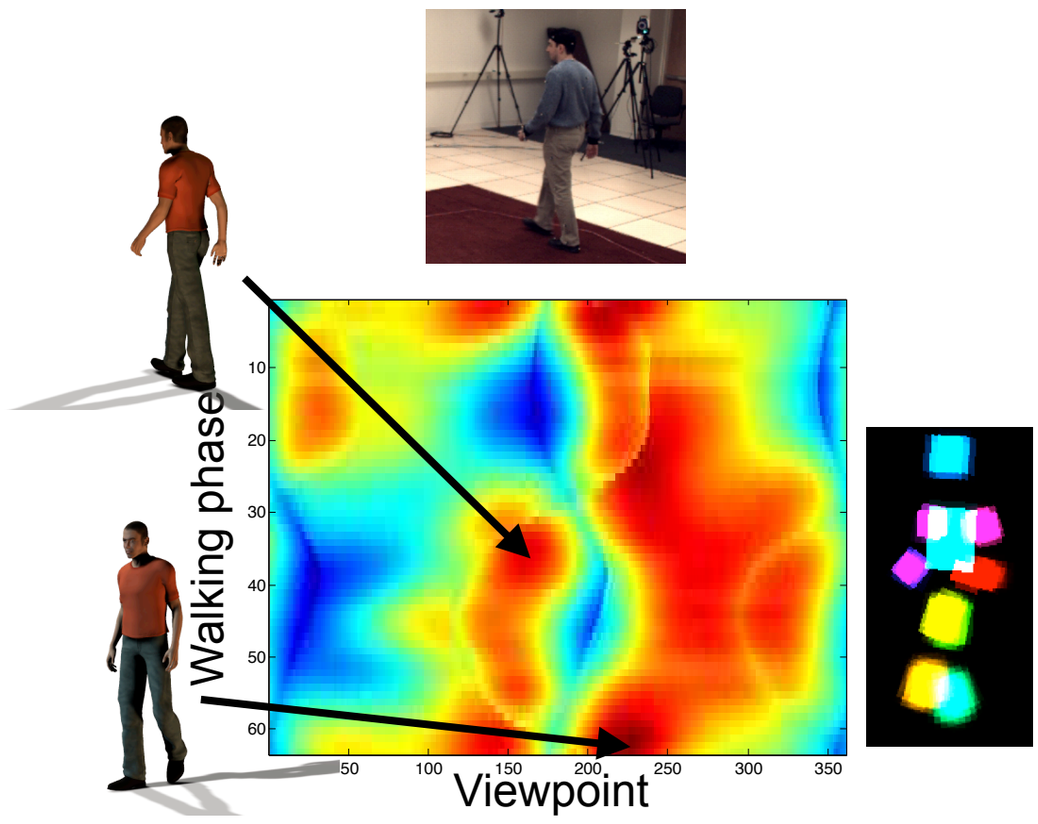
Motivation

- Goal: find people and estimate their poses in 3D from monocular potentially moving camera

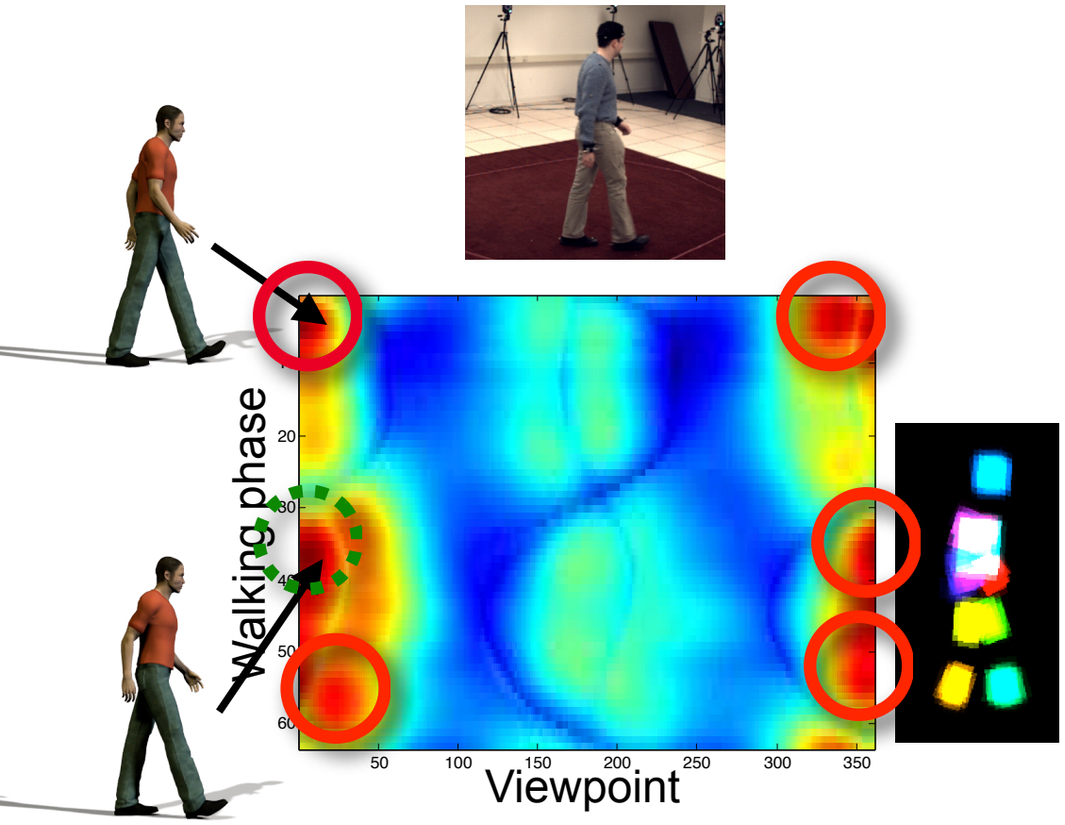
- Challenges:
  - dynamic background
  - multiple people
  - ambiguity in 2D to 3D lifting

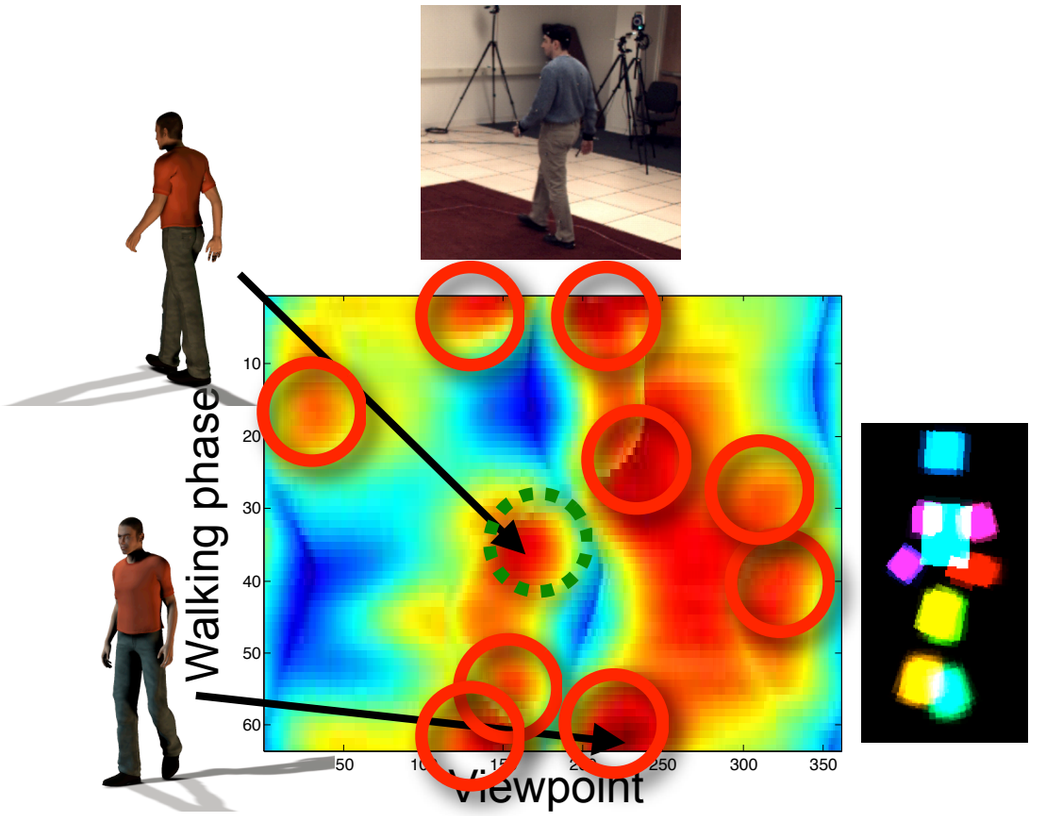
Towards Our Approach





Towards Our Approach

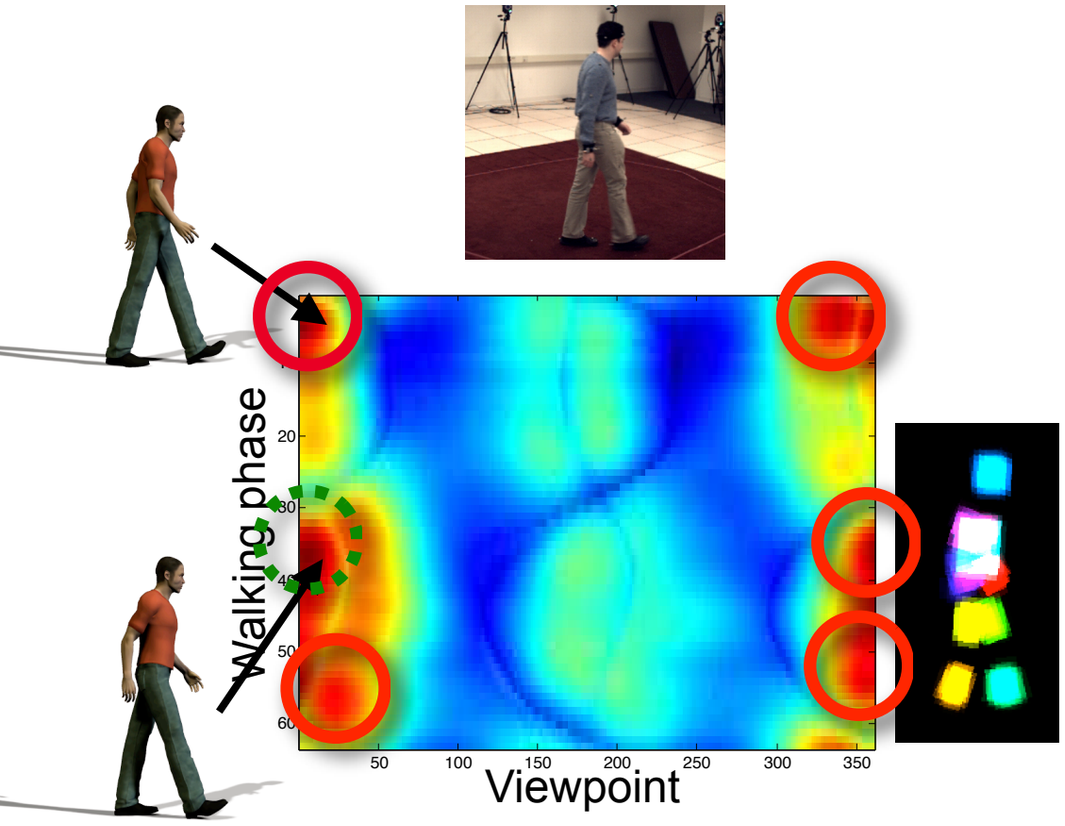


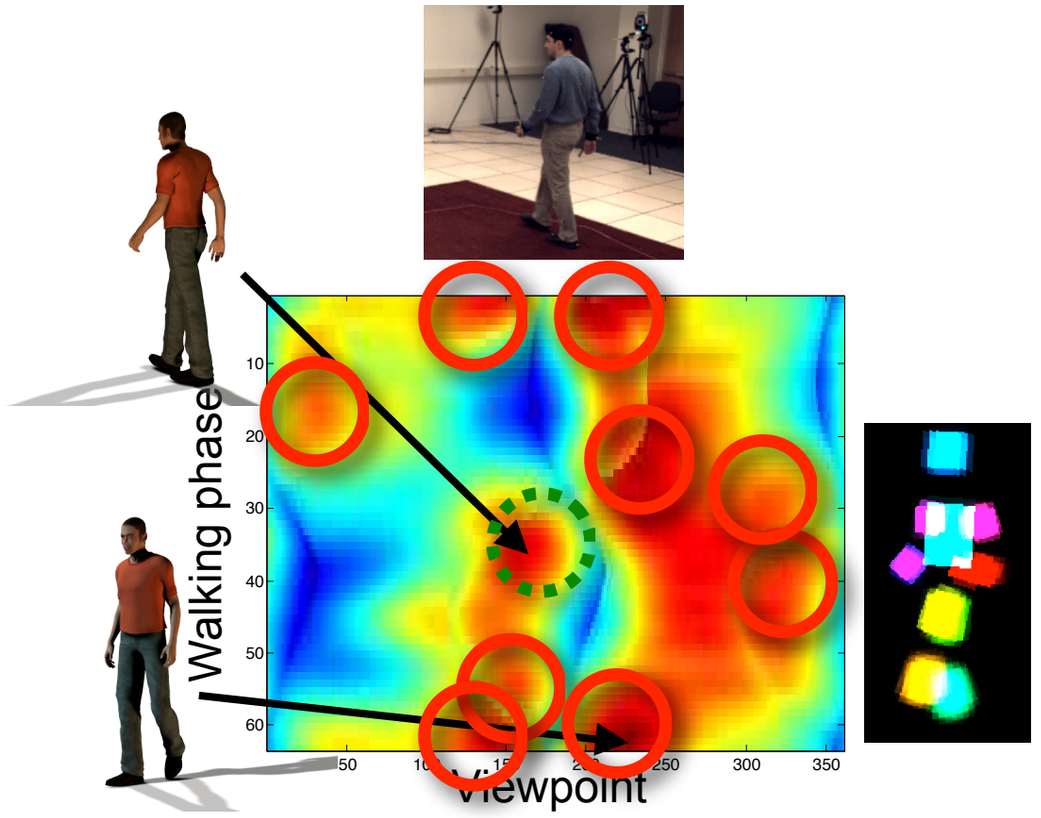


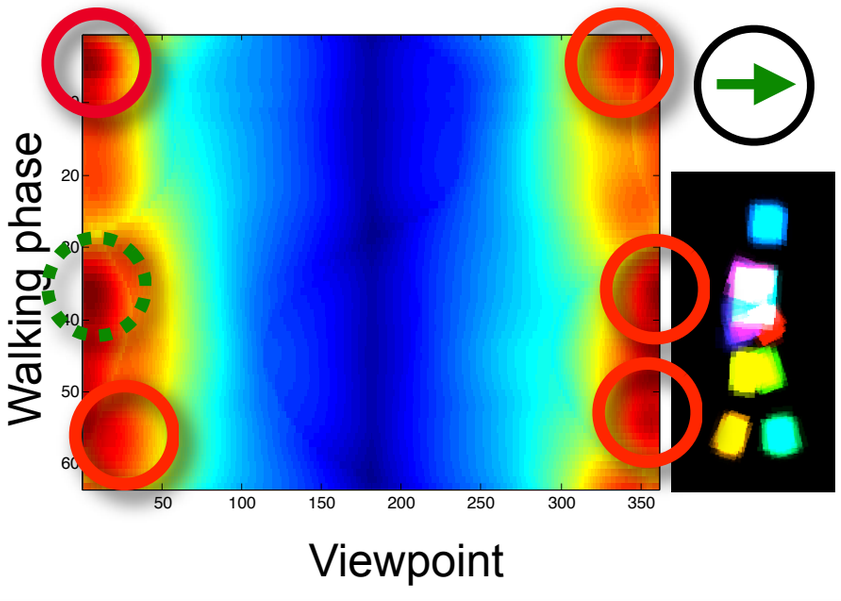
Towards Our Approach

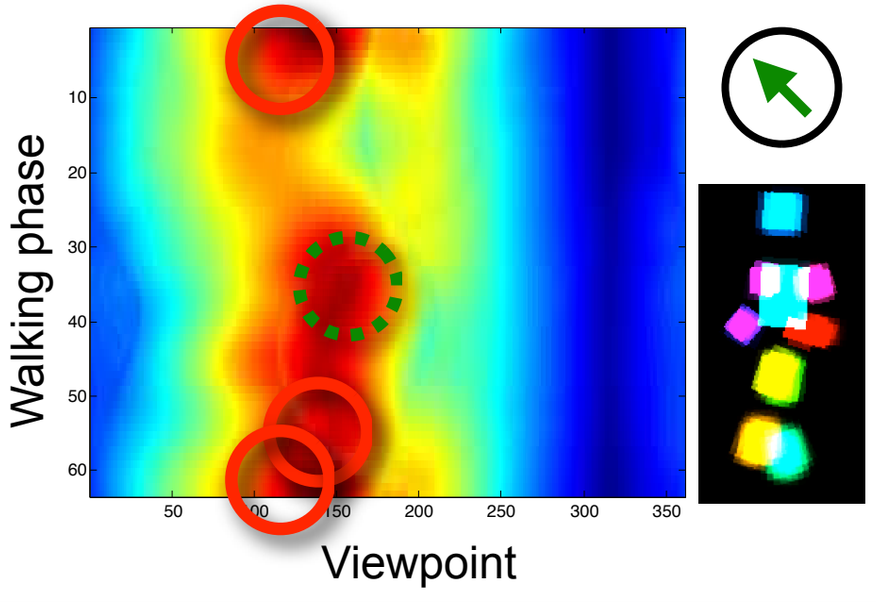
TECHNISCHE UNIVERSITÄT DARMSTADT

max planck institut informatik









Our Approach

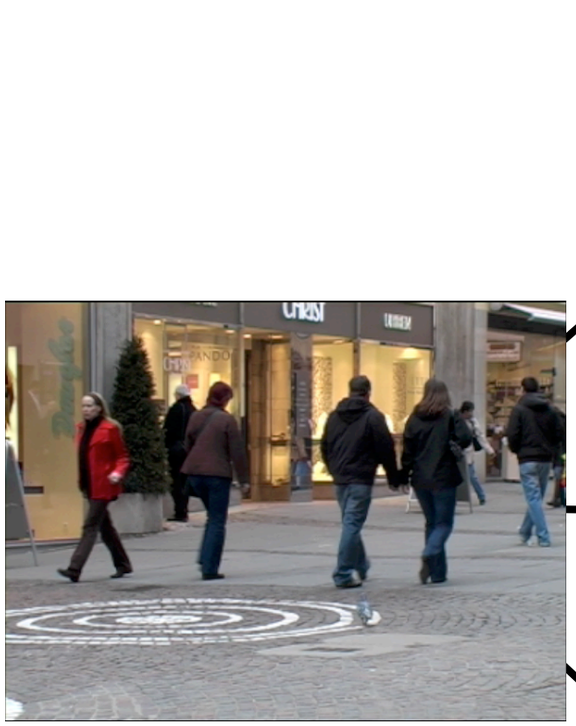
TECHNISCHE UNIVERSITÄT DARMSTADT

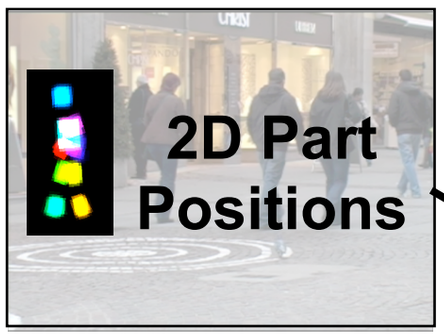
max planck institut informatik

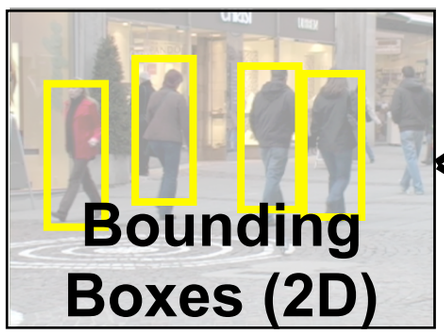
1. Single-frame detection

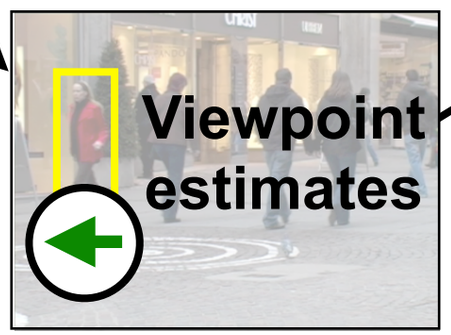
2. 2D-Tracklet detection

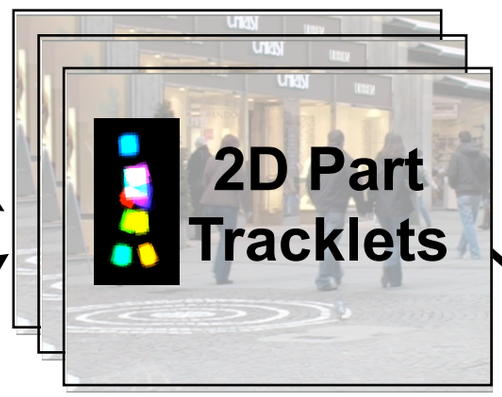
3. 2D-to-3D lifting (tracking & pose estimation)

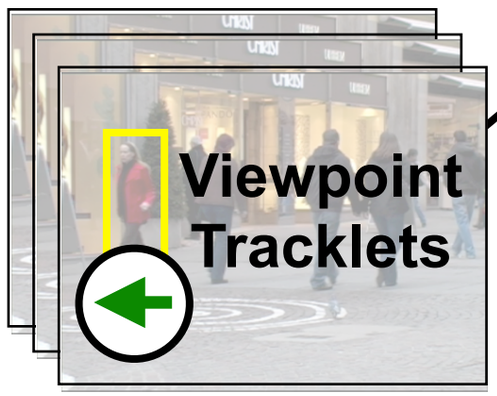


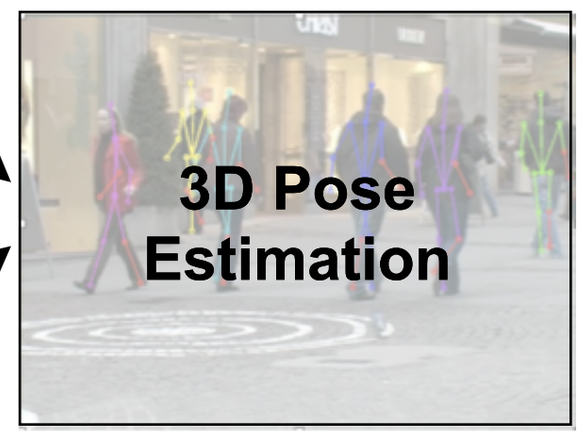












Our Approach

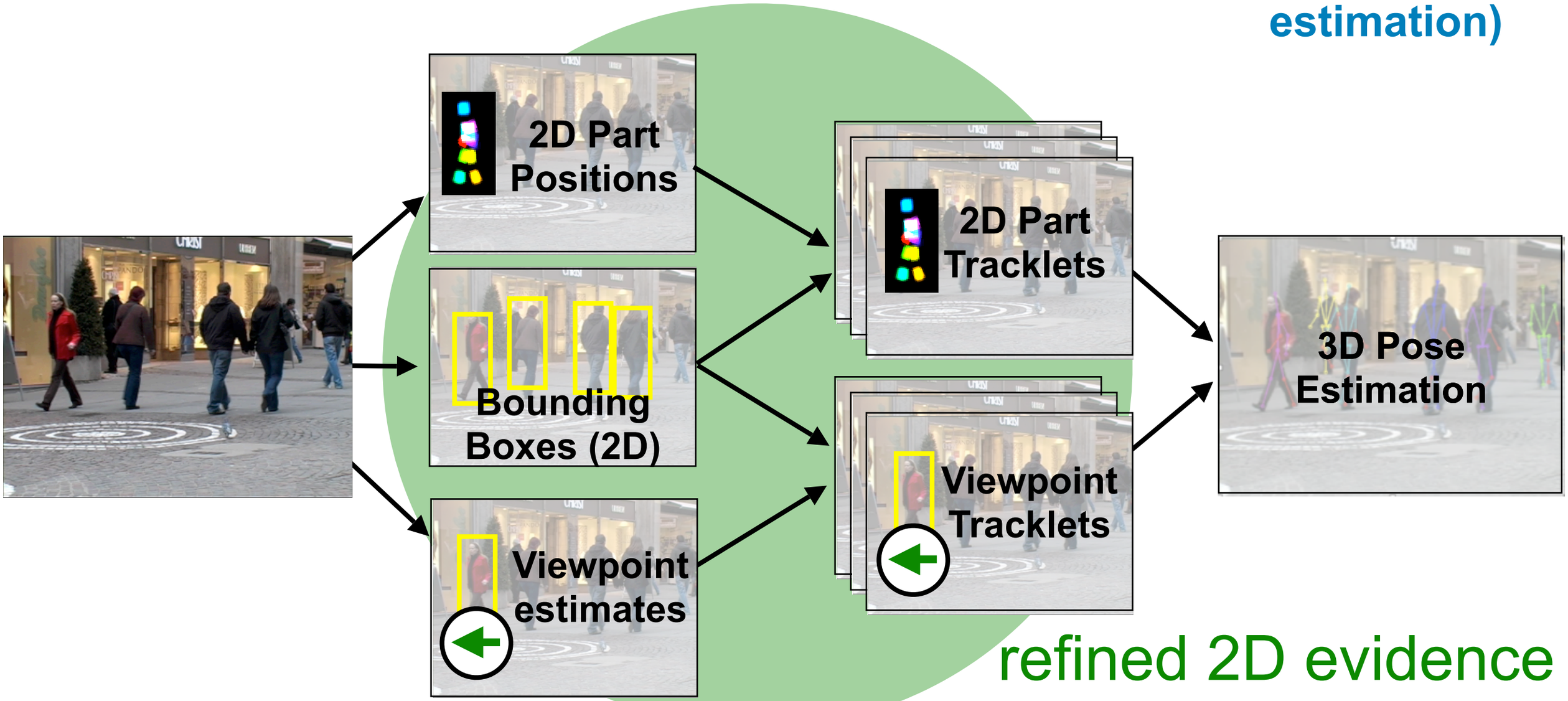
TECHNISCHE UNIVERSITÄT DARMSTADT

max planck institut informatik

1. Single-frame detection

2. 2D-Tracklet detection

3. 2D-to-3D lifting (tracking & pose estimation)



refined 2D evidence over several frames

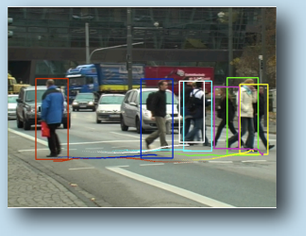
Related Work

1. Single-frame detection

People Detection

[Leibe et al., CVPR'05]
[Dalal&Triggs, CVPR'05]
[Wojek et al., CVPR'09]

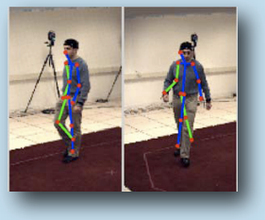
2. 2D-Tracklet detection



People Tracking

[Wu&Nevatia, IJCV'07]
[Andriluka et al., CVPR'08]
[Breitenstein et al., ICCV'09]

3. 2D-to-3D lifting (tracking & pose estimation)



3D Pose Estimation

[Urtasun et al., CVPR'06]
[Rogez et al., CVPR'08]
[Gall et al., IJCV'10]

2D Pose Estimation

[Felzenszwalb&Huttenlocher, IJCV'05]
[Sigal&Black, CVPR'06],
[Eichner&Ferrari, BMVC'09],
[Andriluka et al., CVPR'09]

Our Approach

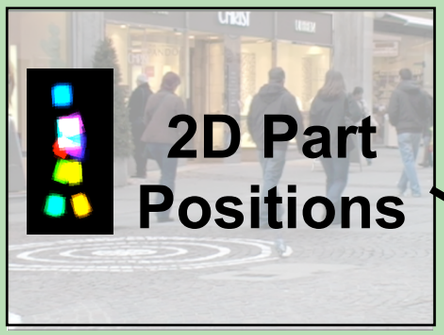
TECHNISCHE UNIVERSITÄT DARMSTADT

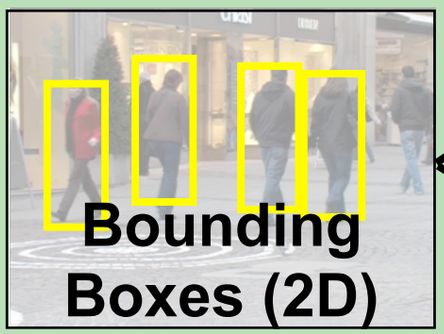
max planck institut informatik

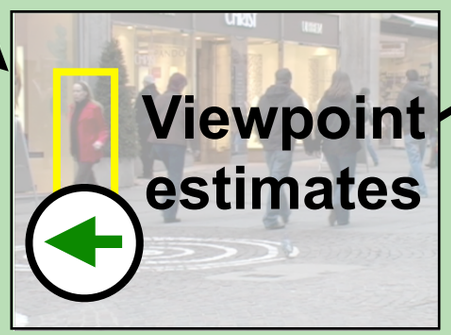
1. Single-frame detection

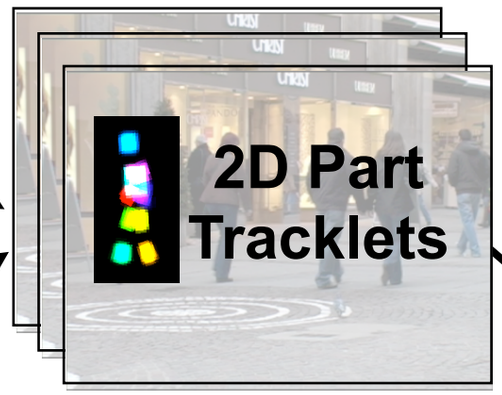
2. 2D-Tracklet detection

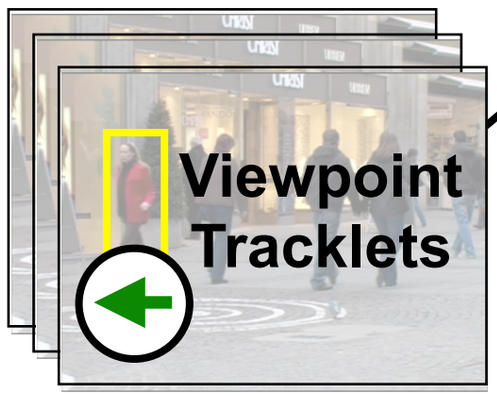
3. 2D-to-3D lifting (tracking & pose estimation)

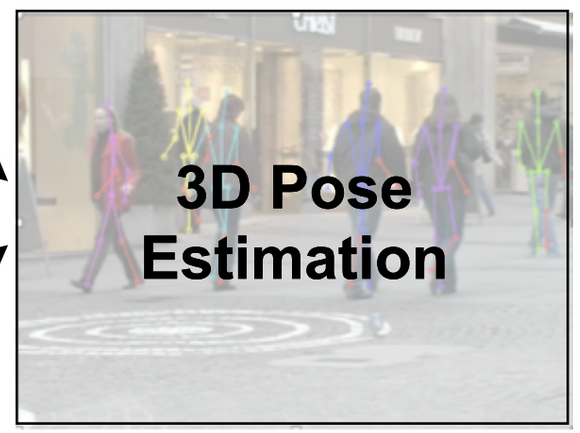












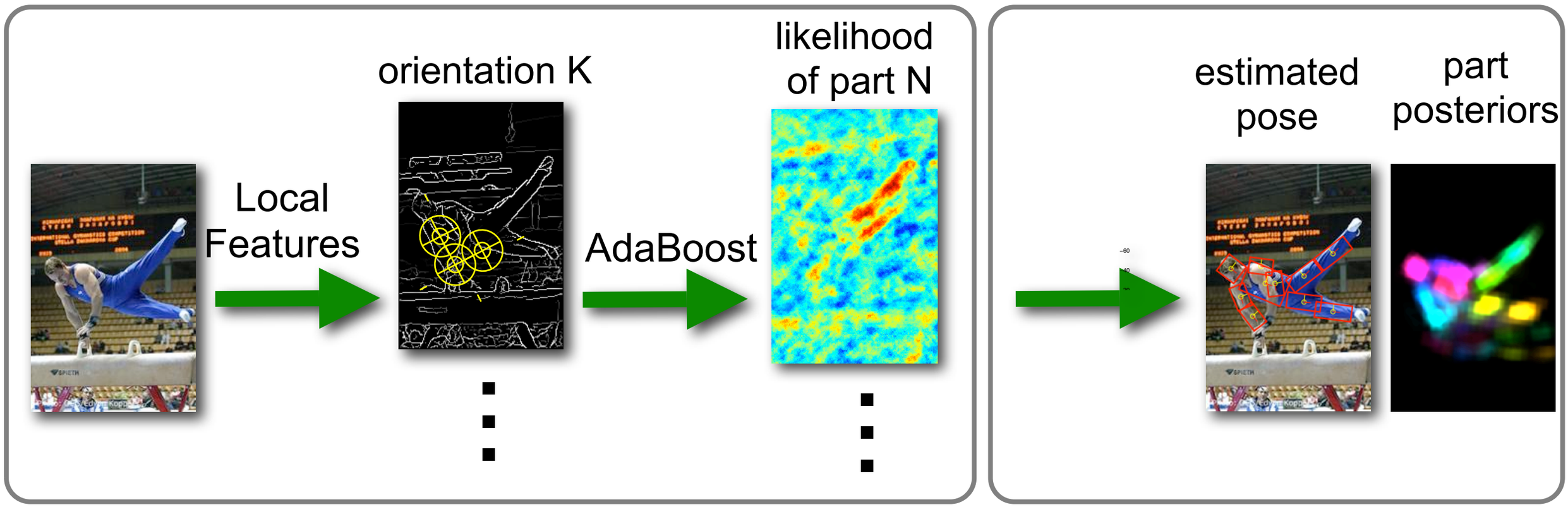
People Detection & 2D Pose Estimation

- Model from [Andriluka et al., CVPR'09]:

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

likelihood of observations (appearance model)

prior on body poses



Combining Pictorial Structures Models

TECHNISCHE UNIVERSITÄT DARMSTADT

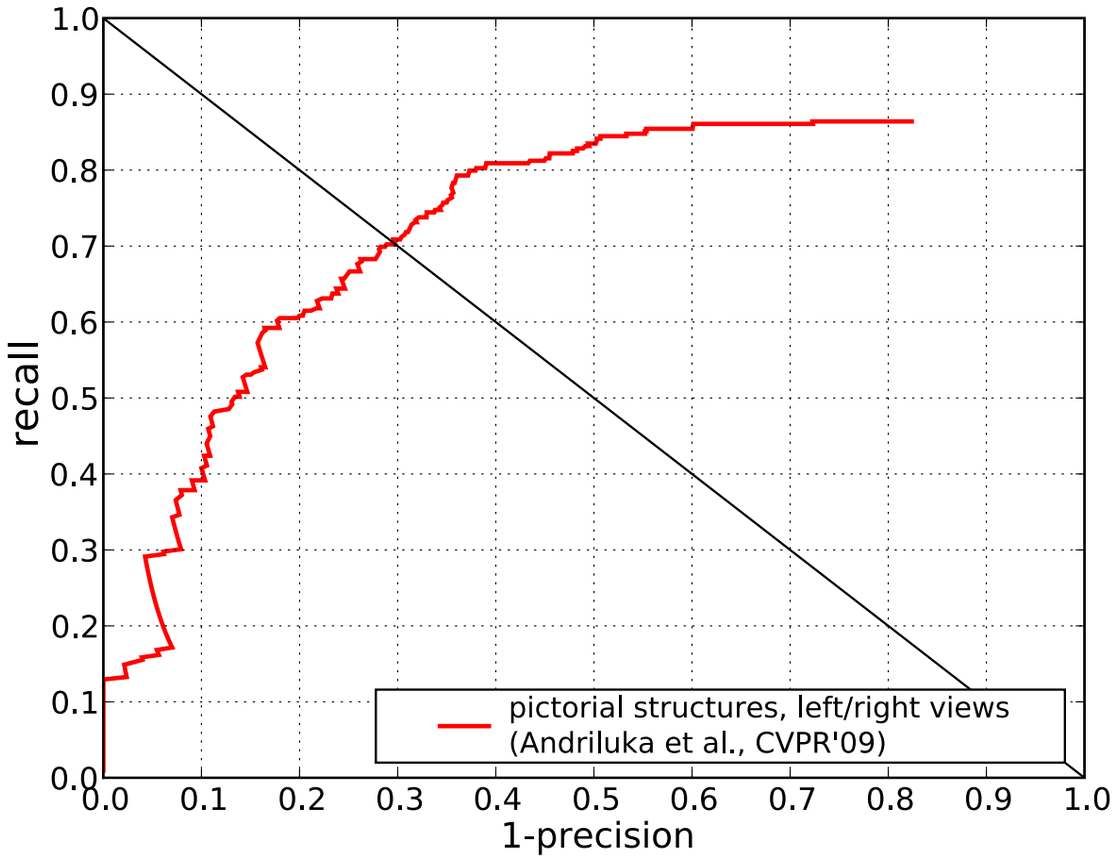
max planck institut informatik

- One pictorial structures model is good but 11 are better
- Each model can focus on specific part of the problem (in our case specific viewpoint)

New Dataset: TUD Multiview Pedestrians
- 2336 training images
- 496 validation and test images
- bounding box, body joints and viewpoint annotations

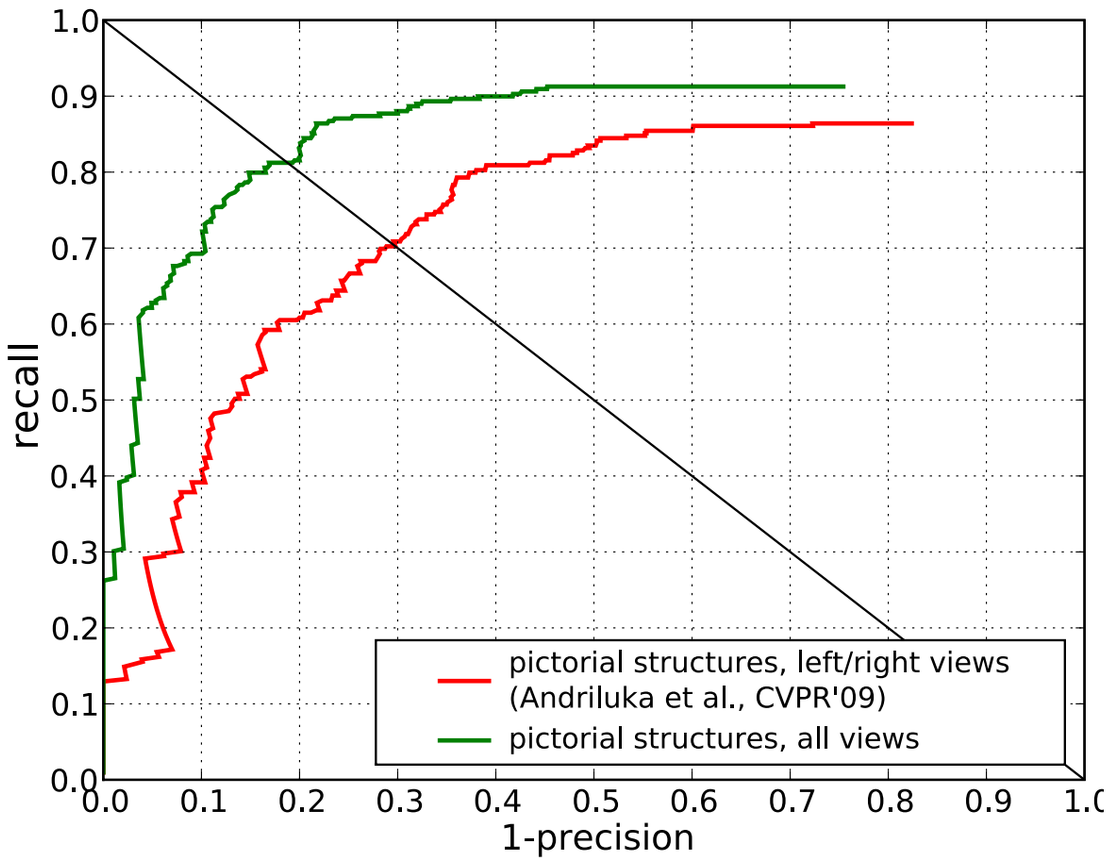
- Employ stacking approach:
  - for detection: train SVM classifier using output of PS models as features
  - for viewpoint classification: train 1-vs-all SVM classifier for each of the 8 viewpoints

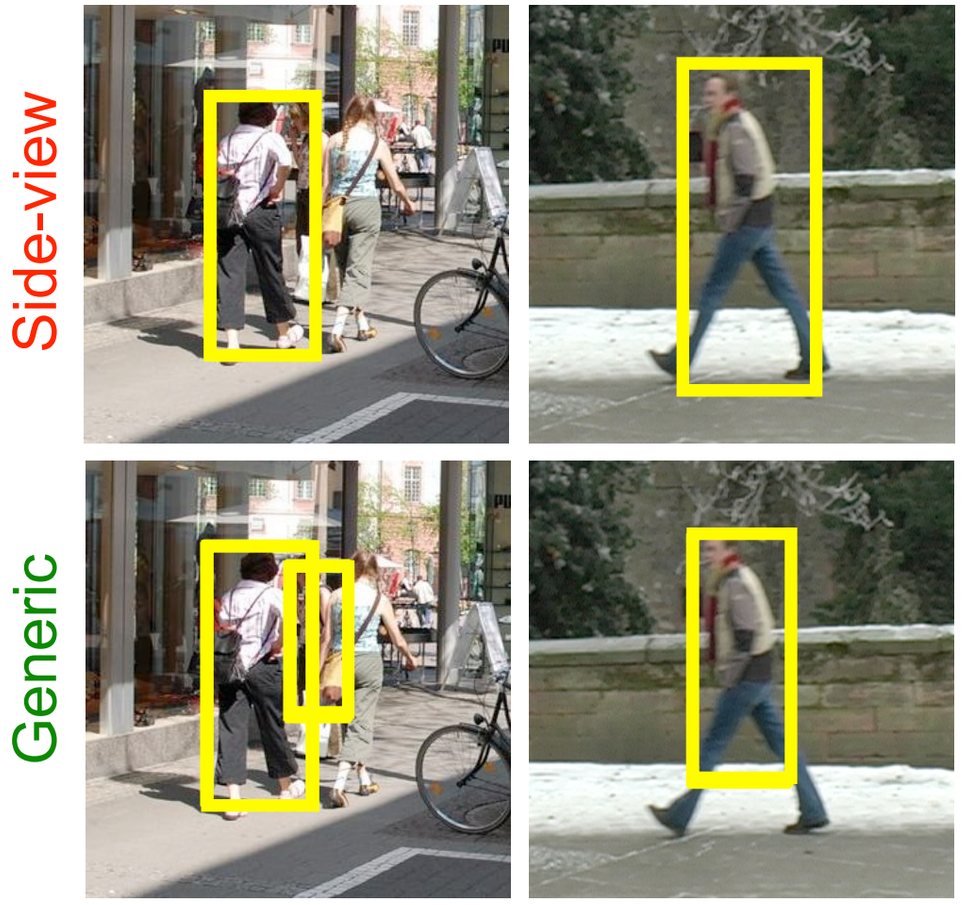
Results: Single-frame Detection



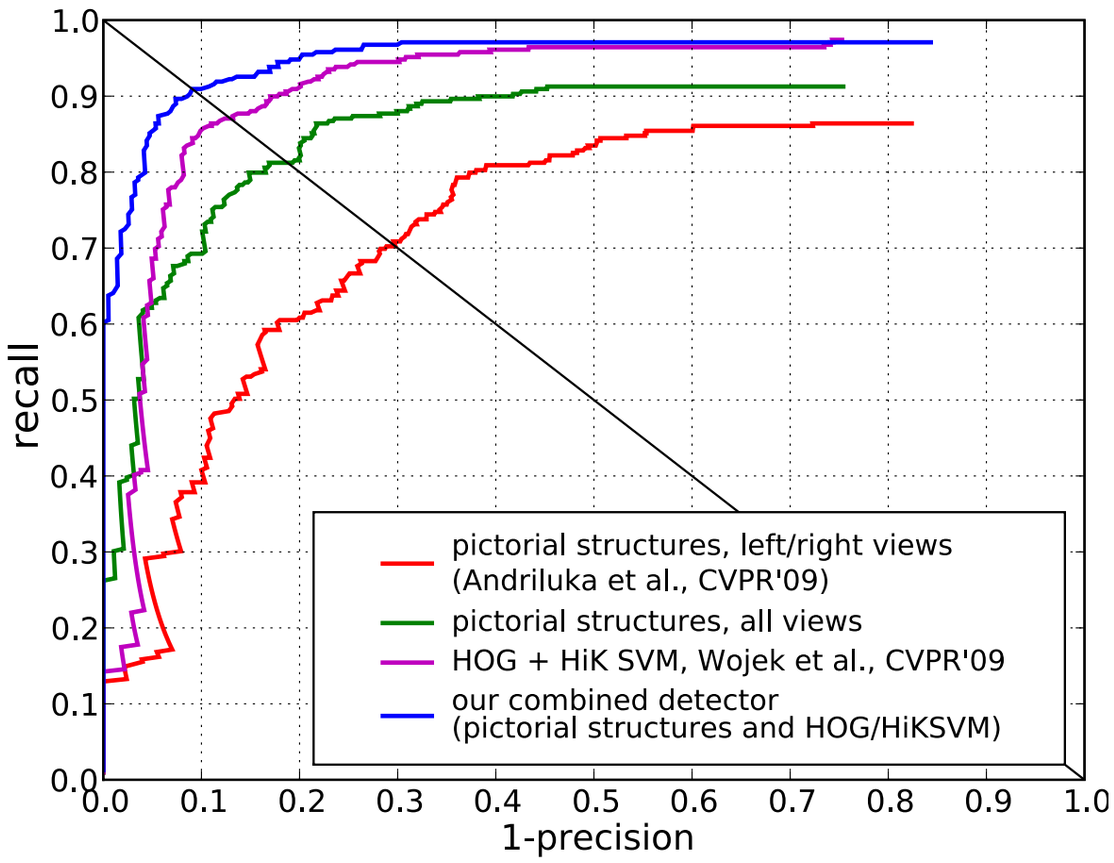
Side-view

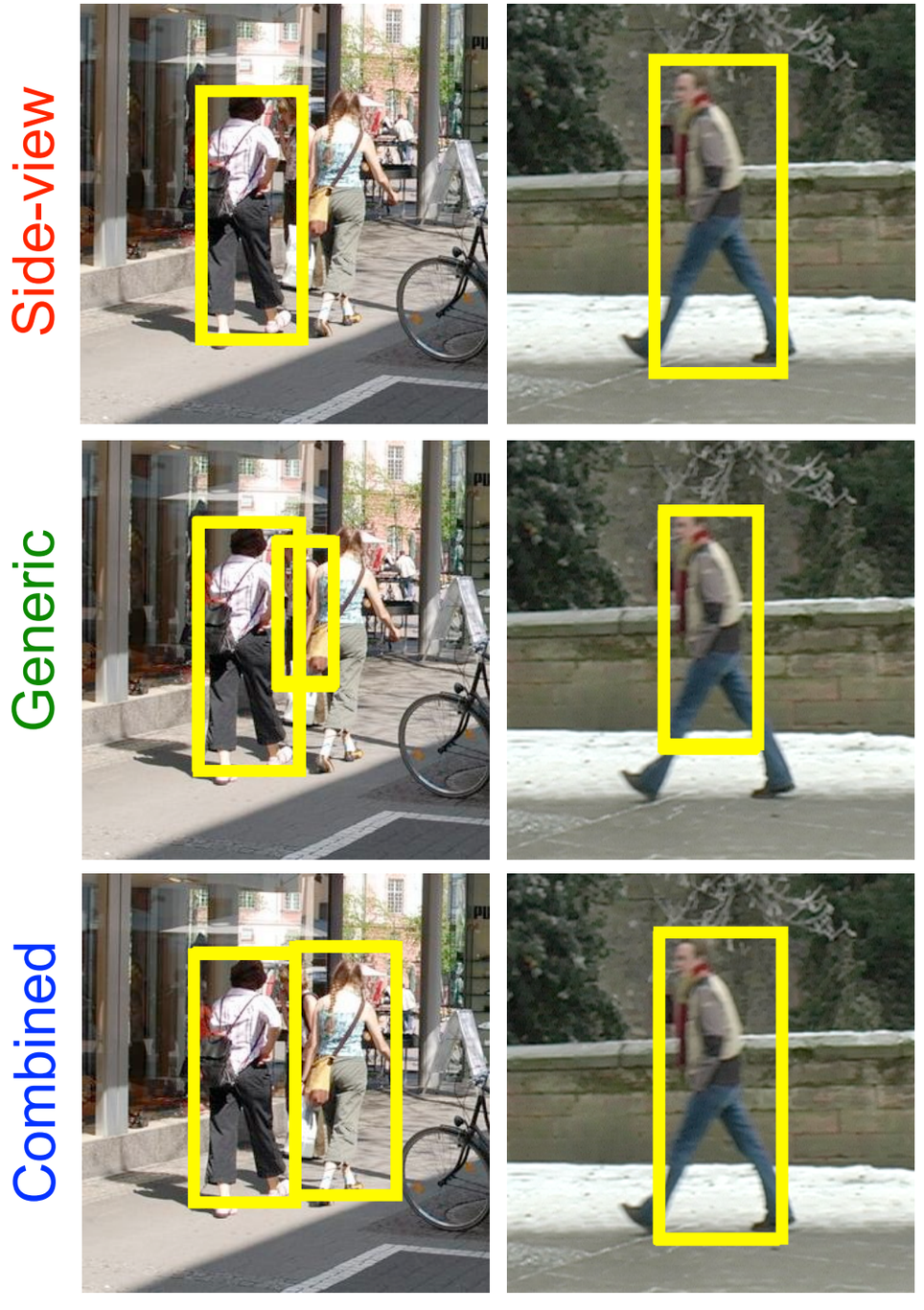
Results: Single-frame Detection





Results: Single-frame Detection





Results: Single-frame Viewpoint Estimation

- Viewpoint classification accuracy:
  - 42.2% in classification of one of the eight viewpoints

- Typical false positives:

Our Approach

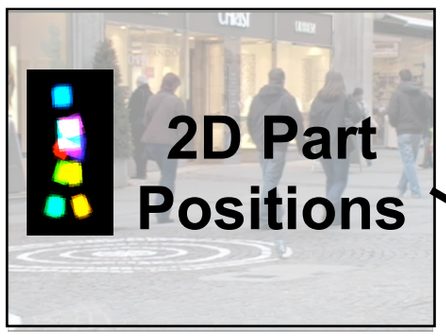
TECHNISCHE UNIVERSITÄT DARMSTADT

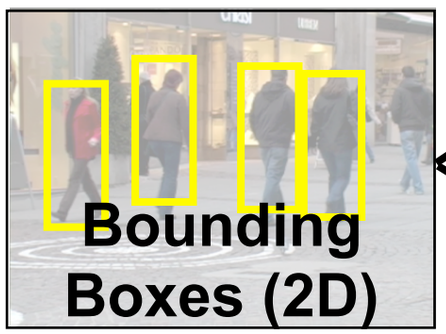
max planck institut informatik

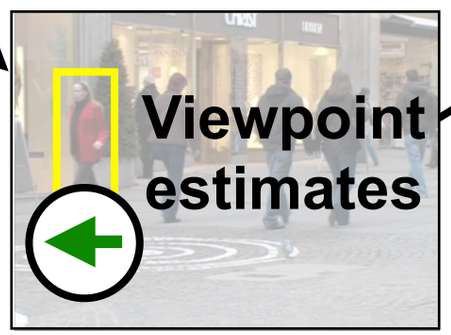
1. Single-frame detection

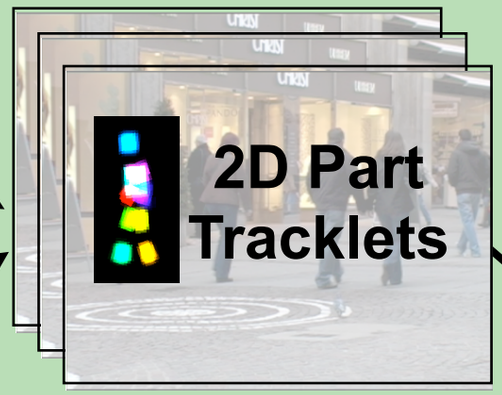
2. 2D-Tracklet detection

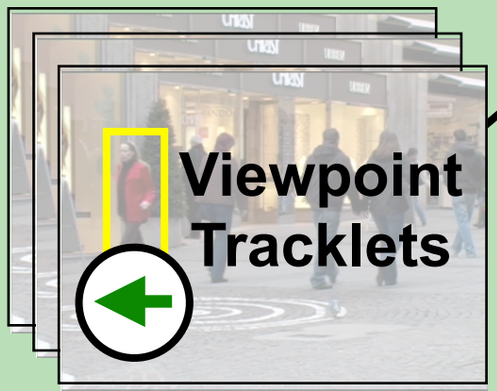
3. 2D-to-3D lifting (tracking & pose estimation)

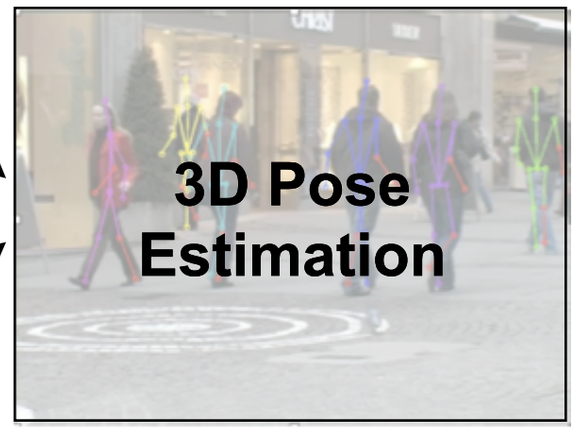






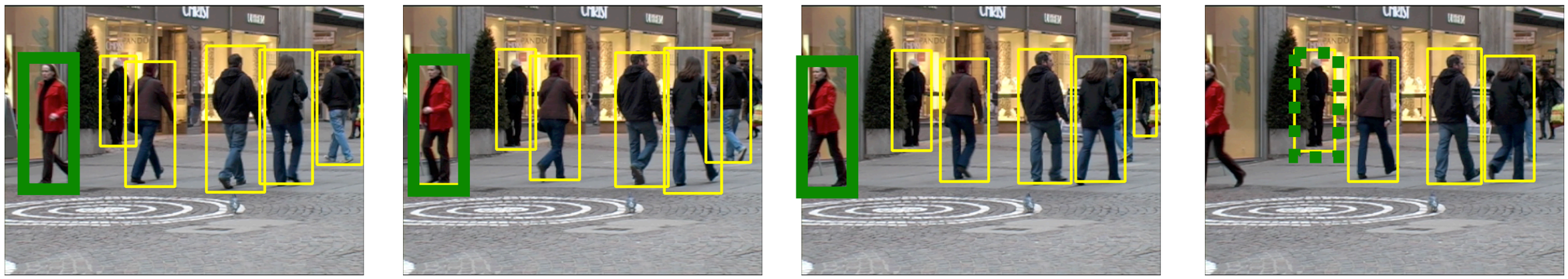


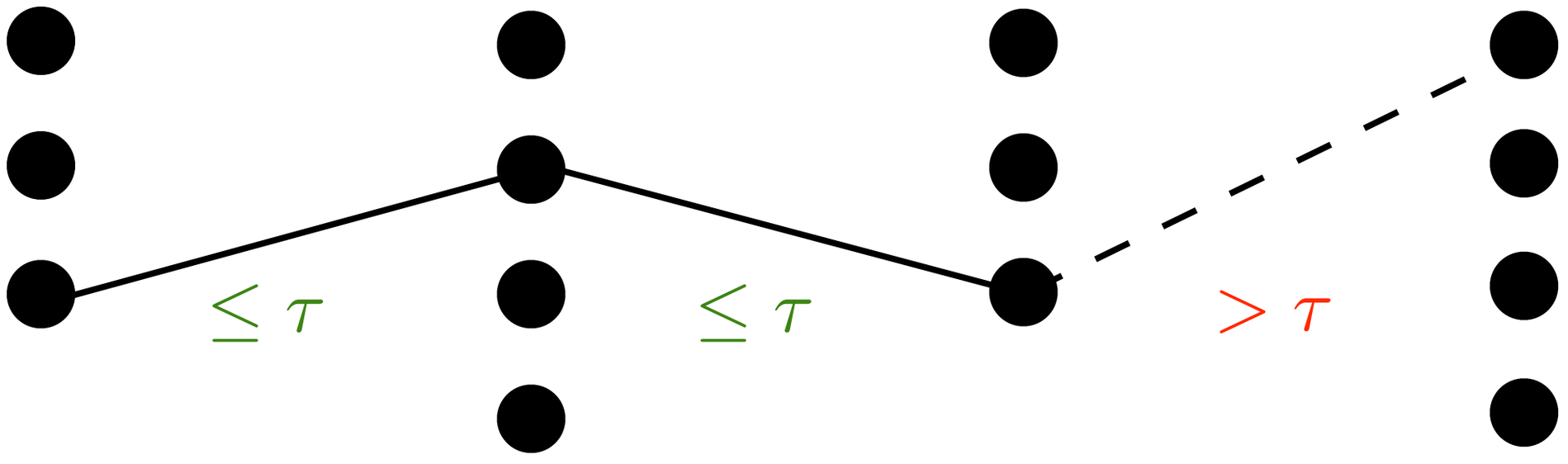




2D Tracking

- Find consistent sequences of 2D hypotheses (tracklets) with Viterbi algorithm [Andriluka et al. CVPR08]:





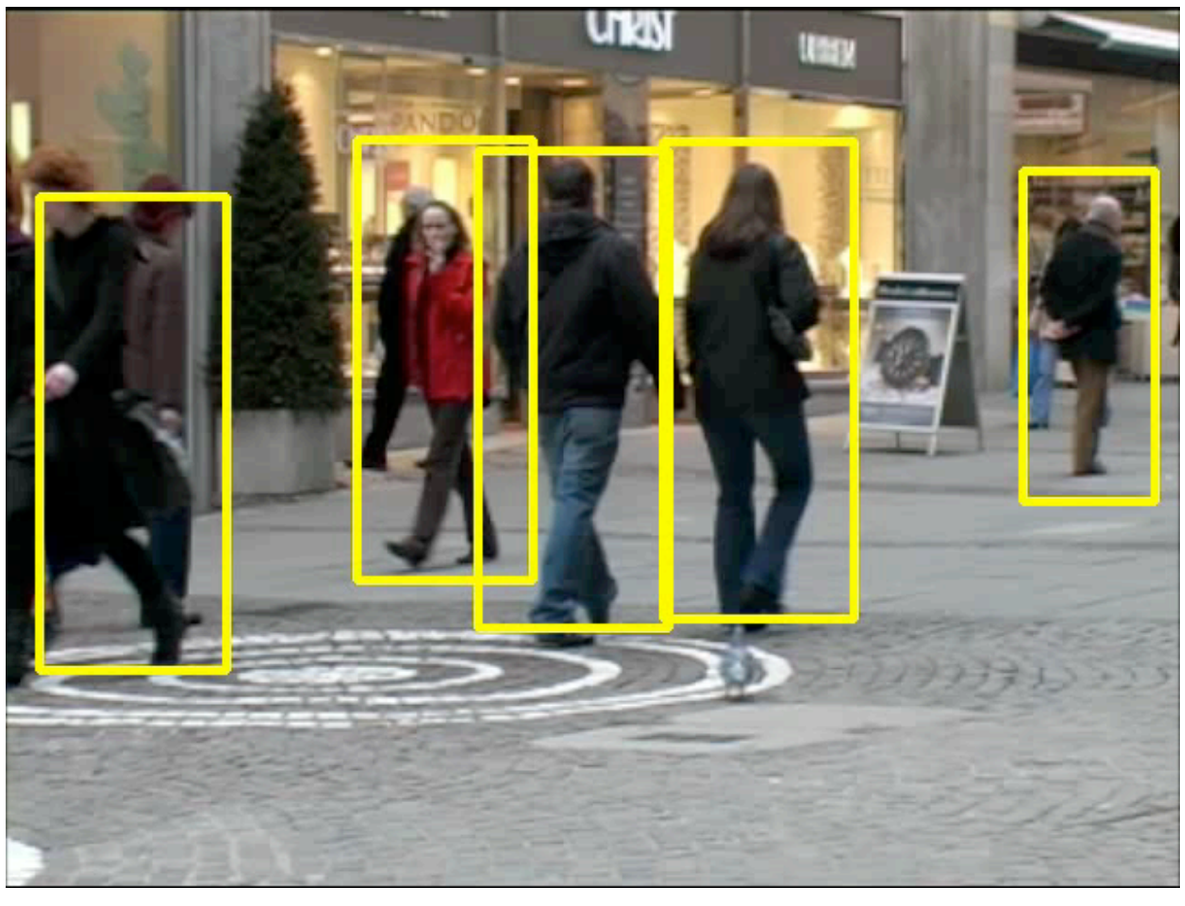
check: $p_{trans}(\mathbf{h}_{mi}, \mathbf{h}_{m-1,j})$

Results: 2D Tracking

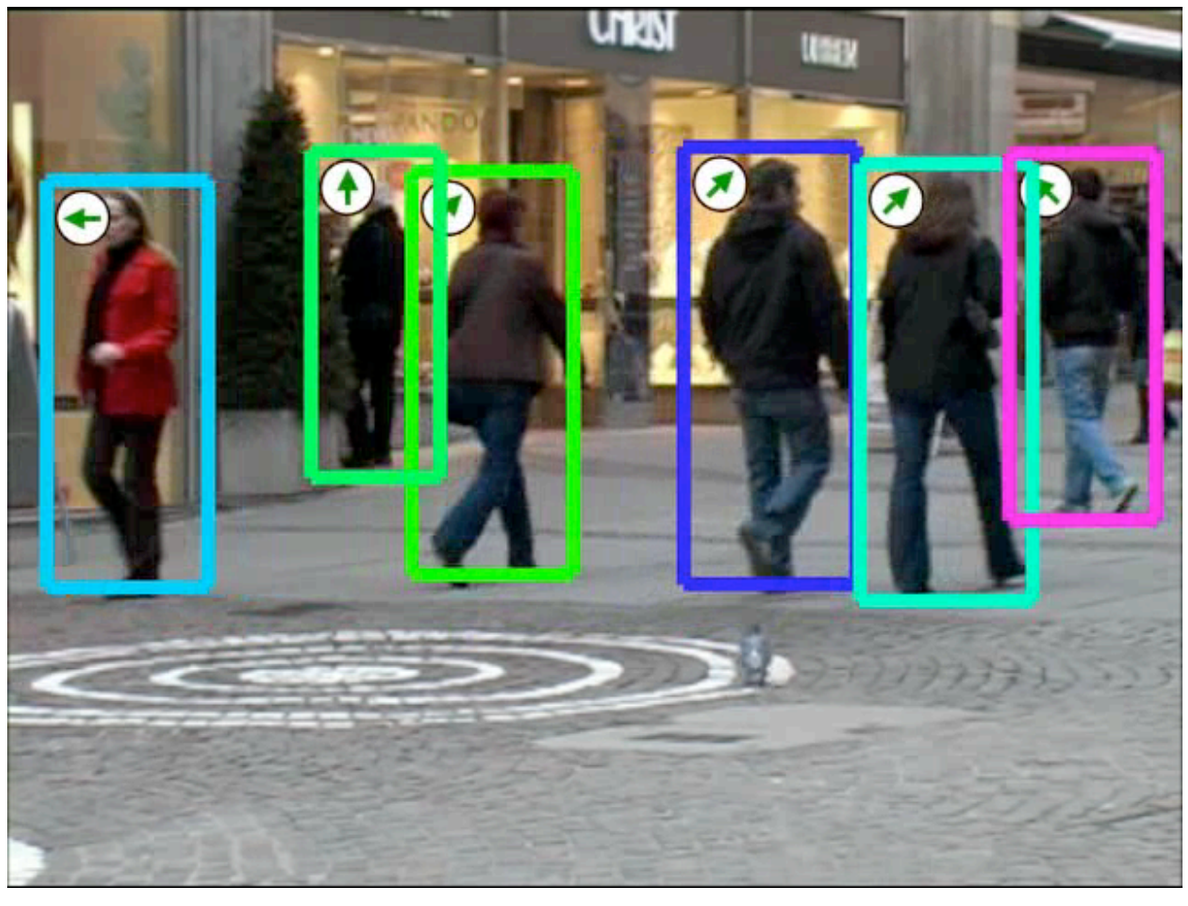
TECHNISCHE UNIVERSITÄT DARMSTADT

max planck institut informatik

1. Single-frame detection



2. 2D-Tracklet detection and viewpoint estimation



Our Approach

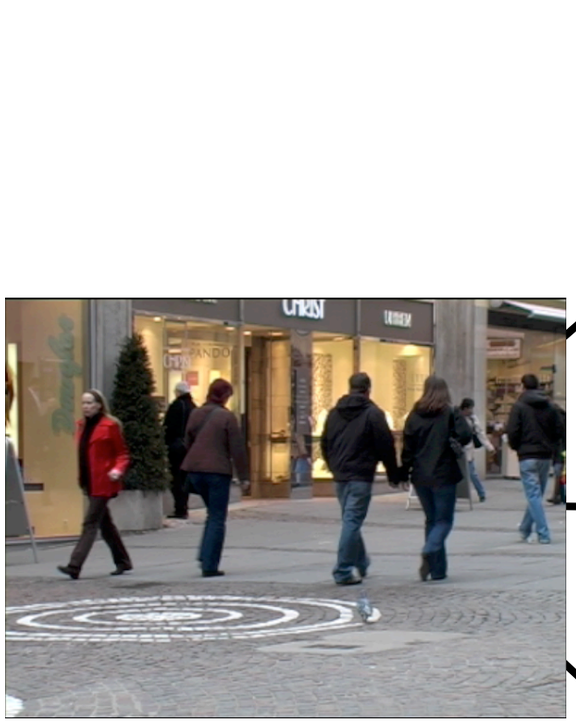
TECHNISCHE UNIVERSITÄT DARMSTADT

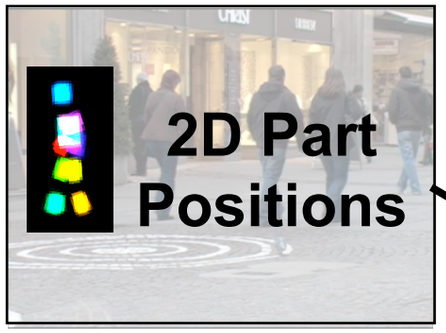
max planck institut informatik

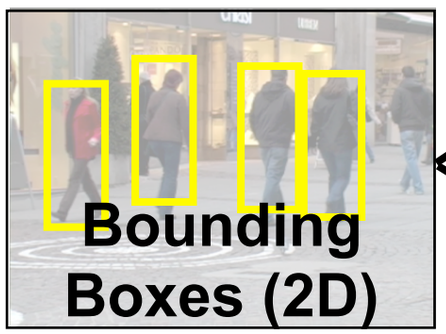
1. Single-frame detection

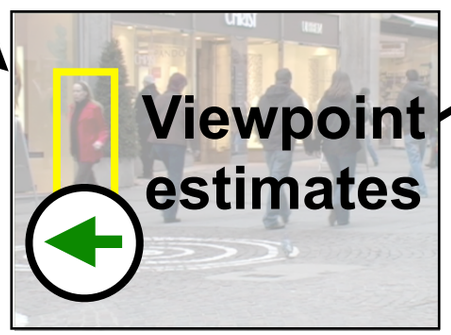
2. 2D-Tracklet detection

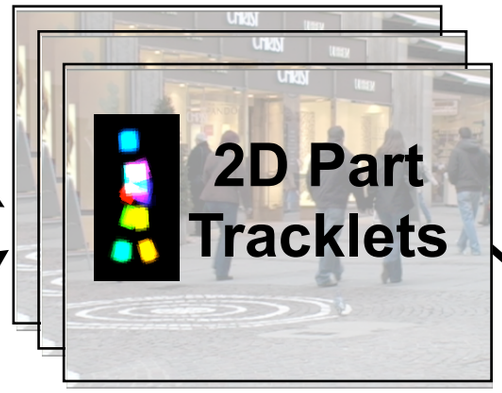
3. 2D-to-3D lifting (tracking & pose estimation)

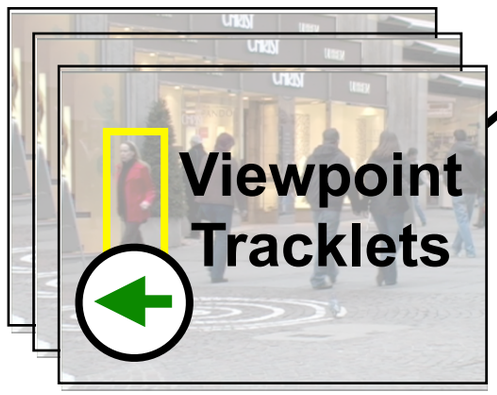


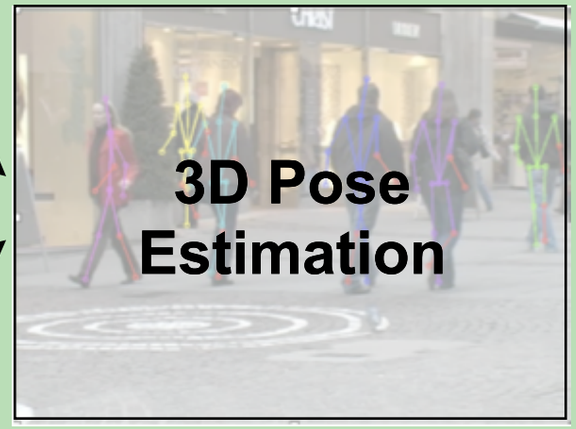












3D Pose Estimation

- Bayesian formulation:

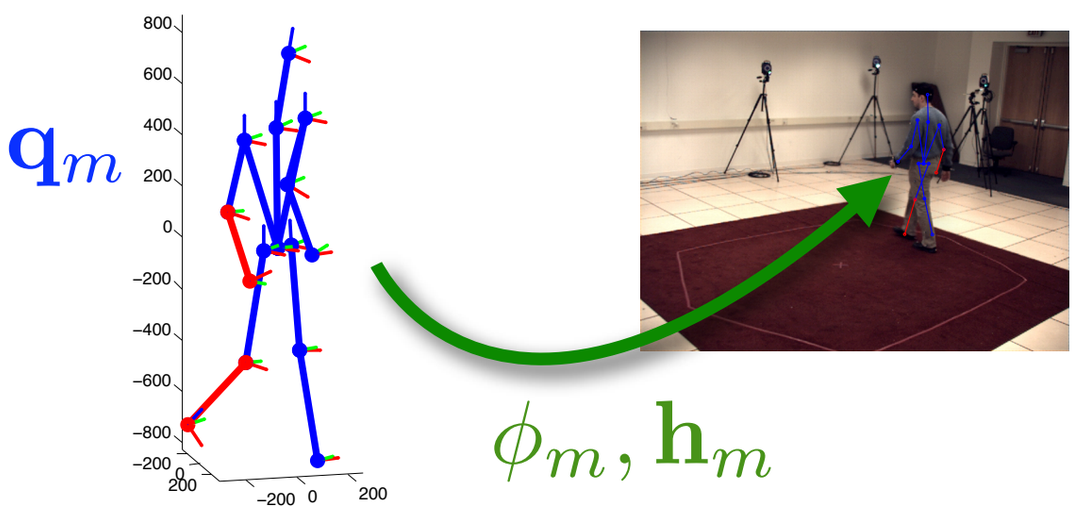
$$p(Q_{1:M}|\mathcal{E}_{1:M}) \propto p(\mathcal{E}_{1:M}|Q_{1:M})p(Q_{1:M})$$

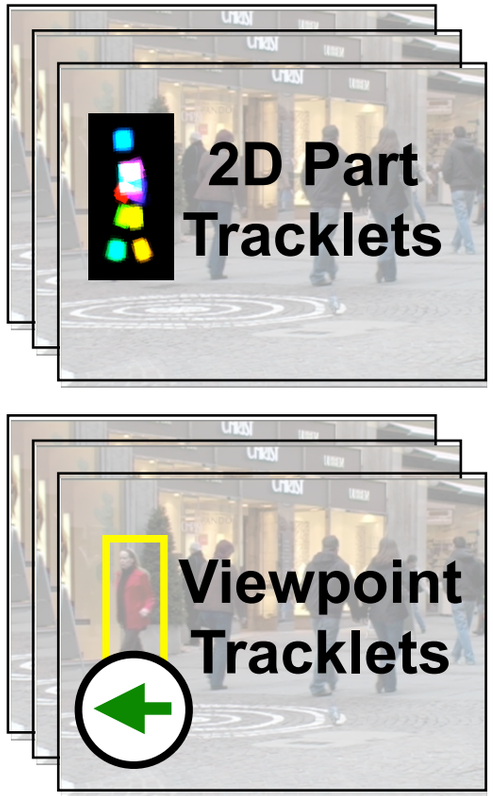
pose and projection

evidence from stages 1 and 2

parameters in each frame:

$$Q_m = \{\mathbf{q}_m, \phi_m, \mathbf{h}_m\}$$





3D Likelihood based on 2D tracklets

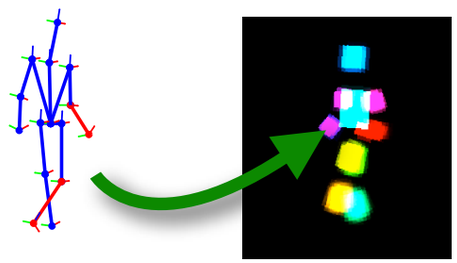
- Bayesian formulation:

$$p(Q_{1:M}|\mathcal{E}_{1:M}) \propto p(\mathcal{E}_{1:M}|Q_{1:M})p(Q_{1:M})$$

Likelihood of 3D parameters

$$p(\mathcal{E}_m|Q_m) = p(\omega_m|Q_m)p(D_m|Q_m)$$

- Projection of 3D pose evaluated under **2D body part posteriors:**



$$p(D_m|Q_m) = \prod_{n=1}^{N} p(\mathrm{proj}_n(Q_m)|D_m)$$

- Restrict orientation to be close to **estimated viewpoint:**

$$p(\omega_m|Q_m) = \mathcal{N}(\omega_m|\phi_m, \sigma_\omega^2)$$

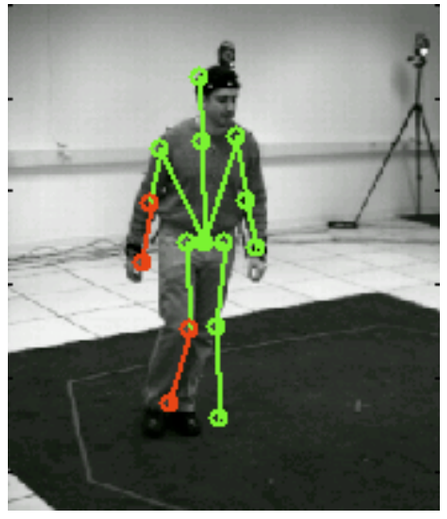
3D Pose Inference

- Bayesian formulation:

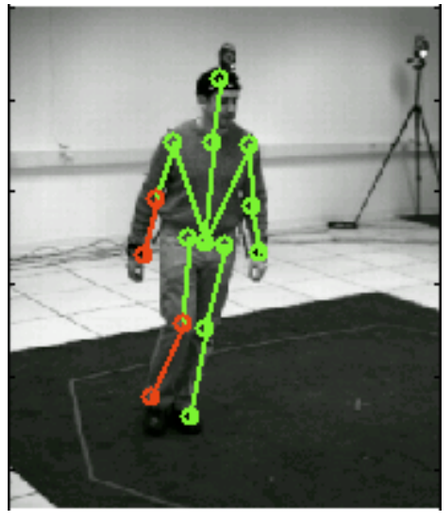
$$p(Q_{1:M}|\mathcal{E}_{1:M}) \propto p(\mathcal{E}_{1:M}|Q_{1:M})p(Q_{1:M})$$

- Find local maxima with scaled conjugate gradients
- Initialization: 2D to 3D lifting by finding sequences of 3D exemplars that fit well to 2D tracklet observations

Initialization



Optimization Result

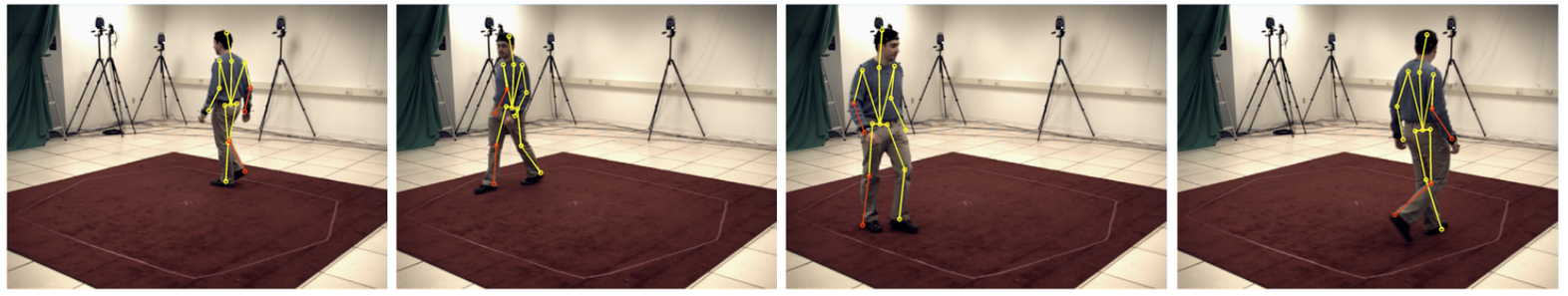


3D Pose Estimation: Quantitative Results

TECHNISCHE UNIVERSITÄT DARMSTADT

max planck institut informatik

- Quantitative evaluation on HumanEva II dataset
- Compare to [Rogez et al. CVPR'08]



Subect/Camera (Frames)	Mean (Std) [Rogez et al. CVPR'08]	Mean (Std) this paper	3D Mean (Std) this paper
S2/C1 (1-350)	12.98 (3.5) px.	**10.49 (2.7) px.**	107 (15) mm.
S2/C2 (1-350	14.18 (4.38) px.	**10.72 (2.44) px.**	101 (9) mm.

3D Pose Estimation: Qualitative Results

TECHNISCHE UNIVERSITÄT DARMSTADT

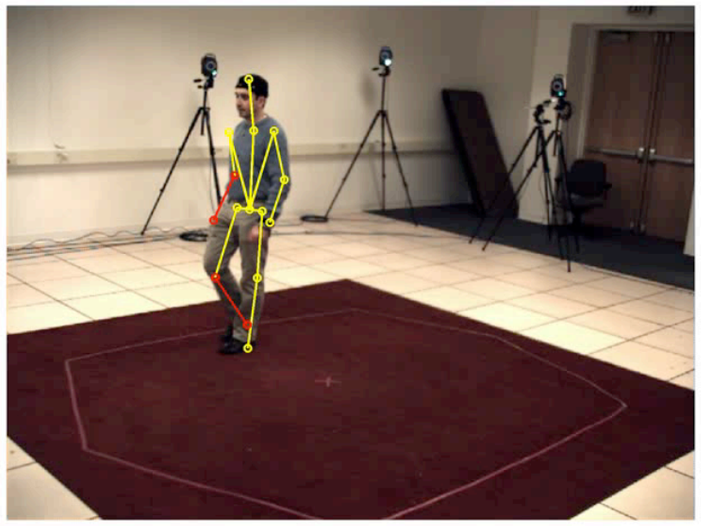
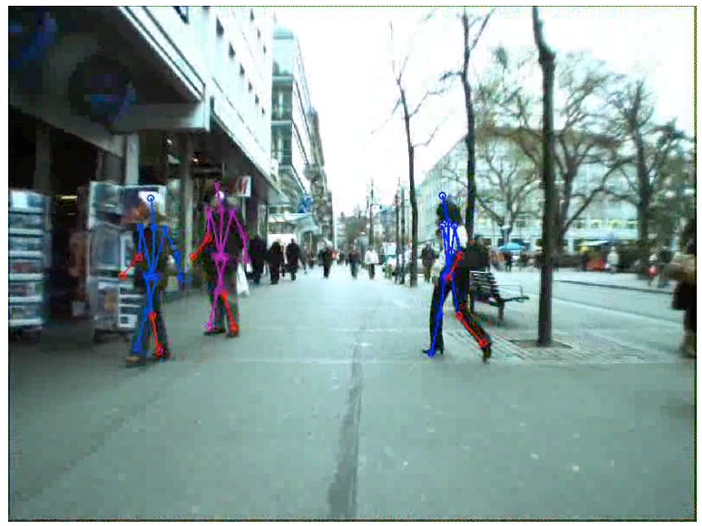
max planck institut informatik

Conclusions and Future Work

- Novel probabilistic approach to 3D pose estimation:
  - powerful 3D likelihood based on evidence from pictorial structures model accumulated over multiple frames
- Leverage recent results in people detection, 2D pose estimation and human motion modeling
- State-of-the-art performance in lab setting
- Applicable in uncontrolled street conditions
- Future work: broader class of motions and more detailed evaluation

See you at the Poster!

- Datasets available at:

  www.mis.tu-darmstadt.de/datasets

- Micha Andriluka gratefully acknowledges scholarship from DFG GRK 1362 “Cooperative, Adaptive and Responsive Monitoring In Mixed-Mode Environments”