People-Tracking-by-Detection and People-Detection-by-Tracking

Mykhaylo Andriluka

Stefan Roth

Bernt Schiele

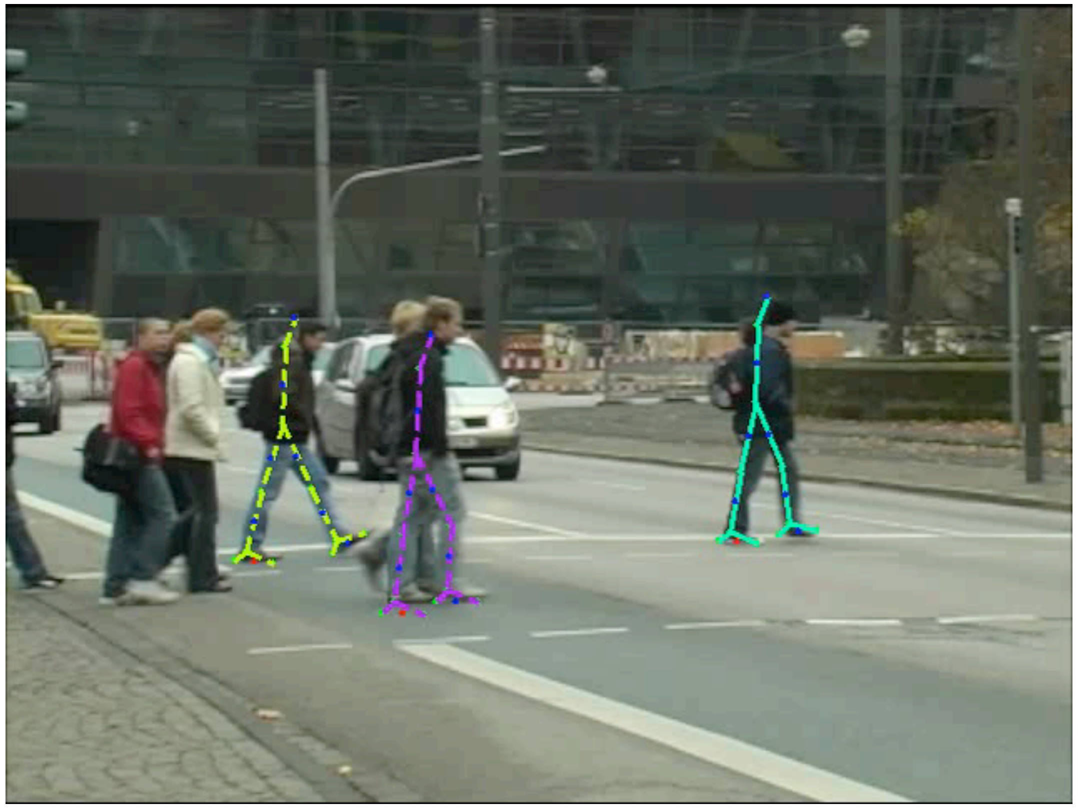
Department of Computer Science
TU Darmstadt

Motivation

- Goal: Detection and tracking of people in complex scenes
- Challenges for detection:
  - Partial occlusions
  - Appearance variation
  - Data association difficult
- Challenges for tracking:
  - Dynamic backgrounds
  - Multiple people
  - Frequent long term occlusions

Motivation

- Goal: Detection and tracking of people in complex scenes
- Challenges for detection:
  - Partial occlusions
  - Appearance variation
  - Data association difficult

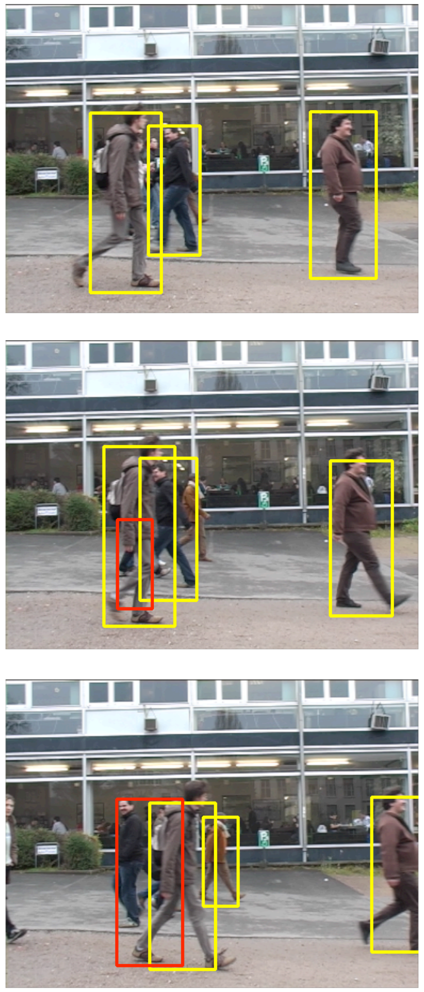


- Challenges for tracking:
  - Dynamic backgrounds
  - Multiple people
  - Frequent long term occlusions

Overview

Three stages of our multi-person detection and tracking system:

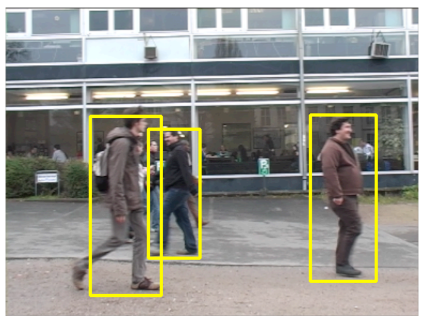
1. Single-frame detection

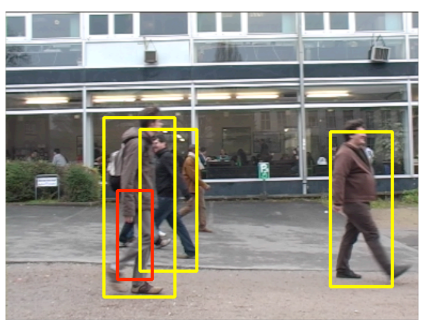


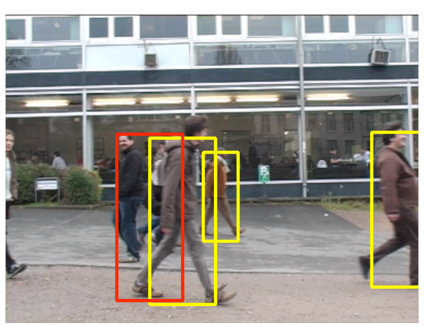
Overview

Three stages of our multi-person detection and tracking system:

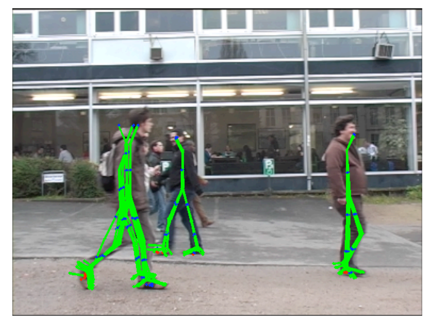
1. Single-frame detection

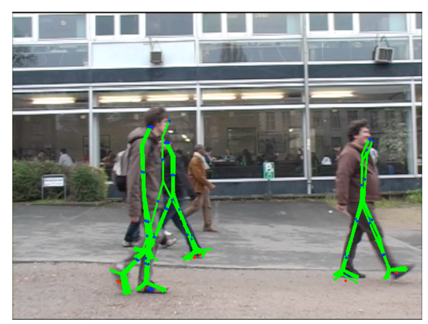


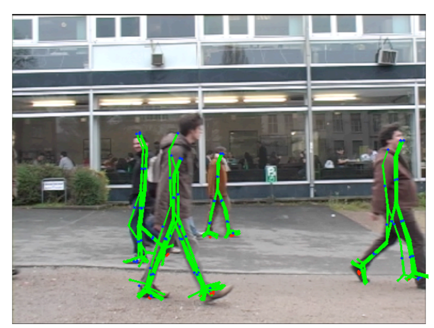




2. Tracklet detection



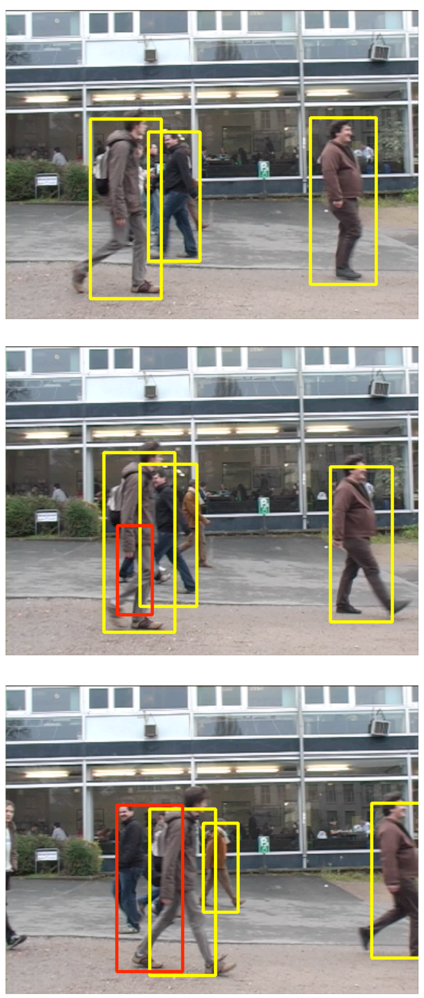




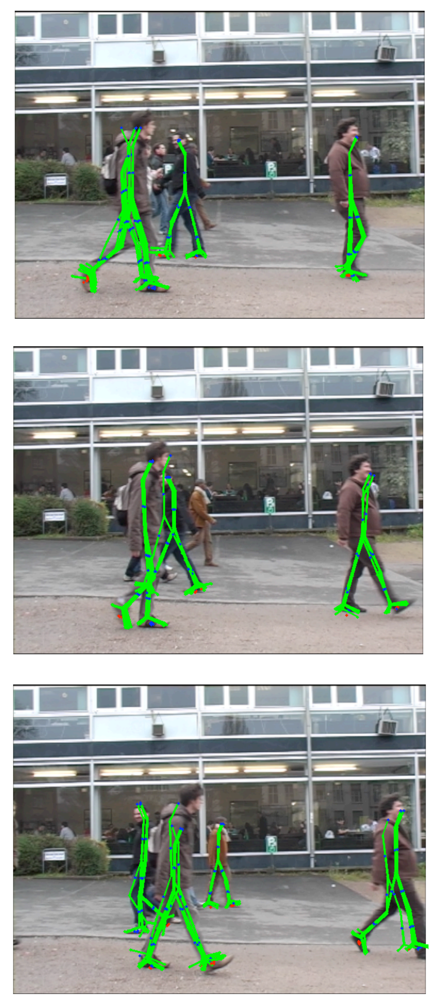
Overview

Three stages of our multi-person detection and tracking system:

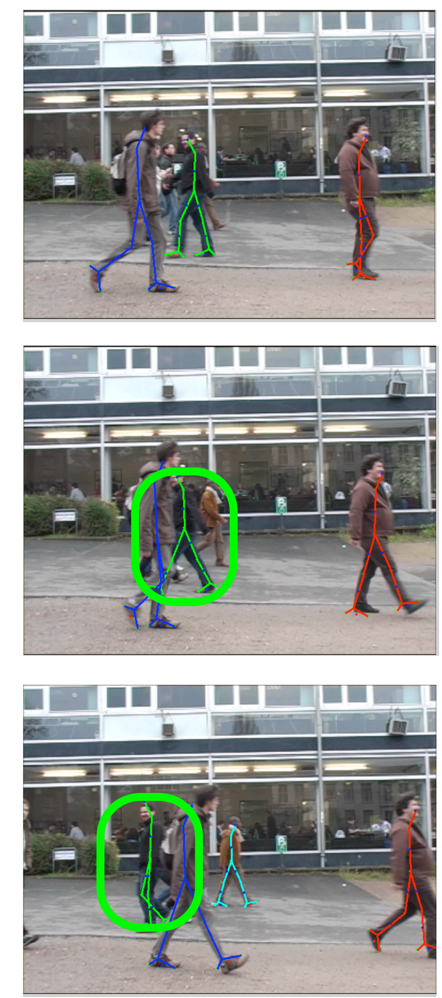
1. Single-frame detection



2. Tracklet detection



3. Tracking through occlusion



Previous Work

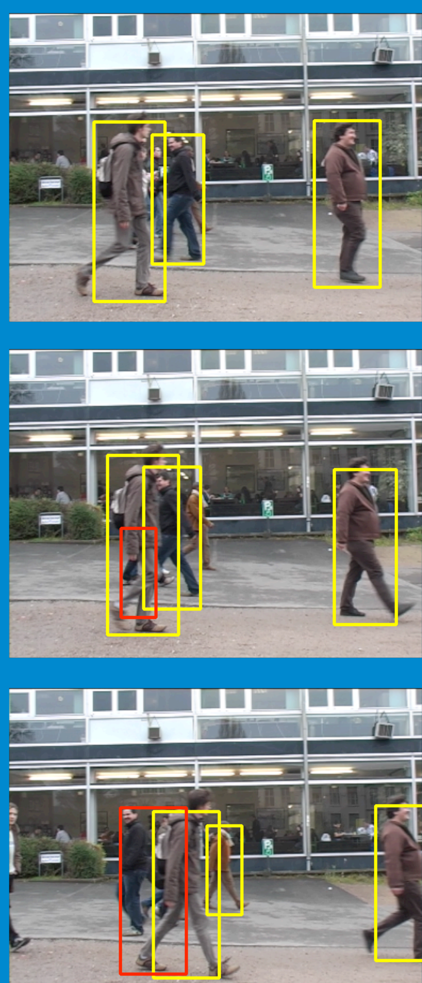
TECHNISCHE UNIVERSITÄT DARMSTADT

- People Detection & Tracking:
  - [Fossati et al., CVPR 2007]: 3D articulated tracking aided by detection, single person, ground plane needed.
  - [Leibe et al., ICCV 2007]: Detection of tracking of multiple people, high viewpoint → no full-body occlusions.
  - [Ramanan et al., PAMI 2007]: Appearance model learned from people detection, then used for tracking and data association.
  - [Wu & Nevatia, IJCV 2007]: Use detection for tracking, works for multiple people → no articulations, detector not aided by tracking.
- Here:
  - More people
  - Significant, long-term full-body occlusions
  - However: more restricted scenario (2-D, people in side views)

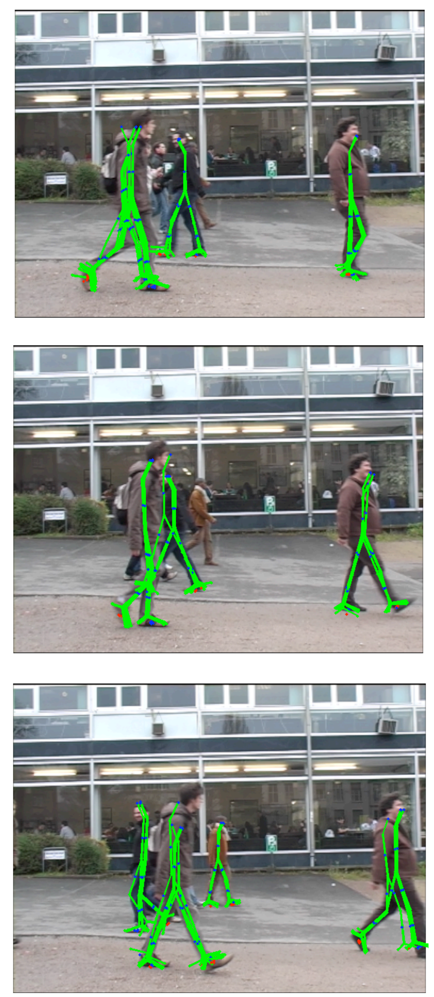
Overview

Three stages of our multi-person detection and tracking system:

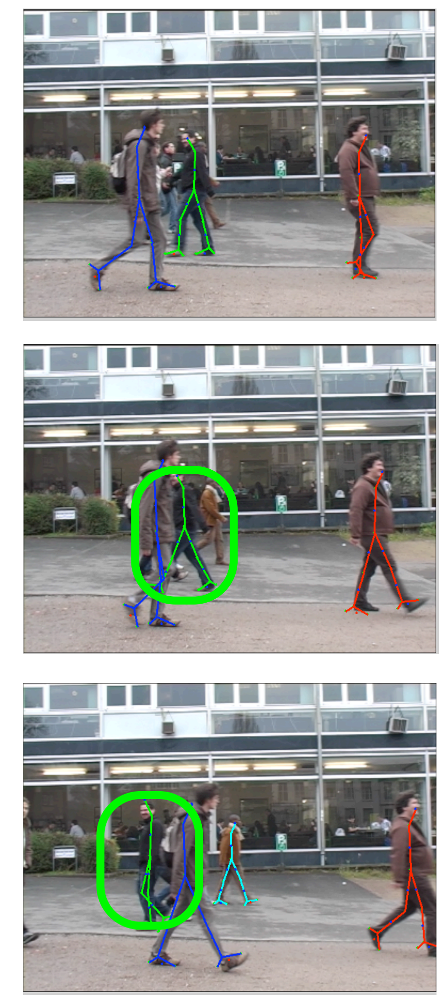
1. Single-frame detection



2. Tracklet detection



3. Tracking through occlusion



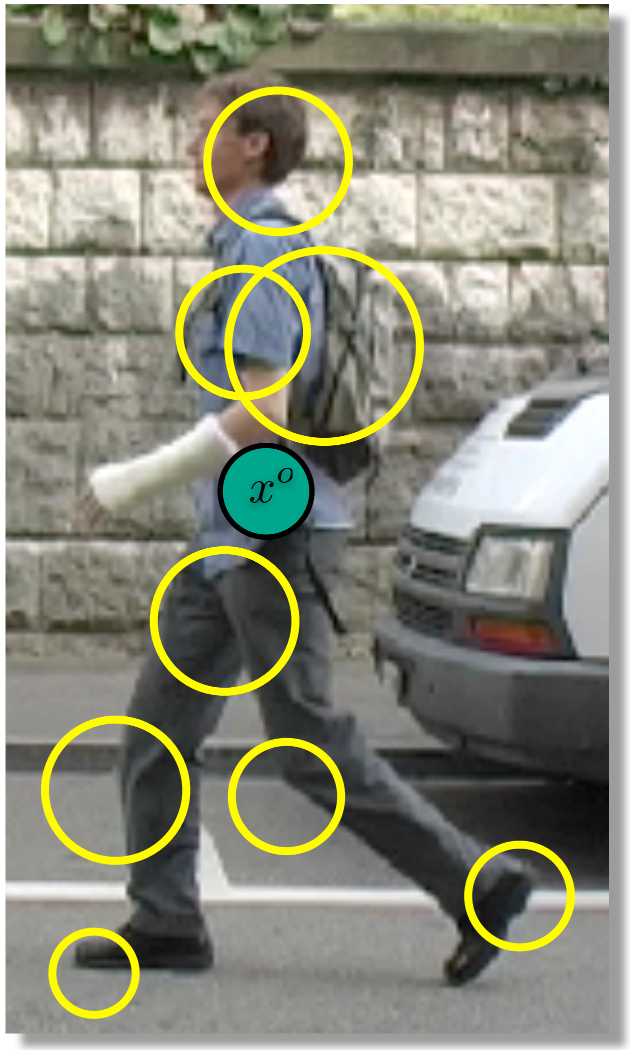
Single-frame Detector: partISM

TECHNISCHE UNIVERSITÄT DARMSTADT

- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]

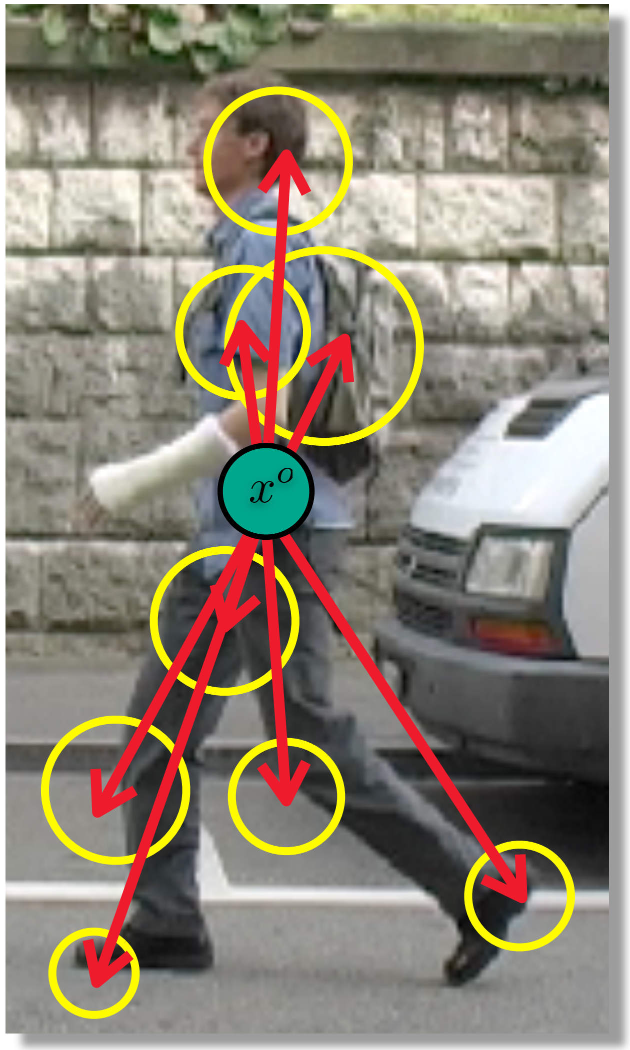
Single-frame Detector: partISM

- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]



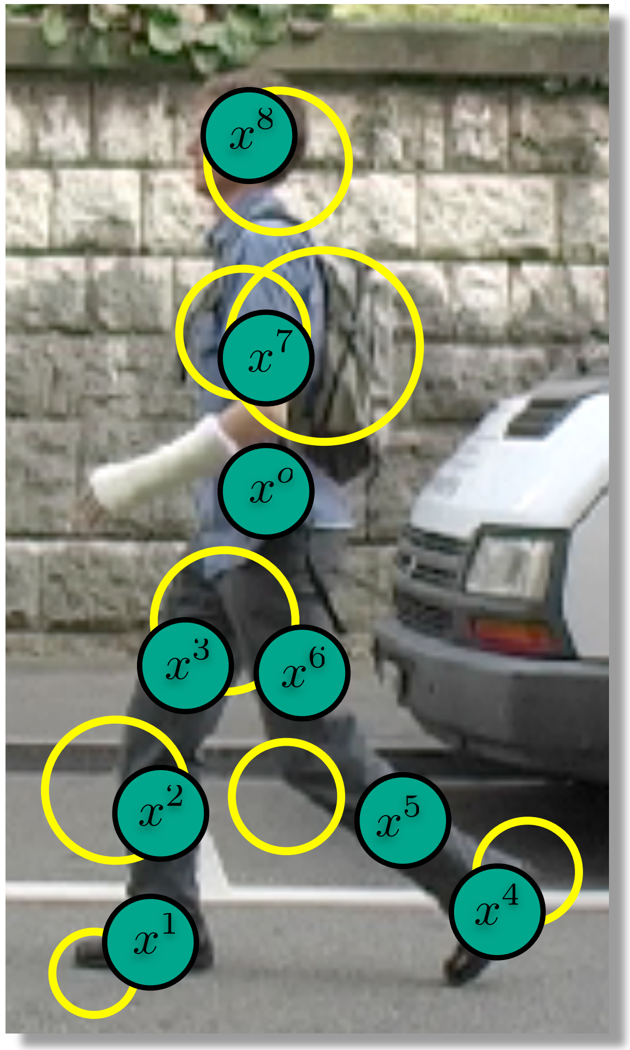
Single-frame Detector: partISM

- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]



Single-frame Detector: partISM

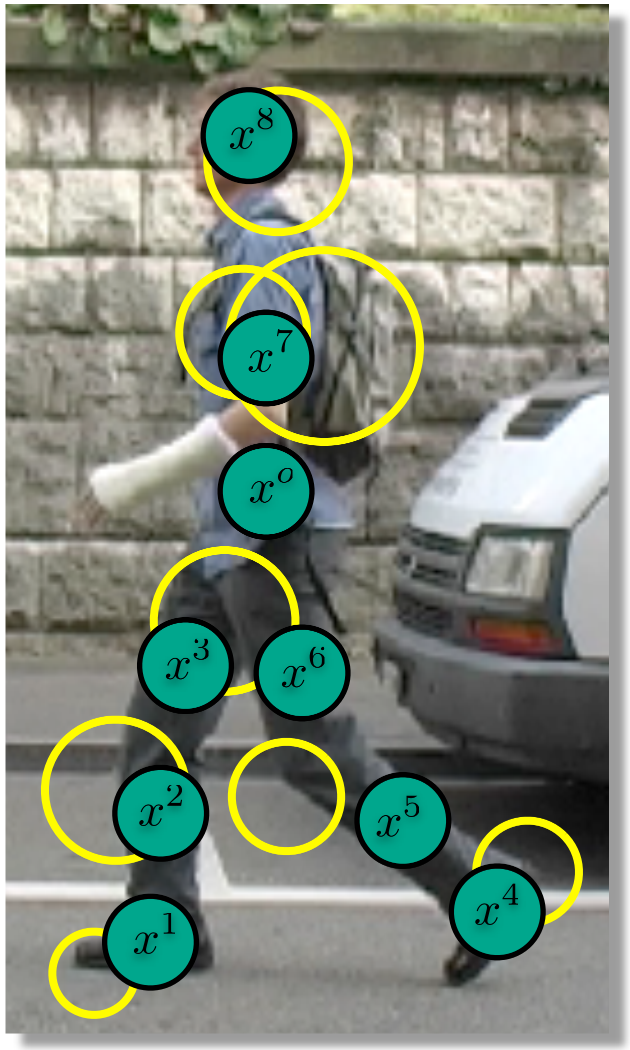
- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]



Single-frame Detector: partISM

TECHNISCHE UNIVERSITÄT DARMSTADT

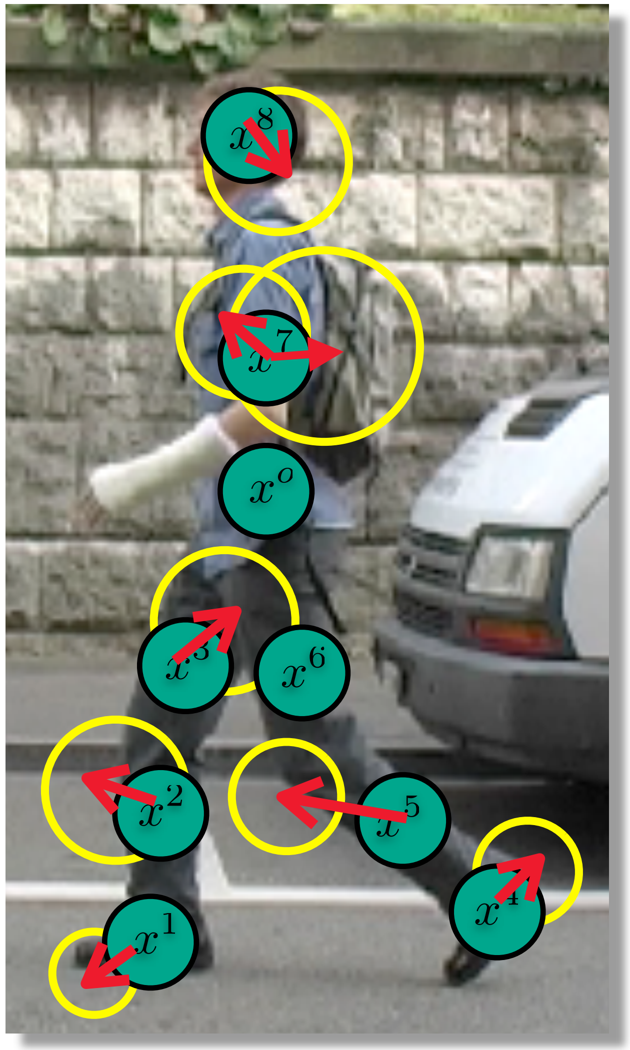
- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]
- Part decomposition and inference:
  Pictorial structures model
  [Felzenszwalb & Huttenlocher, IJCV 2005]



Single-frame Detector: partISM

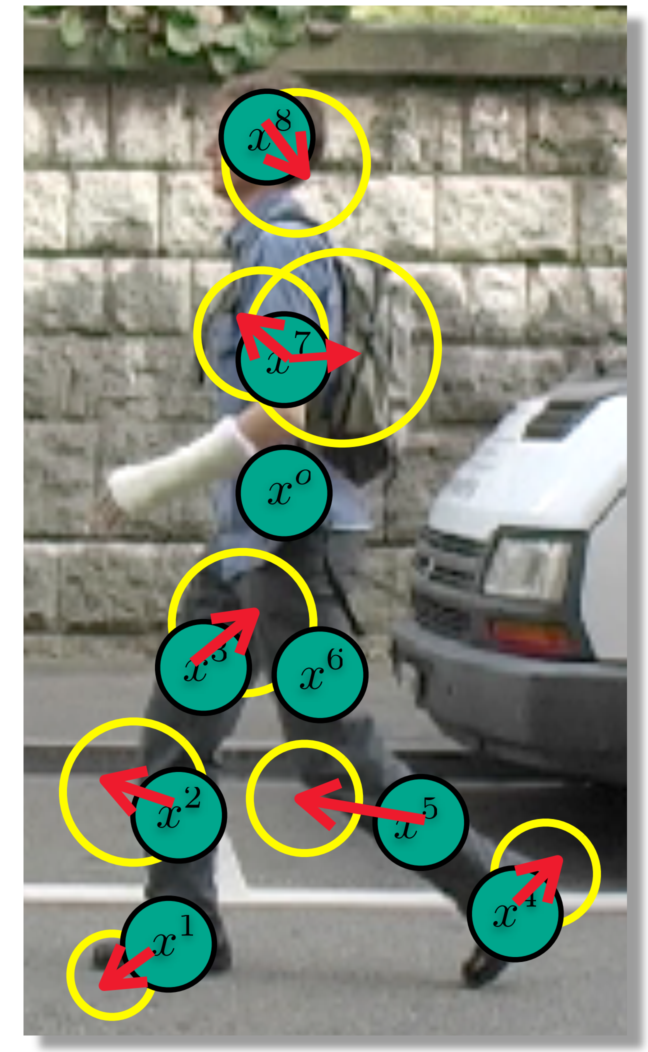
TECHNISCHE UNIVERSITÄT DARMSTADT

- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]
- Part decomposition and inference:
  Pictorial structures model
  [Felzenszwalb & Huttenlocher, IJCV 2005]



Single-frame Detector: partISM

- Appearance of parts:
  Implicit Shape Model (ISM)
  [Leibe, Seemann & Schiele, CVPR 2005]
- Part decomposition and inference:
  Pictorial structures model
  [Felzenszwalb & Huttenlocher, IJCV 2005]

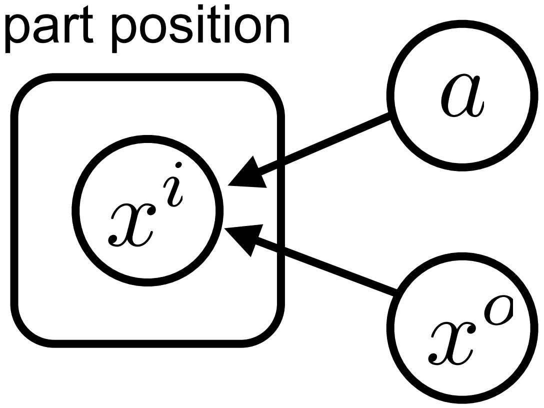


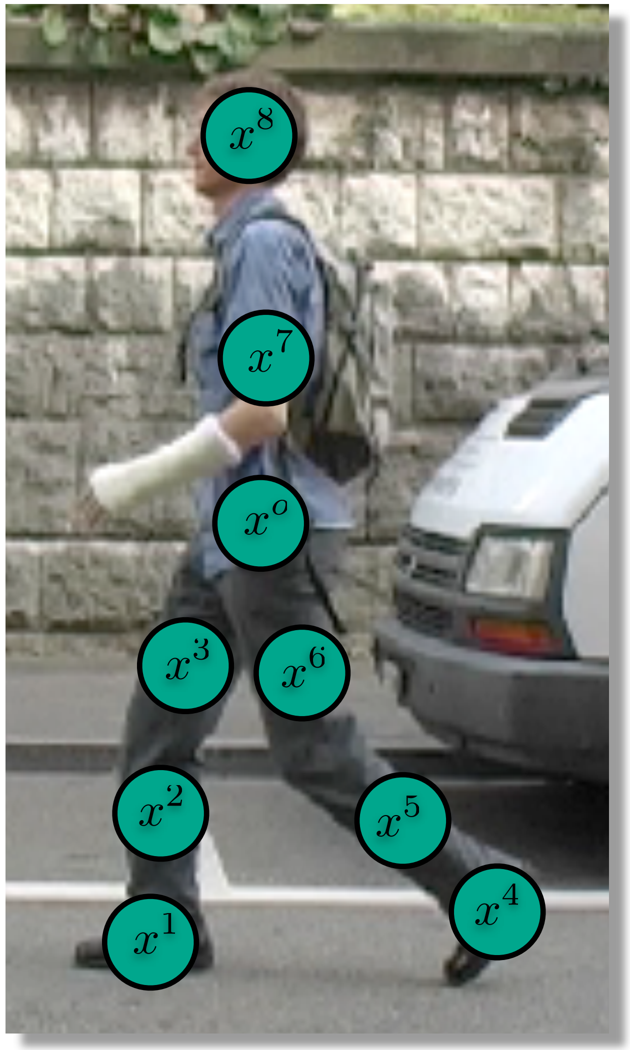
$$p(L|E) \propto p(E|L)p(L)$$

Body-part positions Image evidence

Part Decomposition

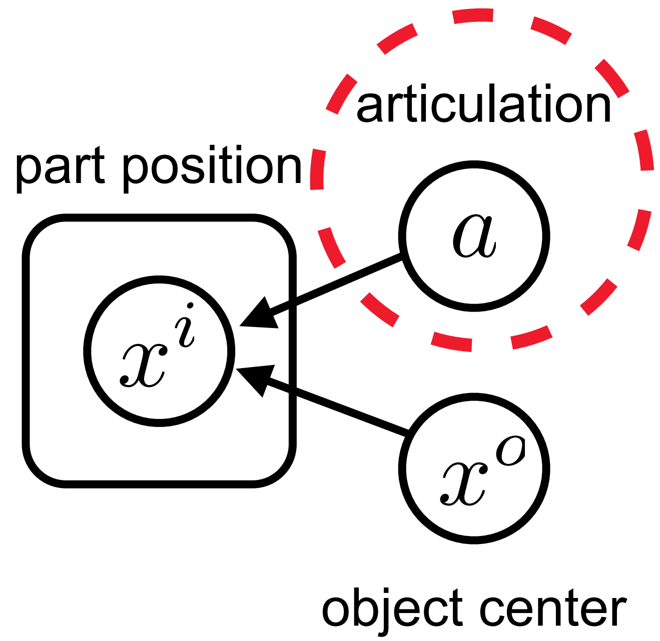
- $L = \{x^o, x^1, \ldots, x^8\}$ - configuration of body parts
- Structure of the prior distribution $p(L)$:
  - Articulation variable a models correlations between part positions.
  - Given articulation, prior on configuration becomes a star model.

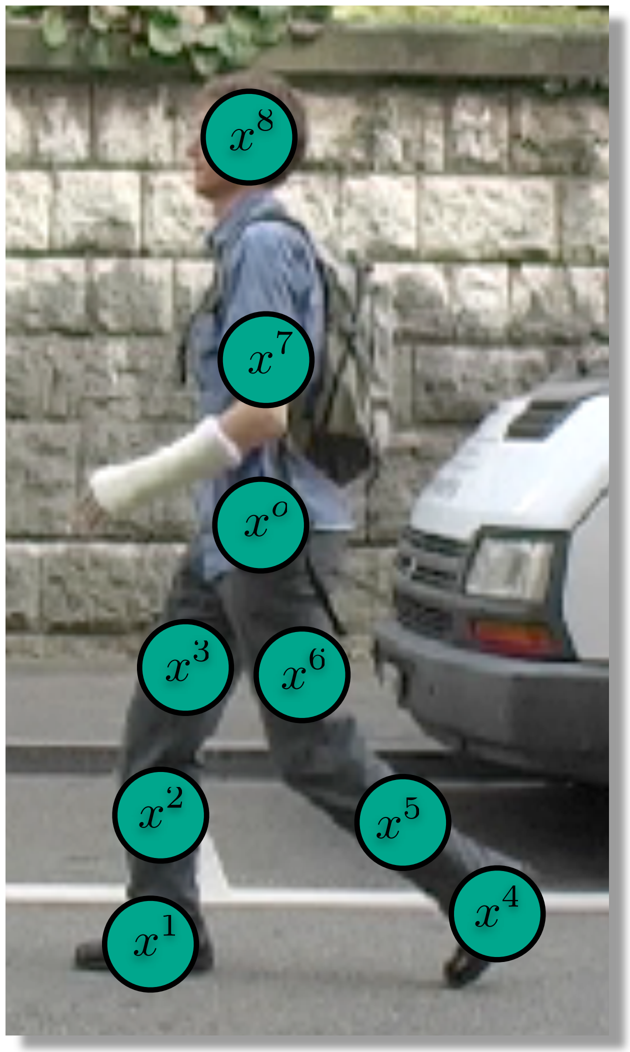




Part Decomposition

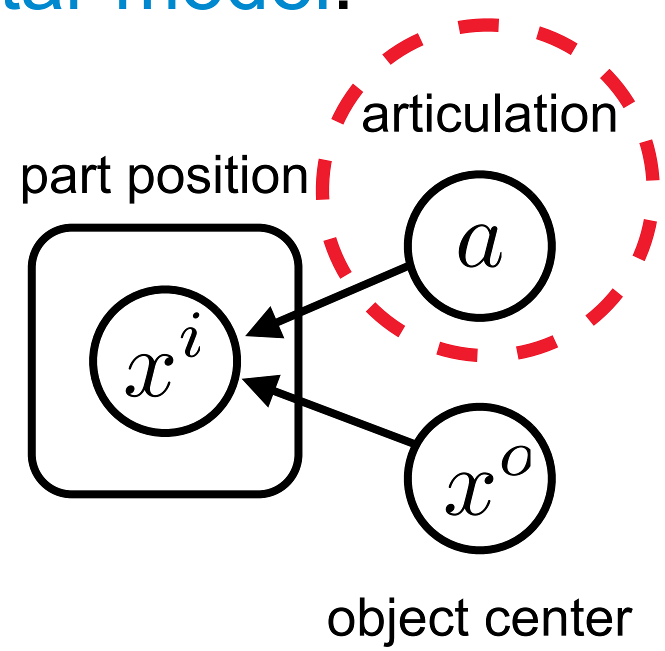
- $L = \{x^o, x^1, \ldots, x^8\}$ - configuration of body parts
- Structure of the prior distribution $p(L)$:
  - Articulation variable a models correlations between part positions.
  - Given articulation, prior on configuration becomes a star model.

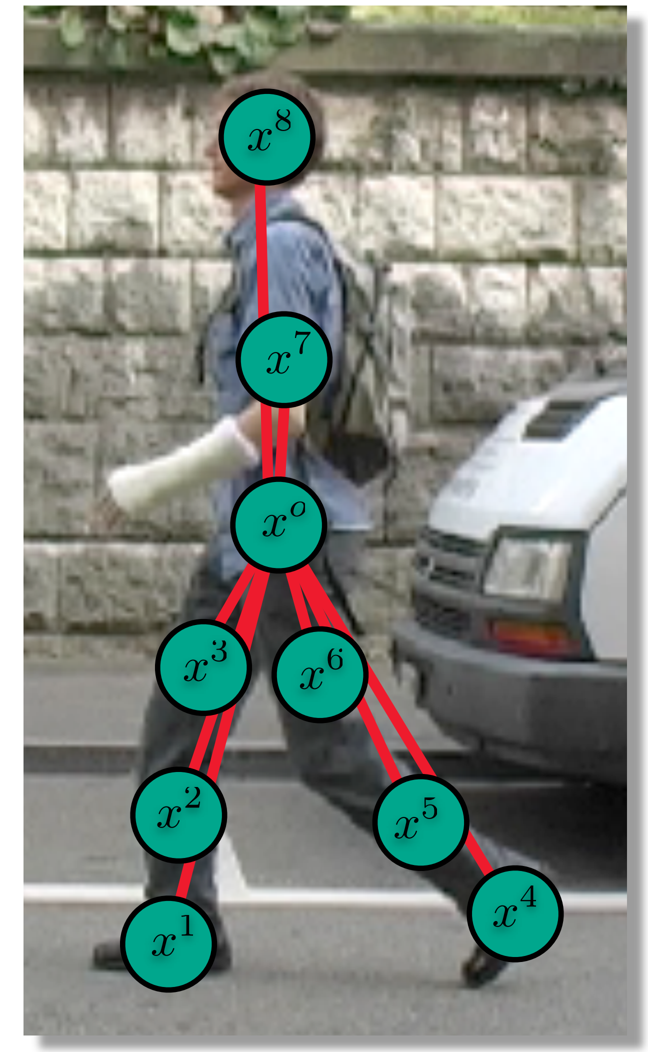




Part Decomposition

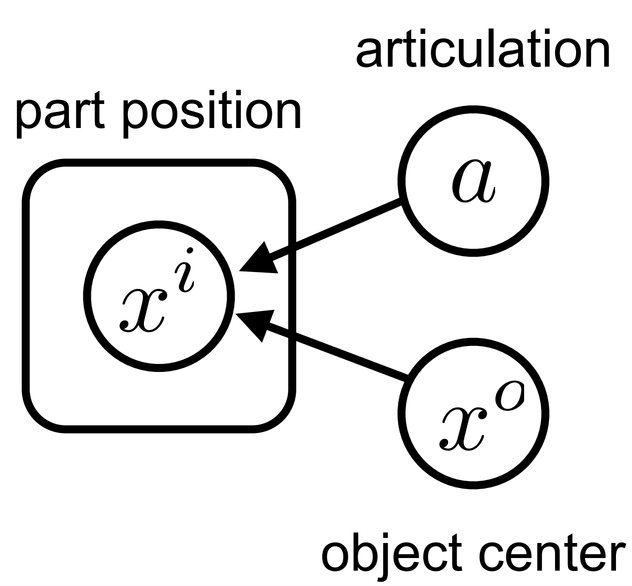
- $L = \{x^o, x^1, \ldots, x^8\}$ - configuration of body parts
- Structure of the prior distribution $p(L)$:
  - Articulation variable a models correlations between part positions.
  - Given articulation, prior on configuration becomes a star model.

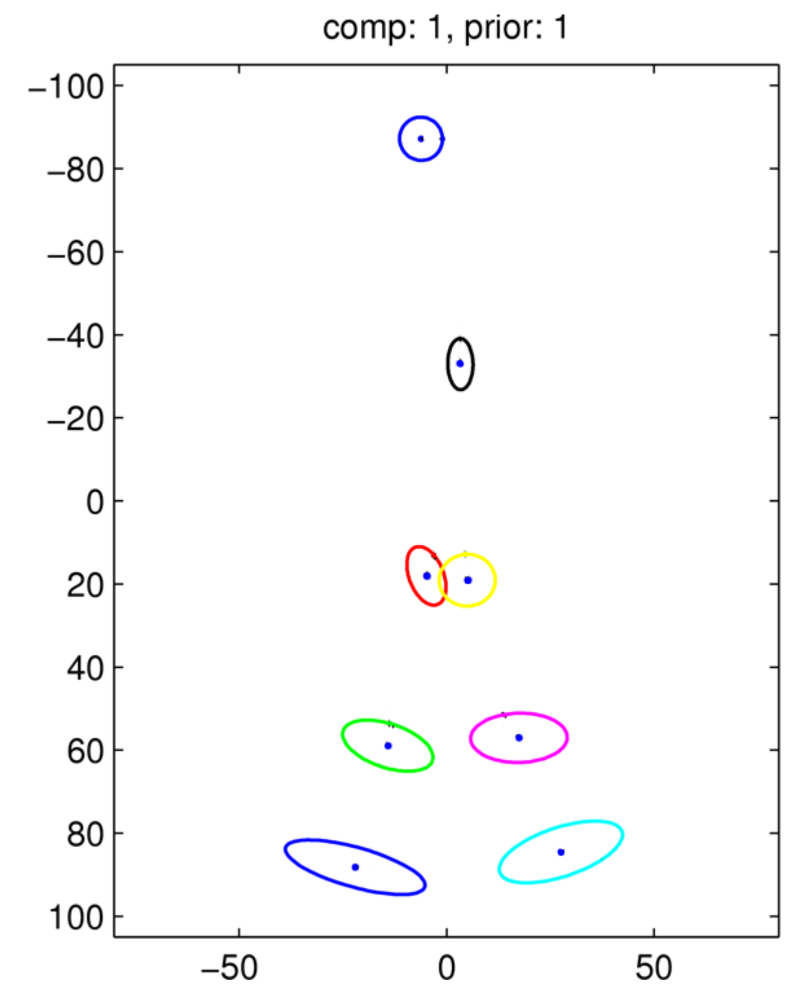




Part Decomposition

- $L = \{x^o, x^1, \ldots, x^8\}$ - configuration of body parts
- Structure of the prior distribution $p(L)$:
  - Articulation variable a models correlations between part positions.
  - Given articulation, prior on configuration becomes a star model.





Covariance and mean part positions for $p(x^i|x^o)$.

Single Frame Detection

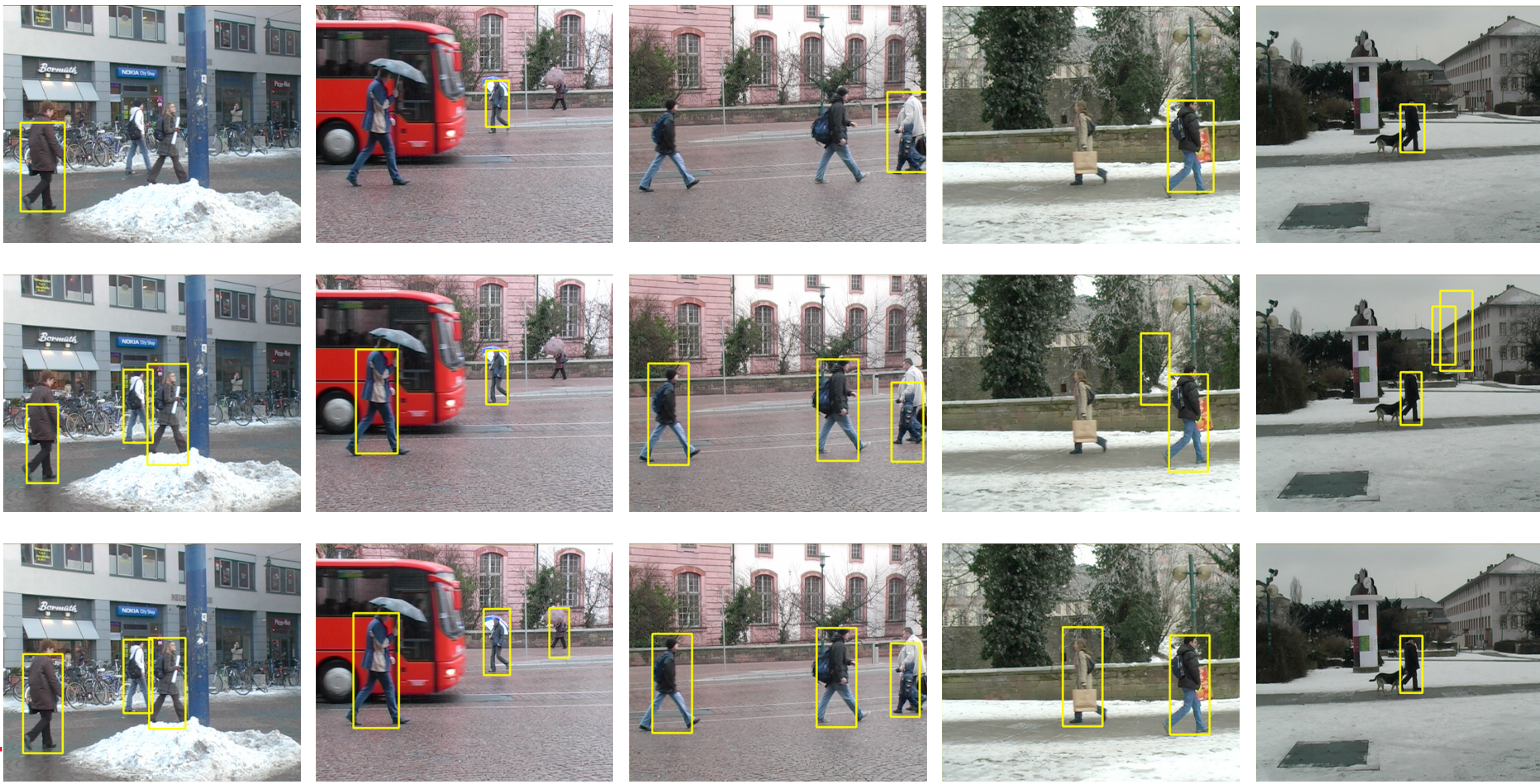
TECHNISCHE UNIVERSITÄT DARMSTADT

- Detections at equal error rate:

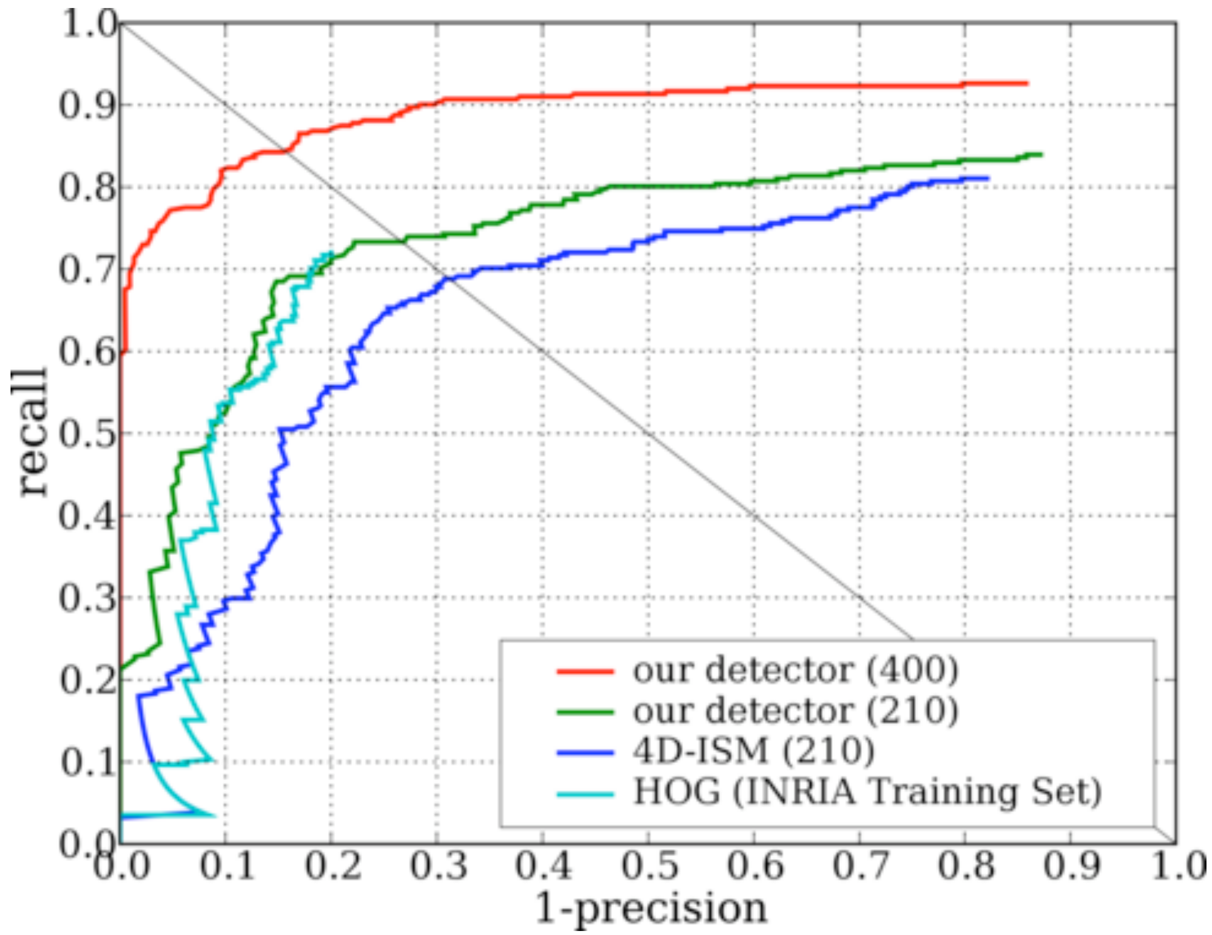
HOG

4D-ISM

partISM



Single-frame Detection Results



TUD pedestrians data

No occlusions

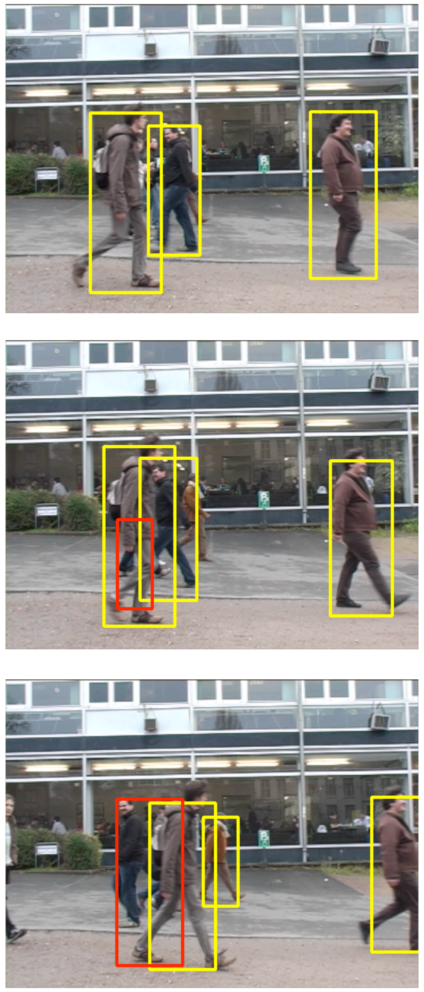
- partISM clearly outperforms 4D-ISM [Seemann et al, DAGM'06].
- Outperforms HOG [Dalal&Triggs, CVPR'05] with much less training data (Note: we only use sideviews).

Overview

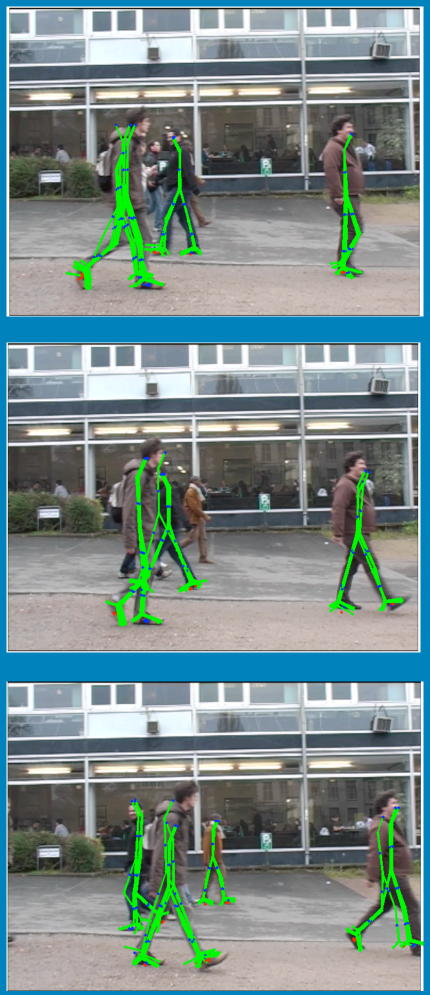
TECHNISCHE UNIVERSITÄT DARMSTADT

Three stages of our multi-person detection and tracking system:

1. Single-frame detection



2. Tracklet detection



3. Tracking through occlusion

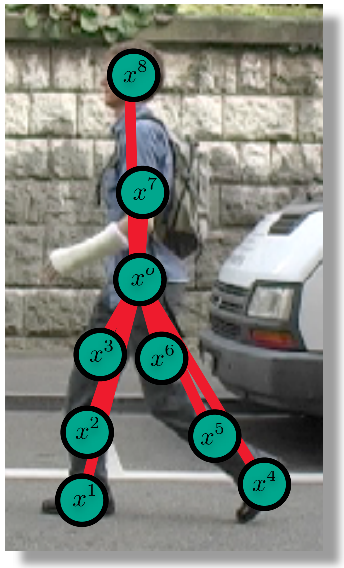
Tracklet Detection in Short Subsequences

- Given: $E = [E_1, \dots, E_m]$

frame 1 frame 2 frame m

overlapping subsequences

- Want:



- Posterior over positions and configurations:

Tracklet Detection in Short Subsequences

- Given: $E = [E_1, \ldots, E_m]$

frame 1 frame 2 frame m

overlapping subsequences

- Want:

$$\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \ldots, \mathbf{x}_m^{o*}]$$

body positions

- Posterior over positions and configurations:

Tracklet Detection in Short Subsequences

- Given: $E = [E_1, \ldots, E_m]$

frame 1 frame 2 frame m

overlapping subsequences

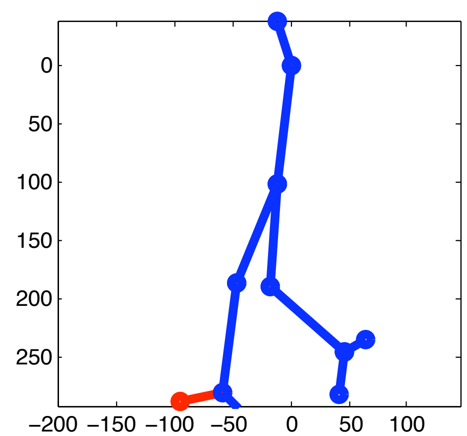
- Want:

$\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \ldots, \mathbf{x}_m^{o*}]$

body positions

$\mathbf{Y}^{*} = [\mathbf{y}_1^{*}, \ldots, \mathbf{y}_m^{*}]$

body configurations



- Posterior over positions and configurations:

Tracklet Detection in Short Subsequences

- Given: $E = [E_1, \ldots, E_m]$

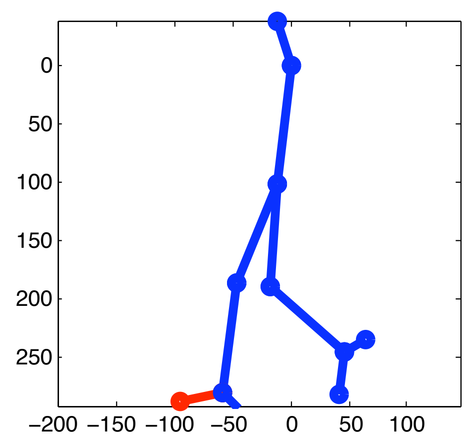
frame 1 frame 2 ... frame m

overlapping subsequences

- Want:

$\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \ldots, \mathbf{x}_m^{o*}]$
body positions

$\mathbf{Y}^* = [\mathbf{y}_1^*, \ldots, \mathbf{y}_m^*]$
body configurations



- Posterior over positions and configurations:

$$p(\mathbf{X}^{o*}, \mathbf{Y}^* | E) \propto p(E | \mathbf{X}^{o*}, \mathbf{Y}^*) p(\mathbf{X}^{o*}) p(\mathbf{Y}^*).$$

Tracklet Detection in Short Subsequences

- Given: $E = [E_1, \ldots, E_m]$

frame 1 frame 2 frame m

overlapping subsequences

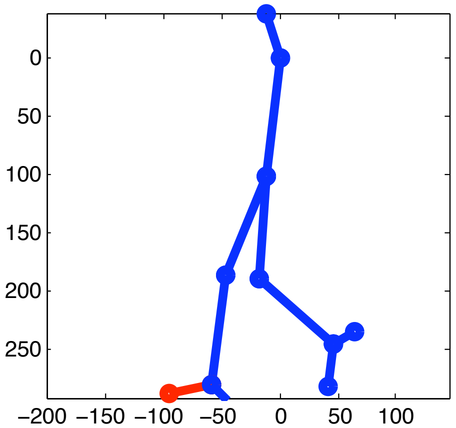
- Want:

$\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \ldots, \mathbf{x}_m^{o*}]$

body positions

$\mathbf{Y}^* = [\mathbf{y}_1^*, \ldots, \mathbf{y}_m^*]$

body configurations



- Posterior over positions and configurations:

$$p(\mathbf{X}^{o*}, \mathbf{Y}^* | E) \propto p(E | \mathbf{X}^{o*}, \mathbf{Y}^*) p(\mathbf{X}^{o*}) p(\mathbf{Y}^*).$$

Likelihood model
(partISM)

Tracklet Detection in Short Subsequences

TECHNISCHE UNIVERSITÄT DARMSTADT

- Given: $E = [E_1, \ldots, E_m]$

frame 1 frame 2 ... frame m

overlapping subsequences

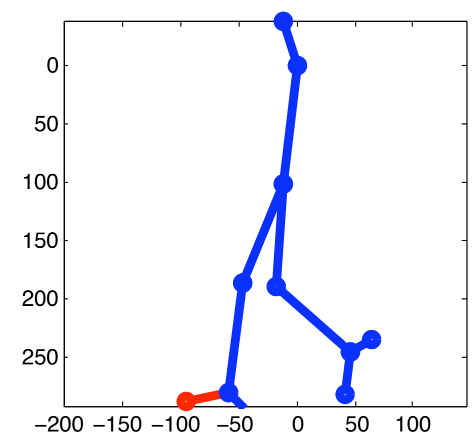
- Want:

$\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \ldots, \mathbf{x}_m^{o*}]$

body positions

$\mathbf{Y}^{*} = [\mathbf{y}_1^{*}, \ldots, \mathbf{y}_m^{*}]$

body configurations



- Posterior over positions and configurations:

$$p(\mathbf{X}^{o*}, \mathbf{Y}^{*} | E) \propto p(E | \mathbf{X}^{o*}, \mathbf{Y}^{*}) p(\mathbf{X}^{o*}) p(\mathbf{Y}^{*}).$$

Likelihood model (partISM)

speed prior (Gaussian)

Tracklet Detection in Short Subsequences

- Given: $E = [E_1, \ldots, E_m]$

frame 1 frame 2 frame m

overlapping subsequences

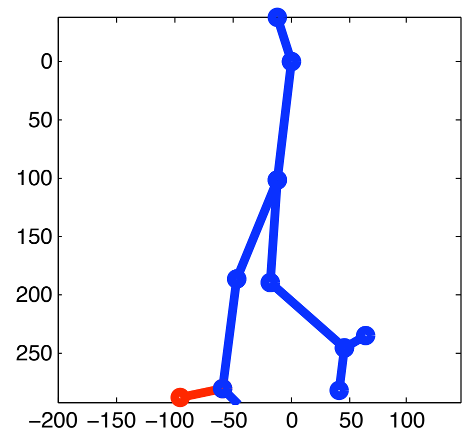
- Want:

$\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \ldots, \mathbf{x}_m^{o*}]$

body positions

$\mathbf{Y}^* = [\mathbf{y}_1^*, \ldots, \mathbf{y}_m^*]$

body configurations



- Posterior over positions and configurations:

$$p(\mathbf{X}^{o*}, \mathbf{Y}^*|E) \propto p(E|\mathbf{X}^{o*}, \mathbf{Y}^*)p(\mathbf{X}^{o*})p(\mathbf{Y}^*).$$

Likelihood model (partISM)

speed prior (Gaussian)

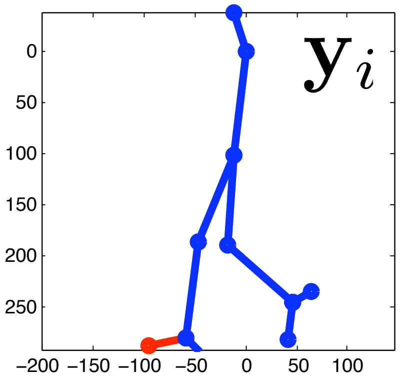
dynamical body model (hGPLVM)

Modeling Body Dynamics

- $\mathbf{Y}^*$ is very high-dimensional: Full body poses in m frames.
- Model the body dynamics using hierarchical Gaussian process latent variable model (hGPLVM) [Lawrence&Moore, ICML 2007]

Modeling Body Dynamics

- $\mathbf{Y}^*$ is very high-dimensional: Full body poses in m frames.
- Model the body dynamics using hierarchical Gaussian process latent variable model (hGPLVM) [Lawrence&Moore, ICML 2007]

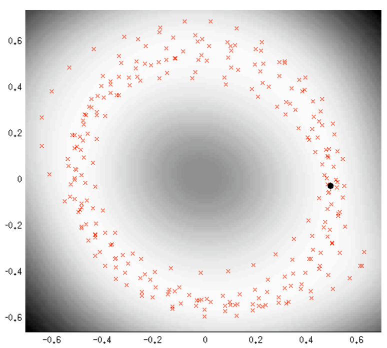


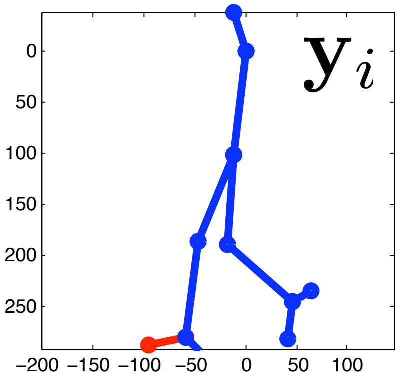
Configuration $\mathbf{Y}$ $\mathbf{Y} = [\mathbf{y}_i \in \mathbb{R}^D]$

Modeling Body Dynamics

TECHNISCHE UNIVERSITÄT DARMSTADT

- $\mathbf{Y}^*$ is very high-dimensional: Full body poses in m frames.
- Model the body dynamics using hierarchical Gaussian process latent variable model (hGPLVM) [Lawrence&Moore, ICML 2007]





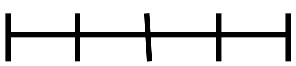
Latent space $\quad \mathbf{Z} \quad \mathbf{Z} = [\mathbf{z}_i \in \mathbb{R}^q]$

$$\mathbf{Z} \longrightarrow \mathbf{Y}$$

Configuration $\quad \mathbf{Y} \quad \mathbf{Y} = [\mathbf{y}_i \in \mathbb{R}^D]$

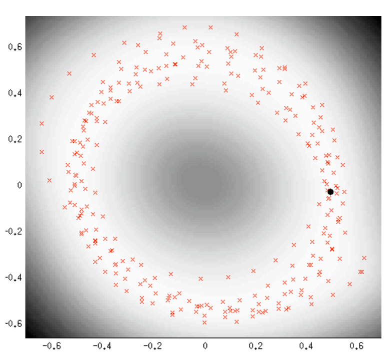
Modeling Body Dynamics

- $\mathbf{Y}^*$ is very high-dimensional: Full body poses in m frames.
- Model the body dynamics using hierarchical Gaussian process latent variable model (hGPLVM) [Lawrence&Moore, ICML 2007]



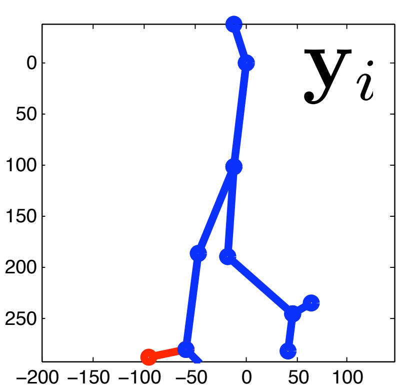
Time (frame #)

$\mathbf{T} = [t_i \in \mathbb{R}]$



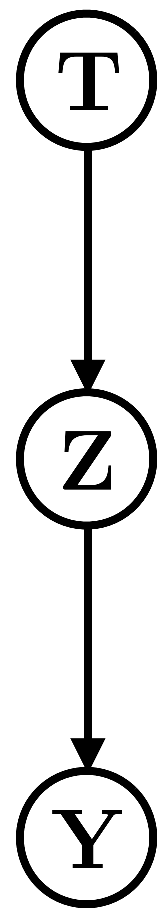
Latent space

$\mathbf{Z} = [\mathbf{z}_i \in \mathbb{R}^q]$



Configuration

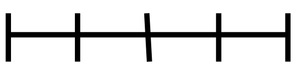
$\mathbf{Y} = [\mathbf{y}_i \in \mathbb{R}^D]$



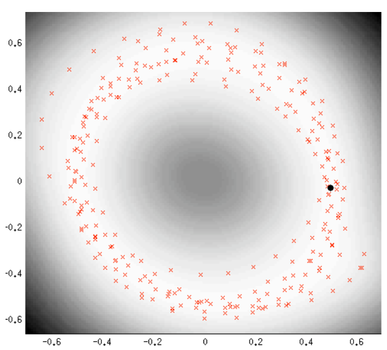
Modeling Body Dynamics

TECHNISCHE UNIVERSITÄT DARMSTADT

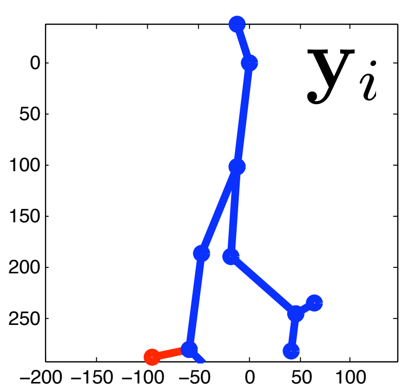
- $\mathbf{Y}^*$ is very high-dimensional: Full body poses in m frames.
- Model the body dynamics using hierarchical Gaussian process latent variable model (hGPLVM) [Lawrence&Moore, ICML 2007]



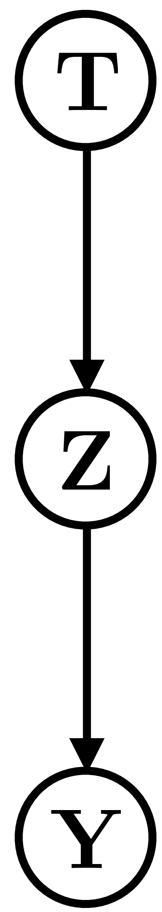
Time (frame #)



Latent space



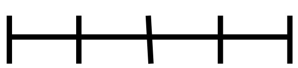
Configuration



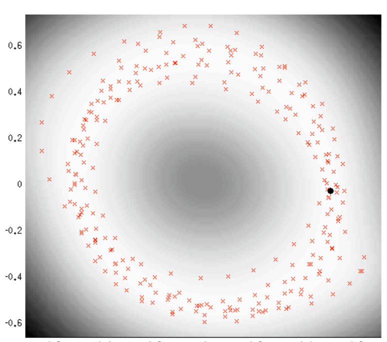
$$p(\mathbf{Y}|\mathbf{Z},\theta) = \prod_{i=1}^{D} \mathcal{N}(\mathbf{Y}_{:,i}|0,\mathbf{K}_{\mathbf{z}})$$

Modeling Body Dynamics

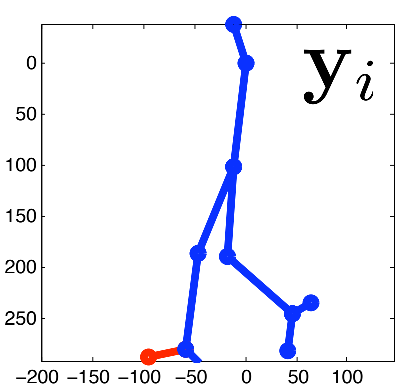
- $\mathbf{Y}^*$ is very high-dimensional: Full body poses in m frames.
- Model the body dynamics using hierarchical Gaussian process latent variable model (hGPLVM) [Lawrence&Moore, ICML 2007]



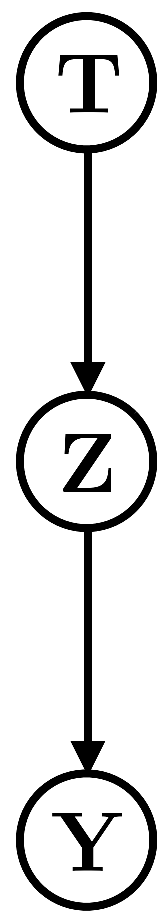
Time (frame #)



Latent space



Configuration

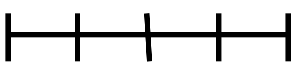


$$p(\mathbf{Z}|\mathbf{T},\hat{\theta}) = \prod_{i=1}^{q} \mathcal{N}(\mathbf{Z}_{:,i}|0,\mathbf{K}_{\mathbf{T}})$$

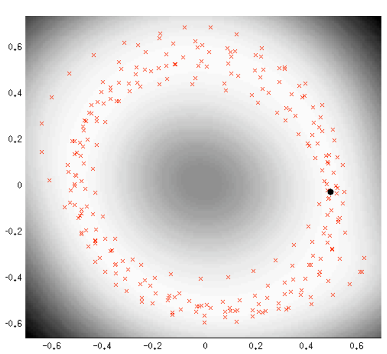
$$p(\mathbf{Y}|\mathbf{Z},\theta) = \prod_{i=1}^{D} \mathcal{N}(\mathbf{Y}_{:,i}|0,\mathbf{K}_{\mathbf{Z}})$$

Modeling Body Dynamics

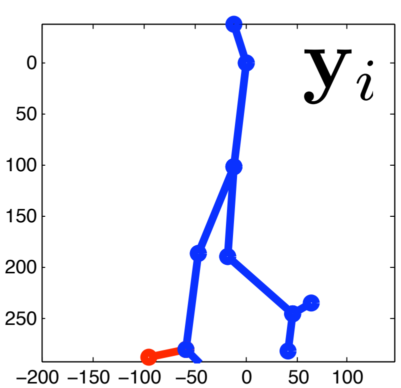
- $\mathbf{Y}^*$ is very high-dimensional: Full body poses in m frames.
- Model the body dynamics using hierarchical Gaussian process latent variable model (hGPLVM) [Lawrence&Moore, ICML 2007]



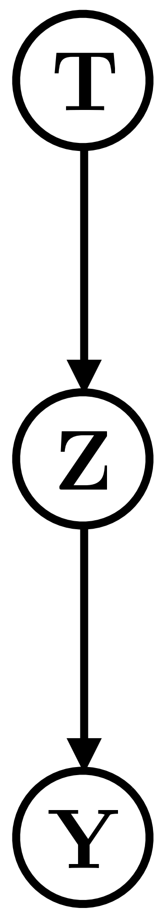
Time (frame #)



Latent space



Configuration



$$p(\mathbf{Z}|\mathbf{T},\hat{\theta}) = \prod_{i=1}^{q} \mathcal{N}(\mathbf{Z}_{:,i}|0,\mathbf{K_T})$$

training

$$p(\mathbf{Y}|\mathbf{Z},\theta) = \prod_{i=1}^{D} \mathcal{N}(\mathbf{Y}_{:,i}|0,\mathbf{K_Z})$$

Tracklet Detection

- Tracklets are local maxima of:

$$p(\mathbf{X}^{o*}, \mathbf{Y}^*|E) \propto p(E|\mathbf{X}^{o*}, \mathbf{Y}^*)p(\mathbf{X}^{o*})p(\mathbf{Y}^*).$$

- Local maxima can be found using standard non-linear optimization (e.g. conjugate gradients).
- **How can we provide good initial hypotheses for optimization?**

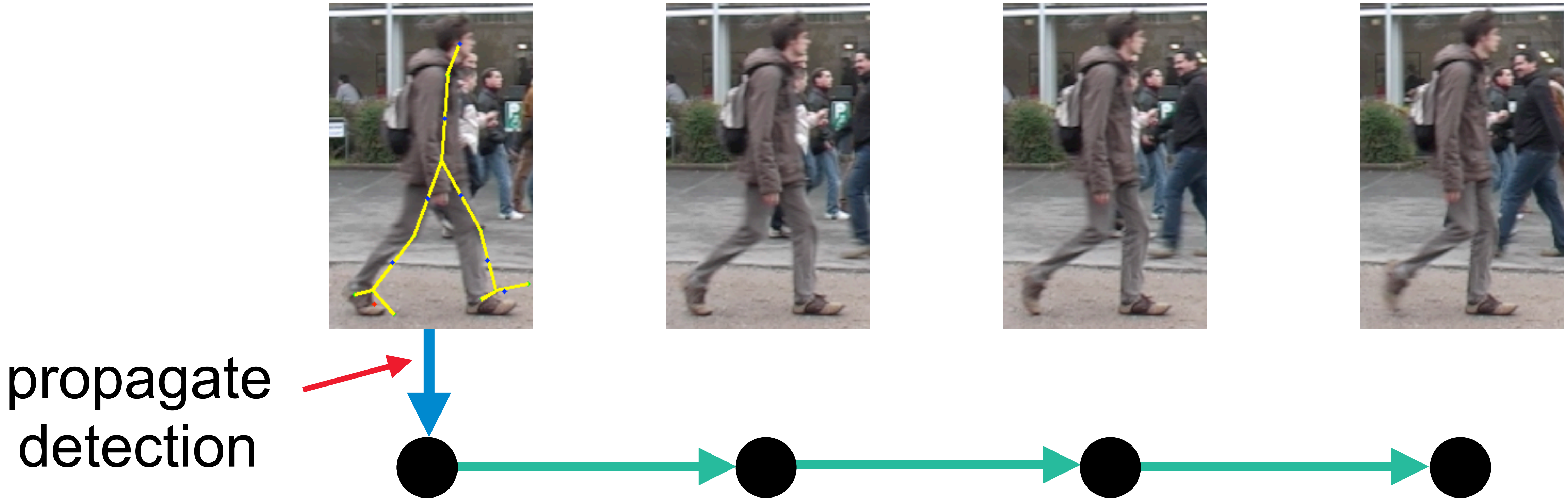
Tracklet Detection

Tracklet Detection

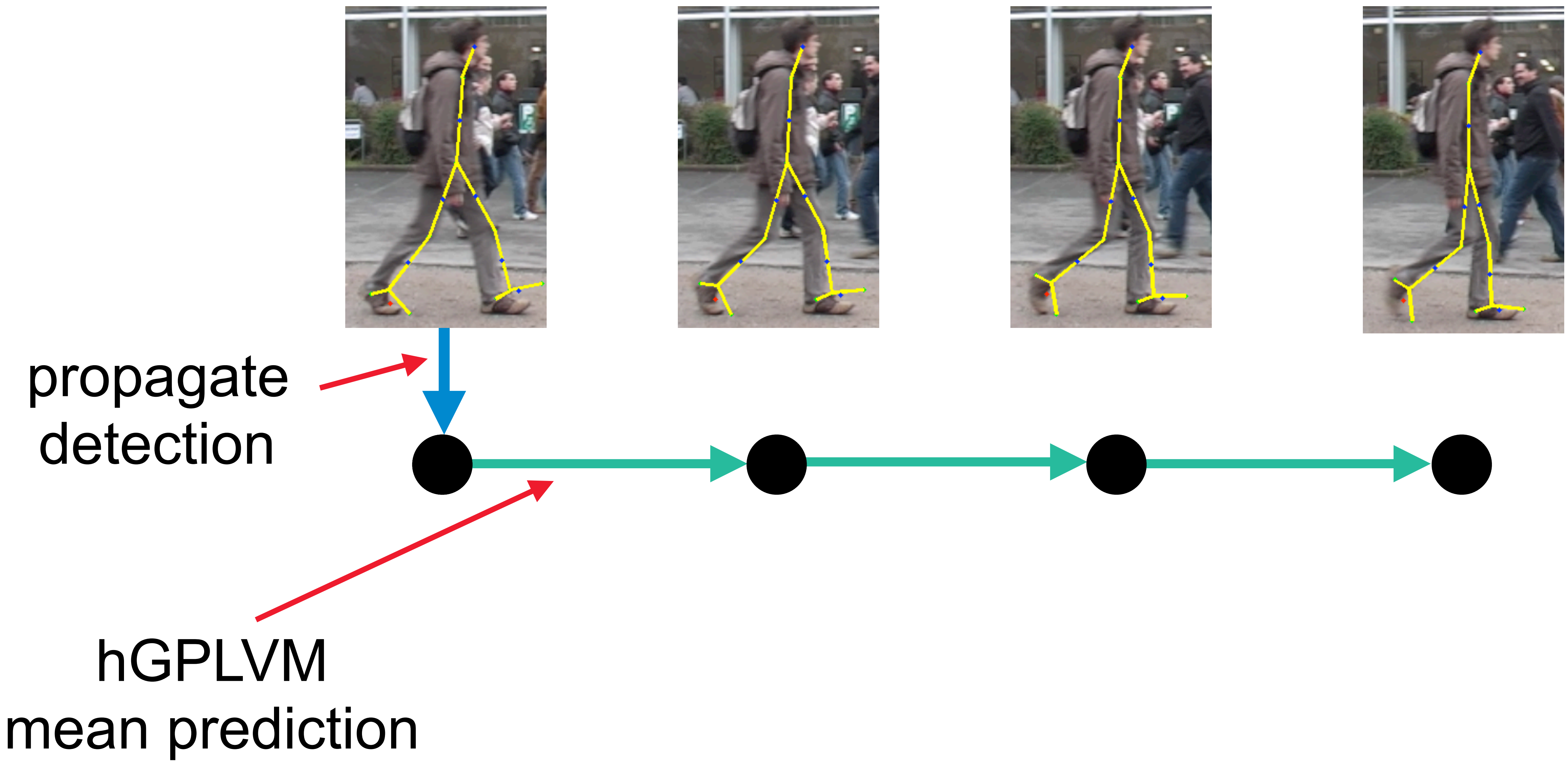
Tracklet Detection

Tracklet Detection

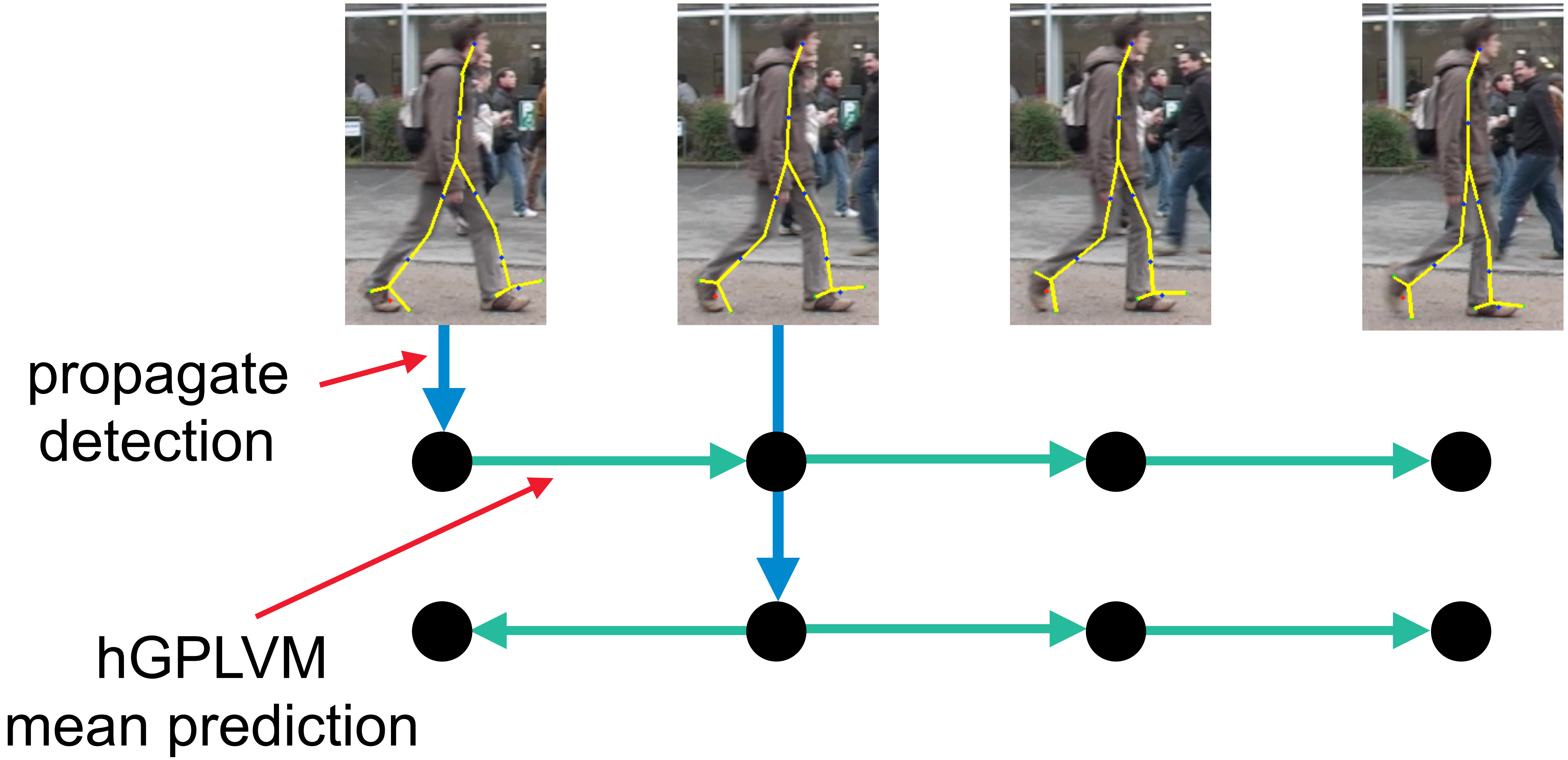
TECHNISCHE UNIVERSITÄT DARMSTADT



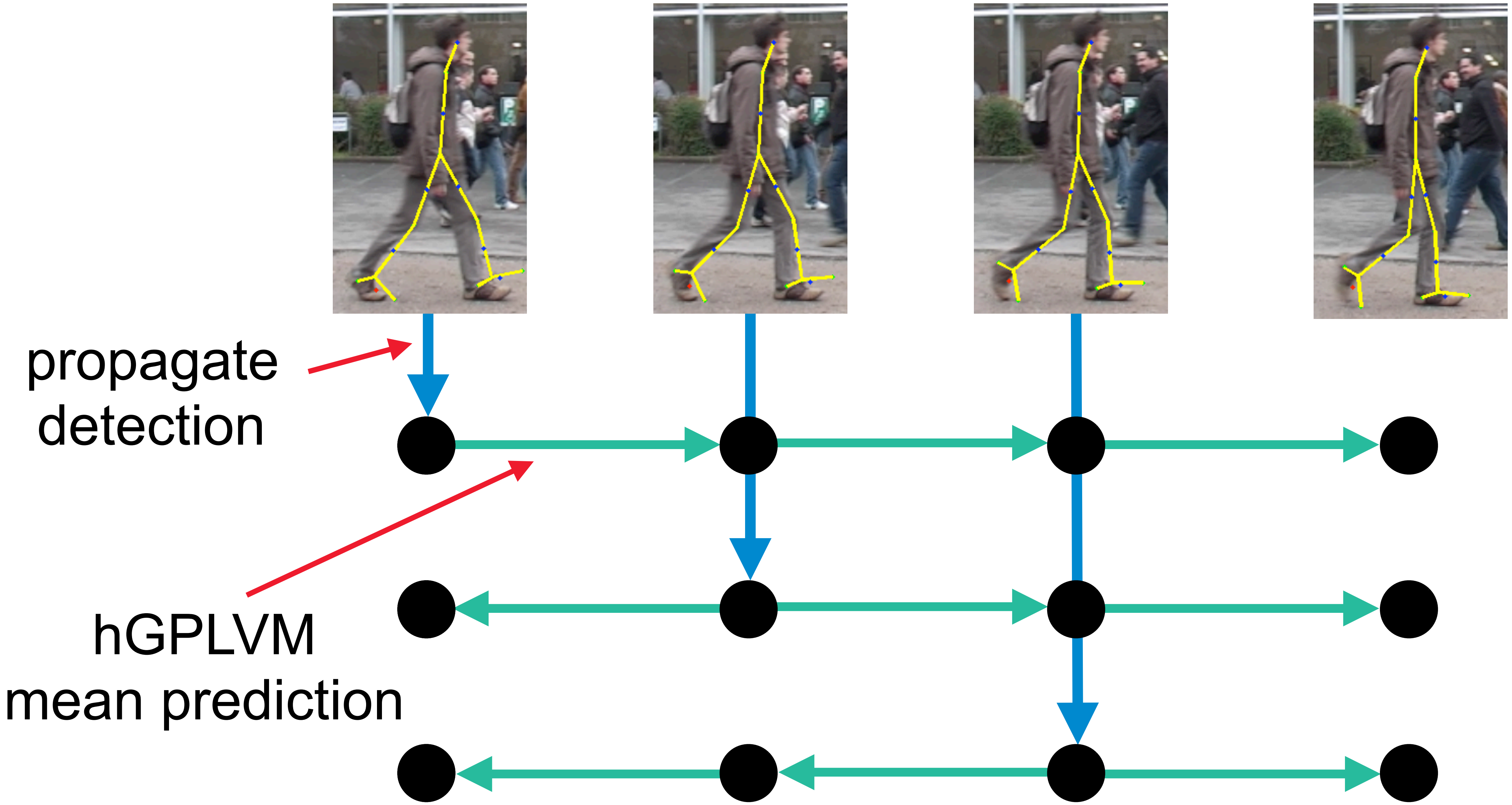
Tracklet Detection



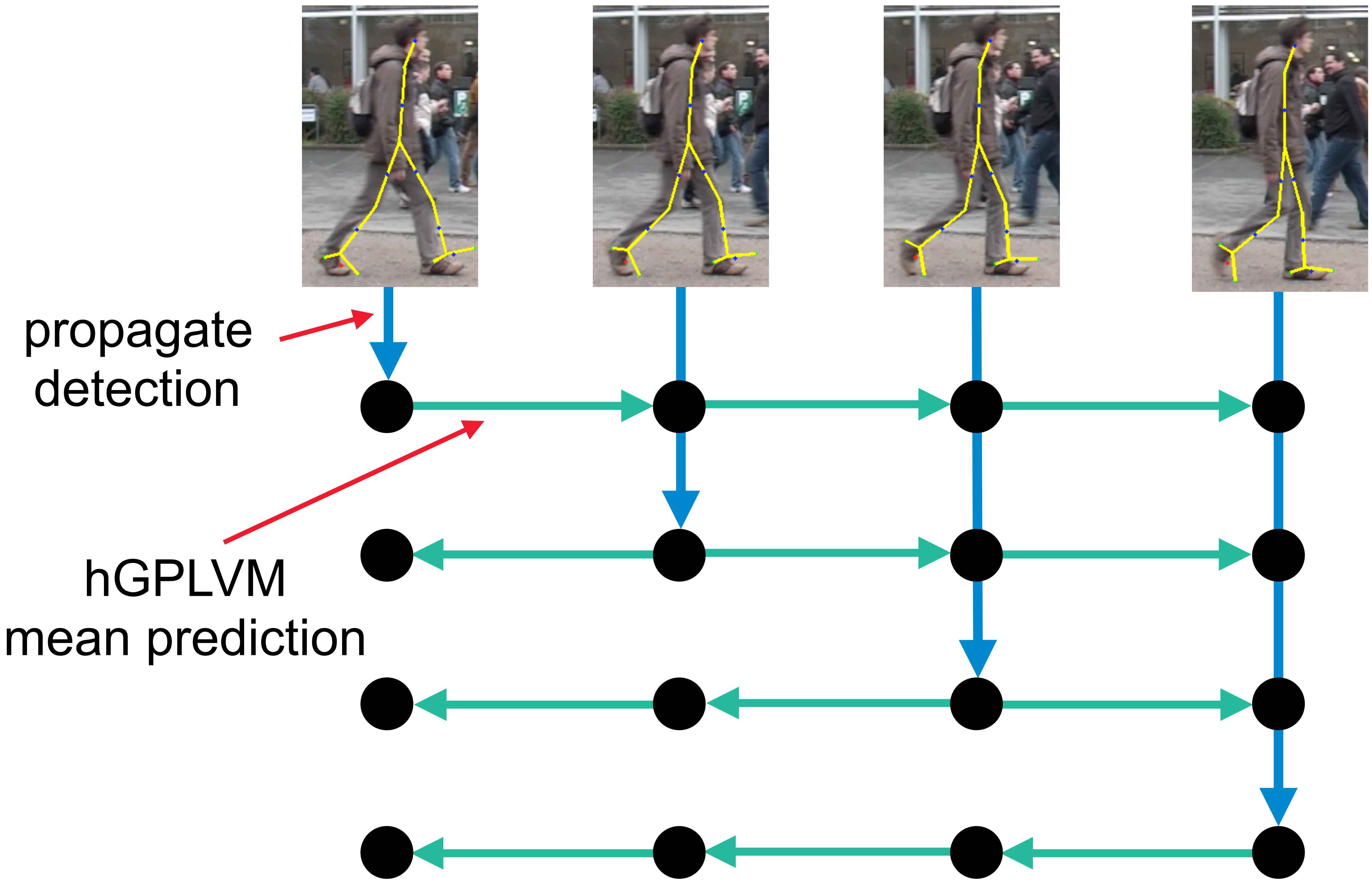
Tracklet Detection



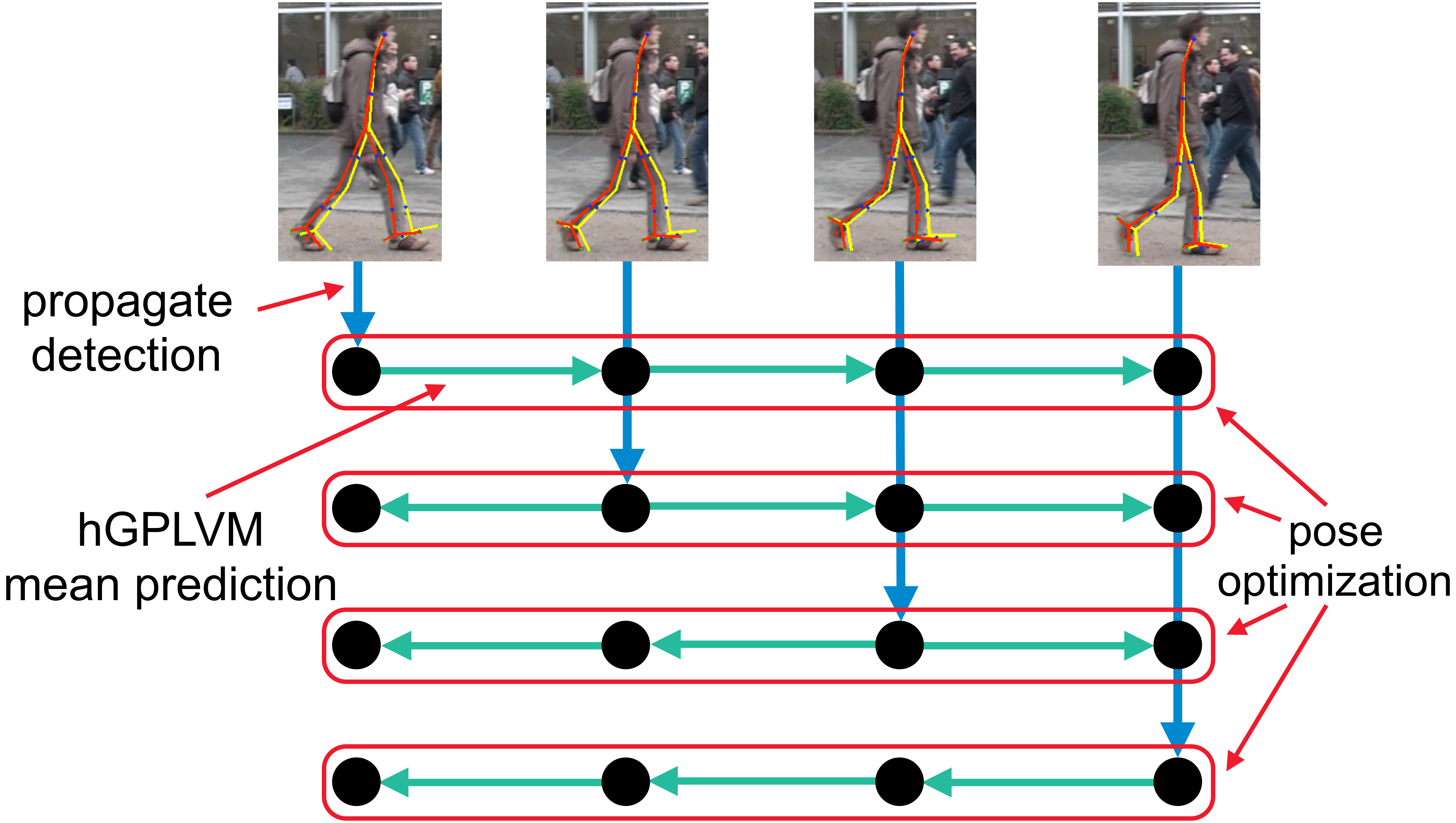
Tracklet Detection



Tracklet Detection



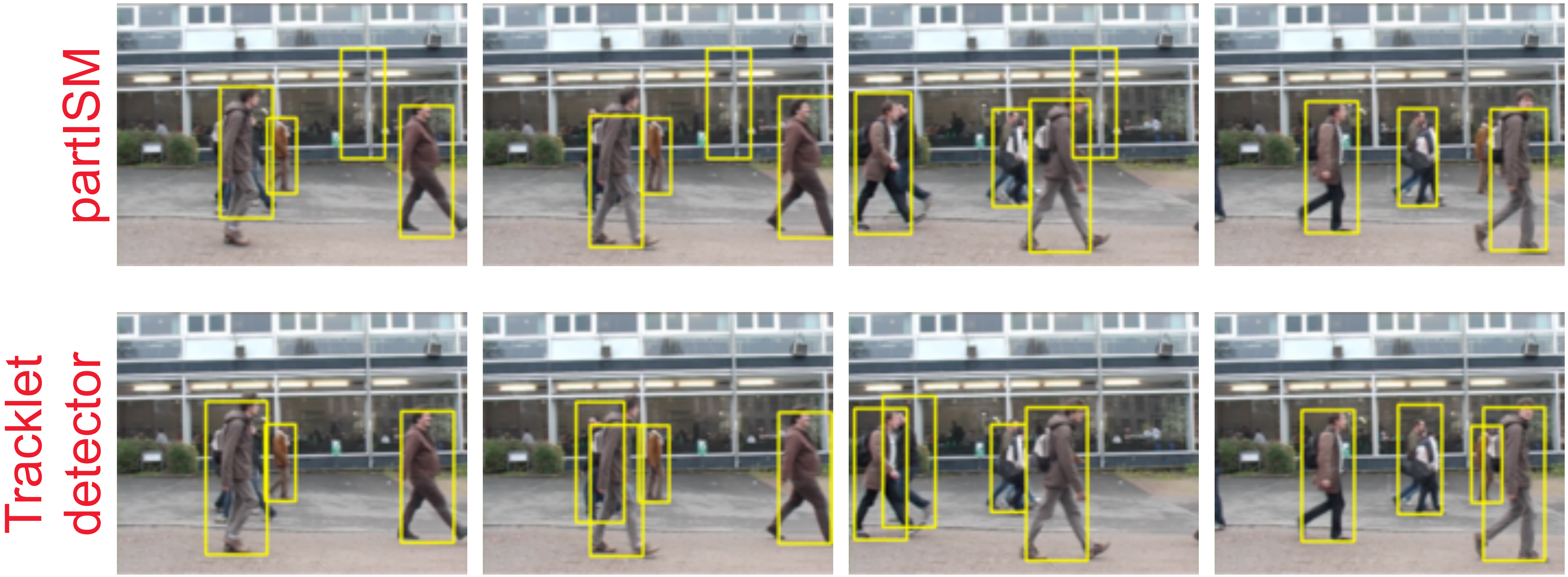
Tracklet Detection



Single-Frame Detector vs. Tracklet Detector

TECHNISCHE UNIVERSITÄT DARMSTADT

- At equal error rate:



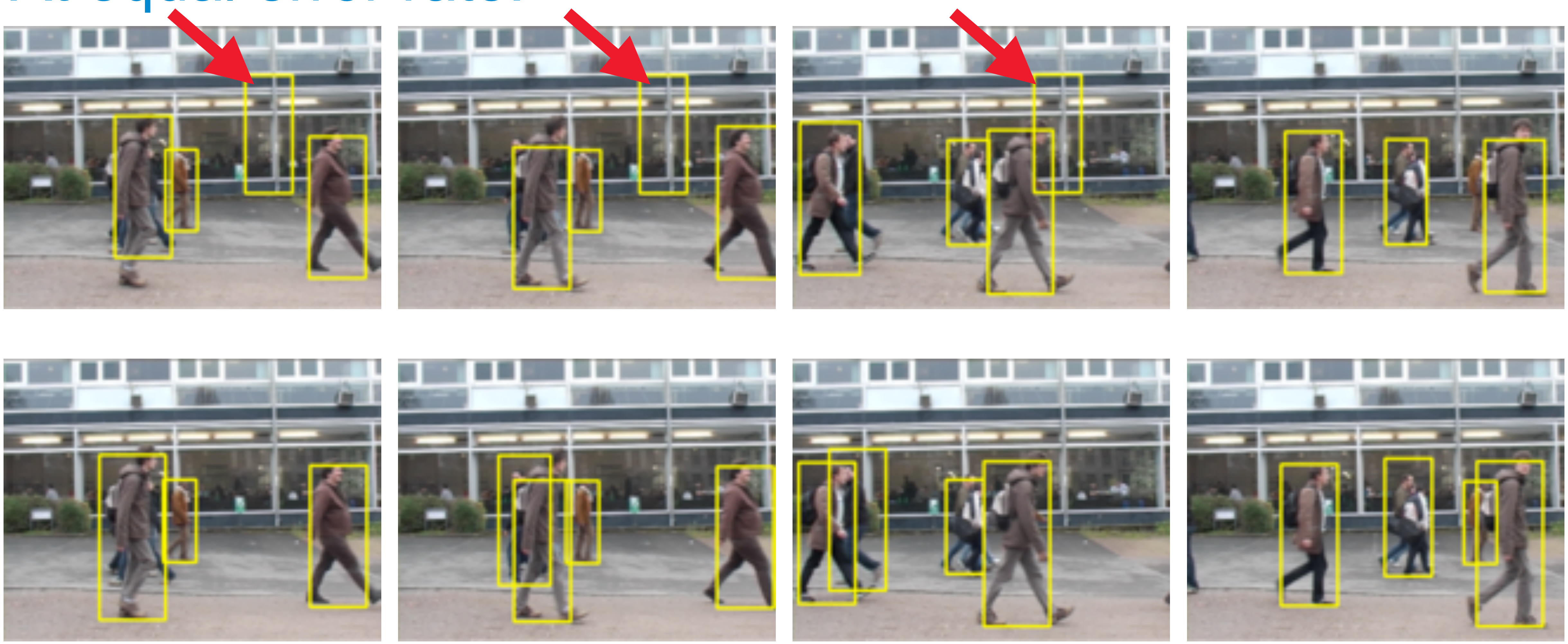
  - Fewer false positives.
  - More robust detection of partially occluded people.

Single-Frame Detector vs. Tracklet Detector

- At equal error rate:

partISM

Tracklet detector



- Fewer false positives.
- More robust detection of partially occluded people.

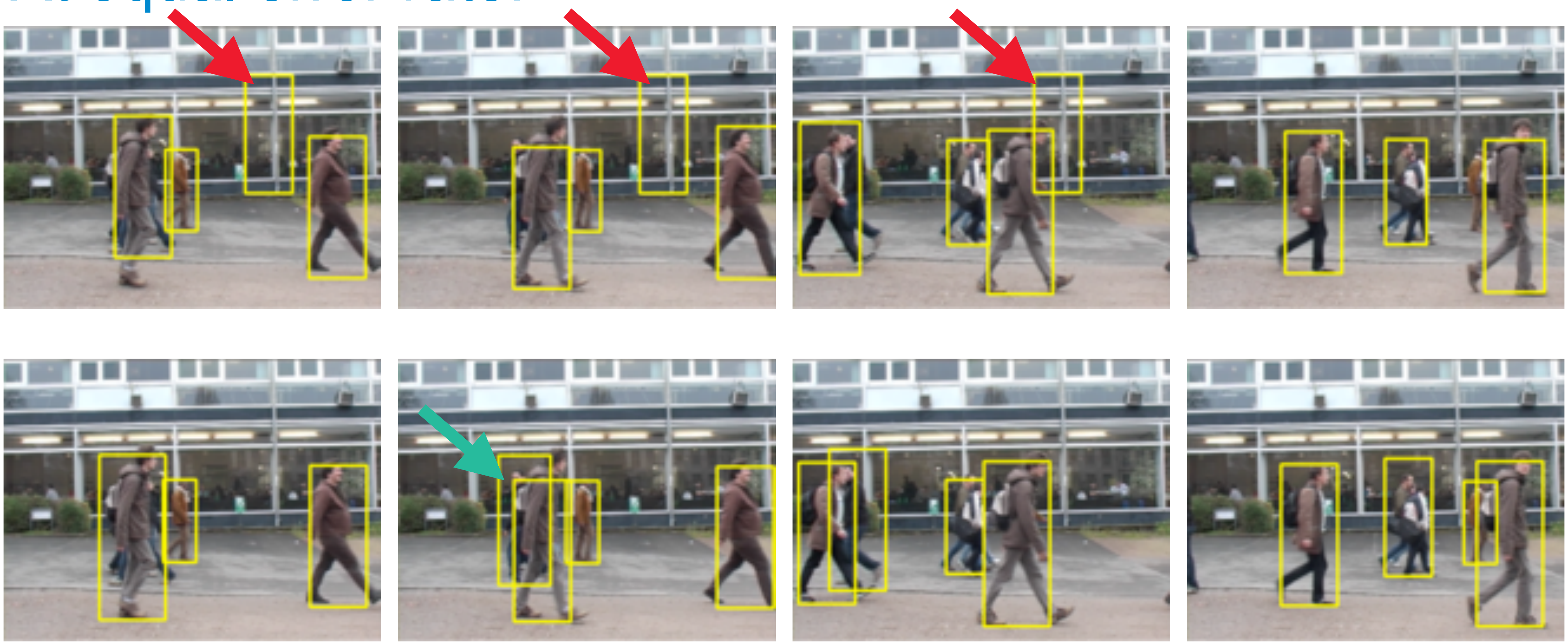
Single-Frame Detector vs. Tracklet Detector

TECHNISCHE UNIVERSITÄT DARMSTADT

- At equal error rate:

partISM

Tracklet detector



  - Fewer false positives.
  - More robust detection of partially occluded people.

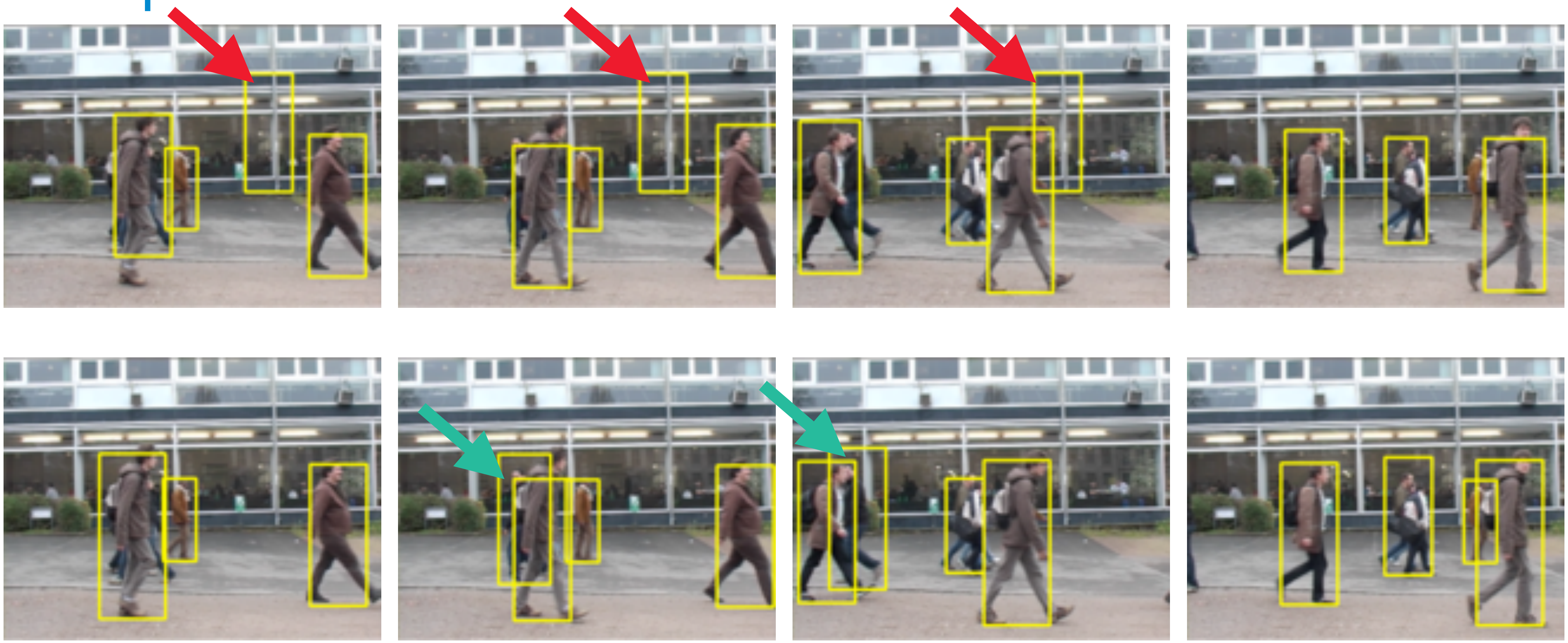
Single-Frame Detector vs. Tracklet Detector

TECHNISCHE UNIVERSITÄT DARMSTADT

- At equal error rate:

partISM

Tracklet detector



  - Fewer false positives.
  - More robust detection of partially occluded people.

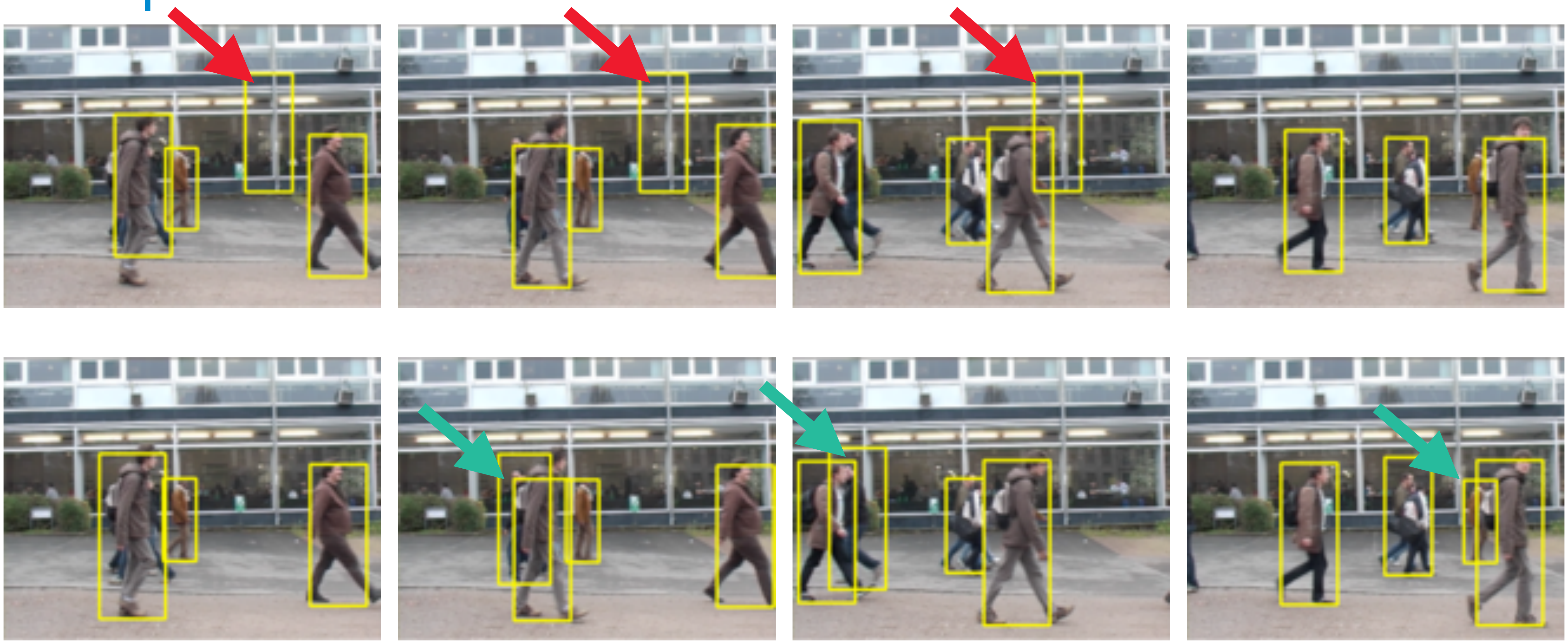
Single-Frame Detector vs. Tracklet Detector

TECHNISCHE UNIVERSITÄT DARMSTADT

- At equal error rate:

partISM

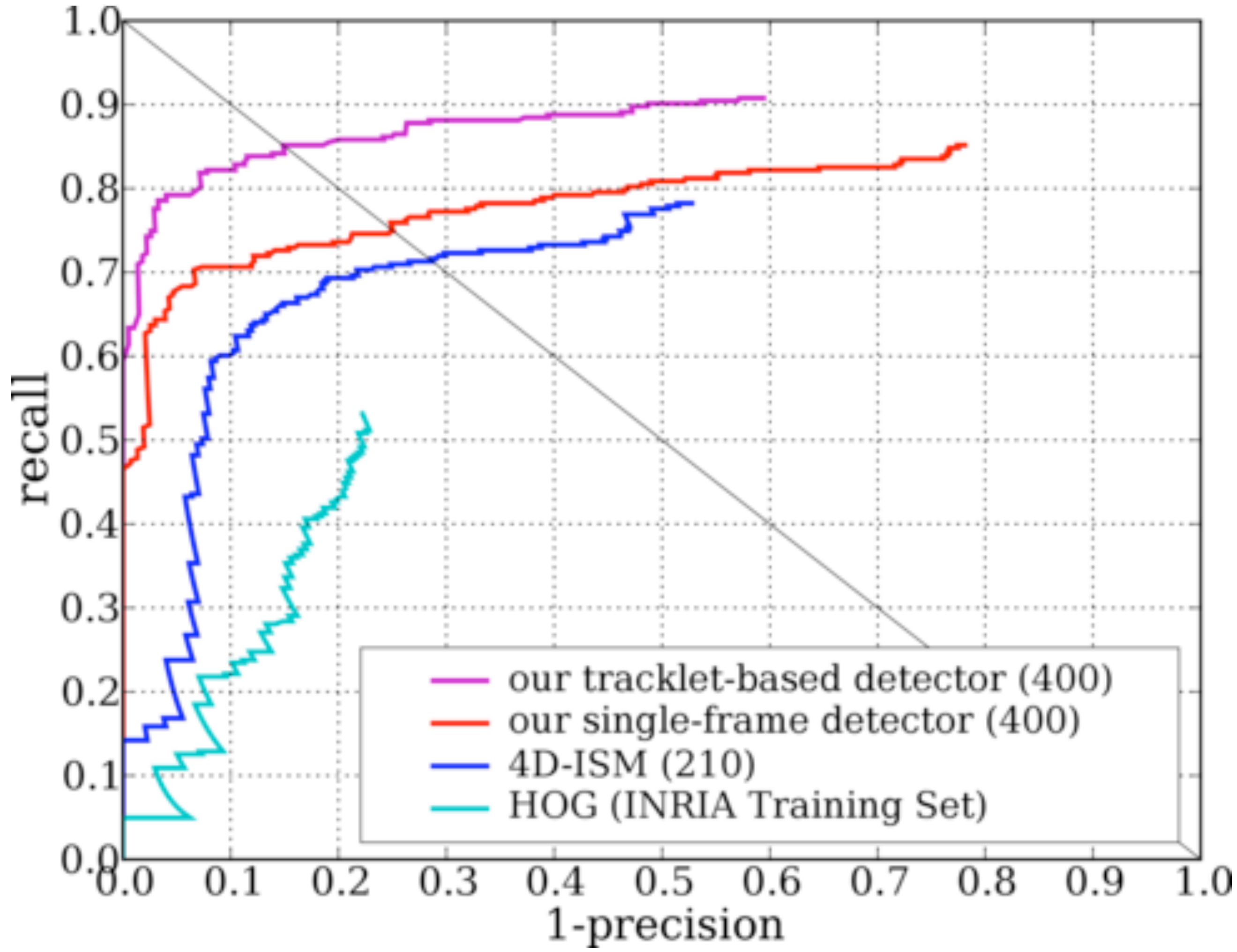
Tracklet detector



  - Fewer false positives.
  - More robust detection of partially occluded people.

Detection Performance

TECHNISCHE
UNIVERSITÄT
DARMSTADT



TUD campus data

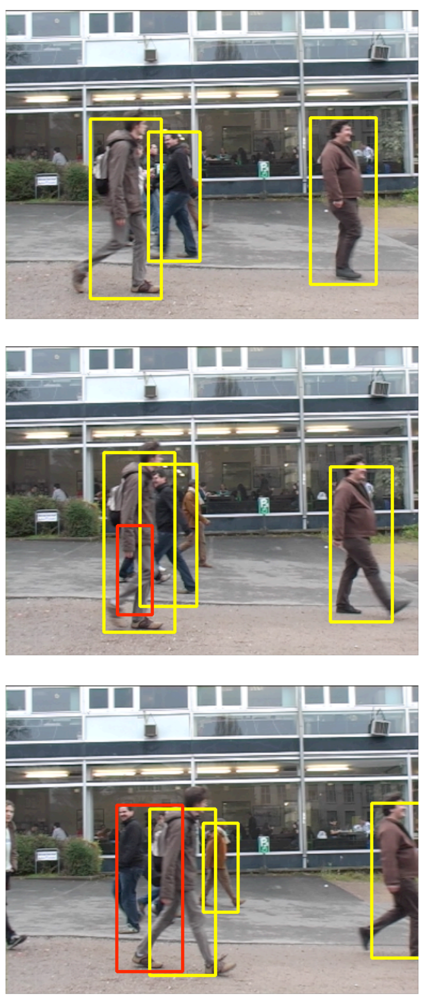
With occlusions (up to 50%)

- Significant improvement over single-frame detector.
  - Also at high precision levels.

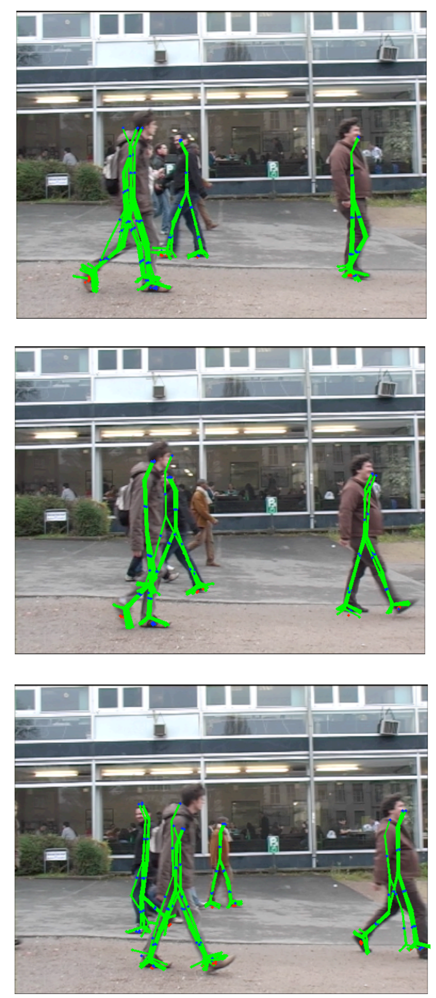
Overview

Three stages of our multi-person detection and tracking system:

1. Single-frame detection



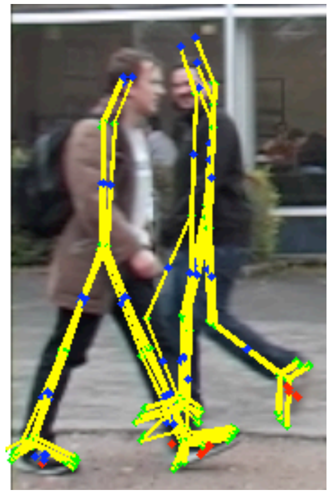
2. Tracklet detection



3. Tracking through occlusion

Tracks from Overlapping Tracklets

t



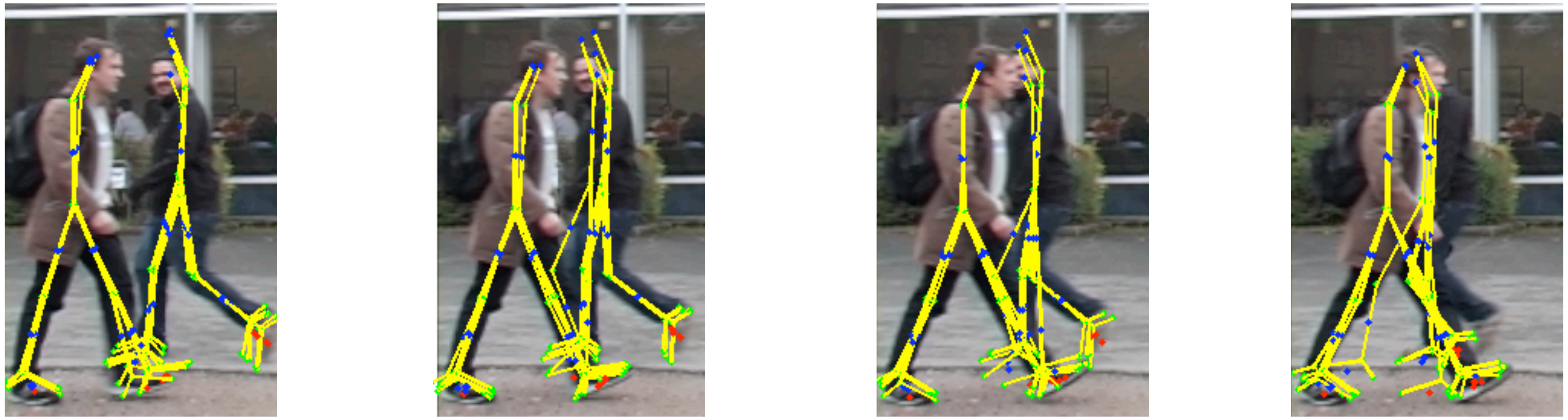
$t+1$

$t+2$

$t+3$

...

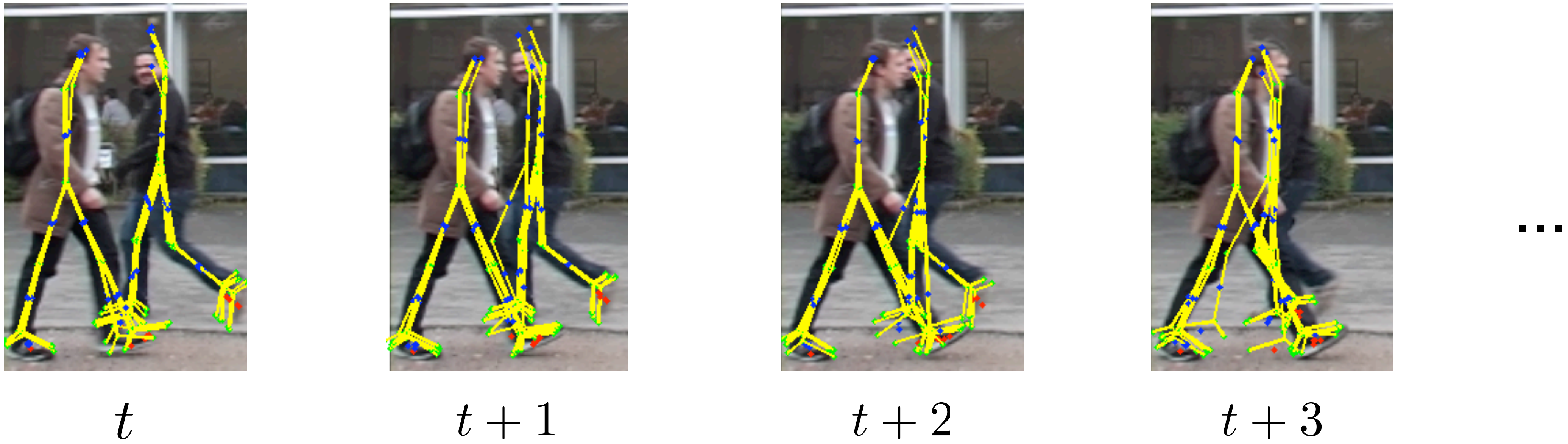
Tracks from Overlapping Tracklets

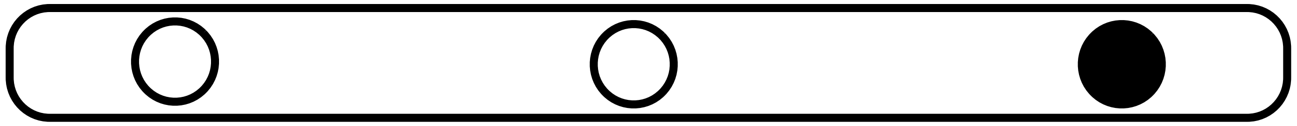


t $t+1$ $t+2$ $t+3$ …

Candidate poses from all overlapping tracklets

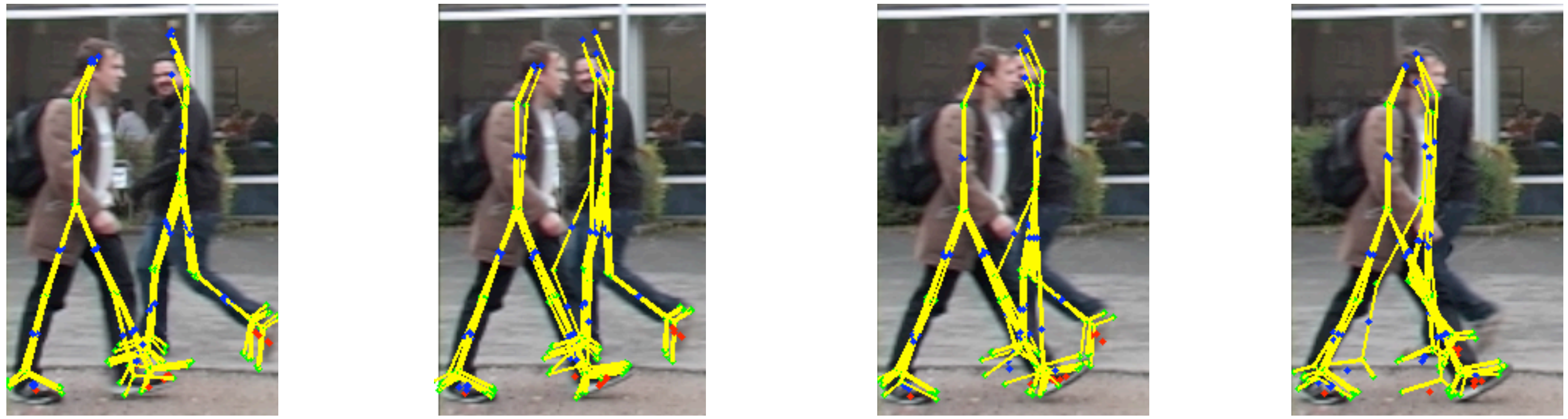
Tracks from Overlapping Tracklets



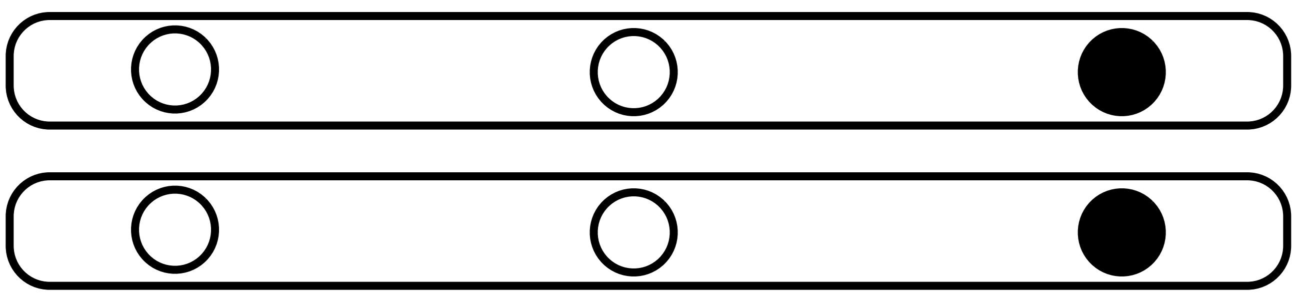


Candidate poses from all overlapping tracklets

Tracks from Overlapping Tracklets

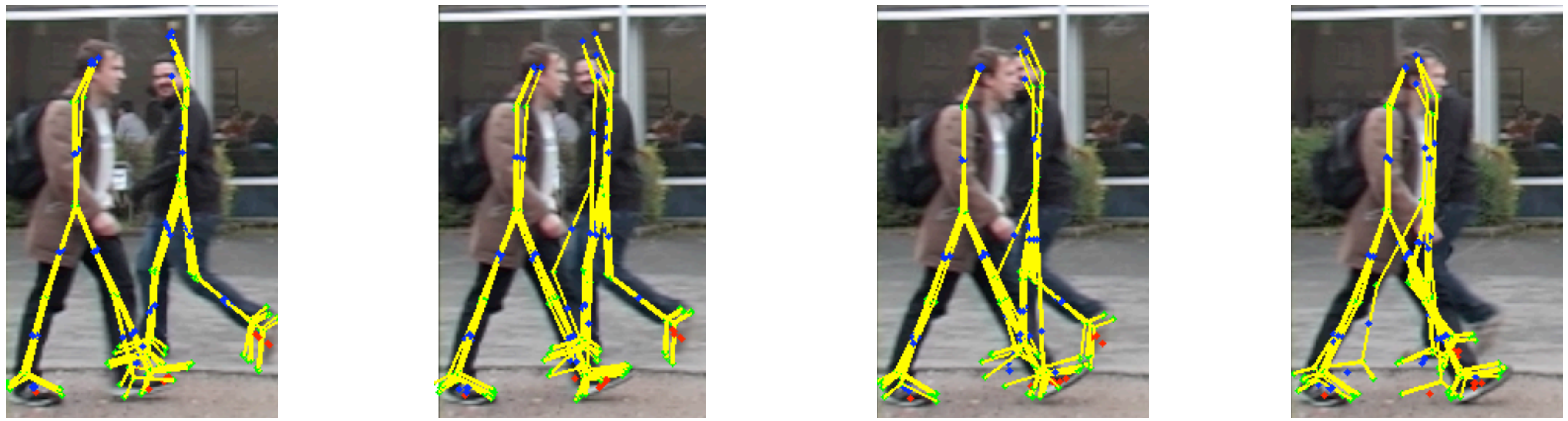


t $t+1$ $t+2$ $t+3$ ...

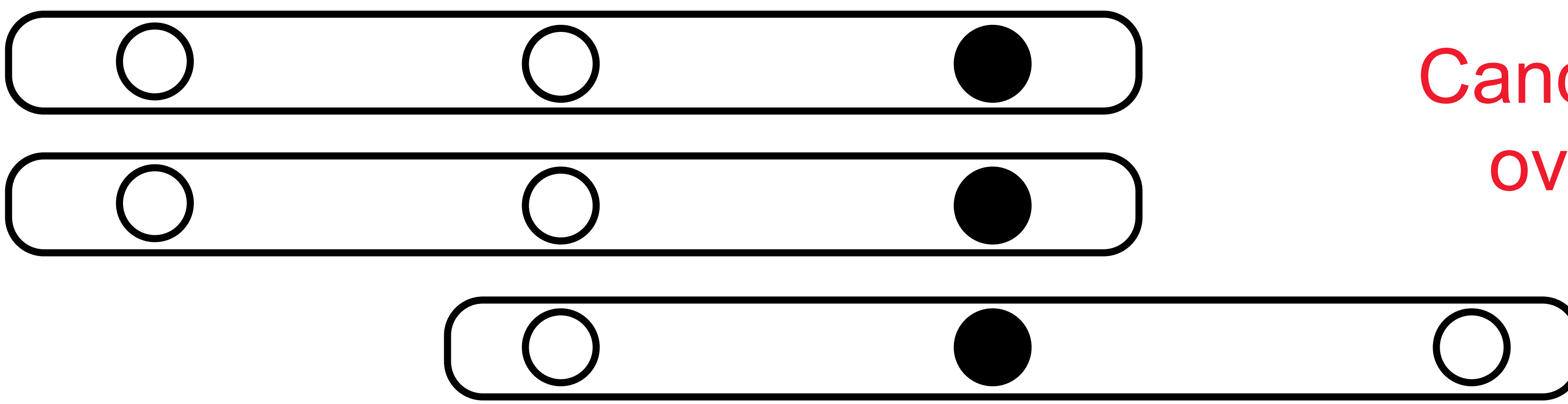


Candidate poses from all overlapping tracklets

Tracks from Overlapping Tracklets

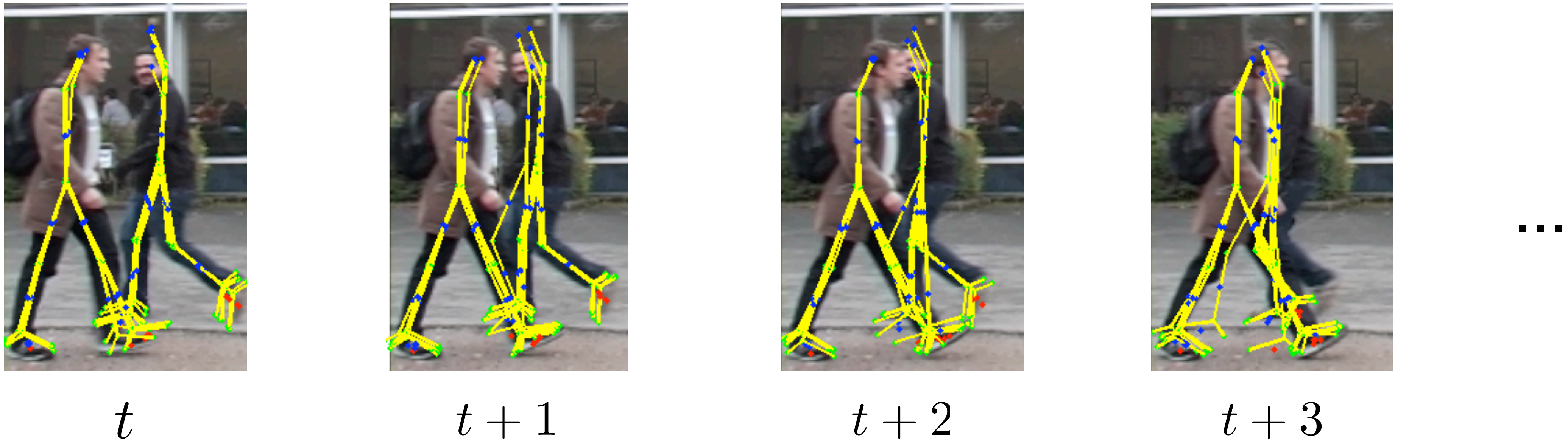


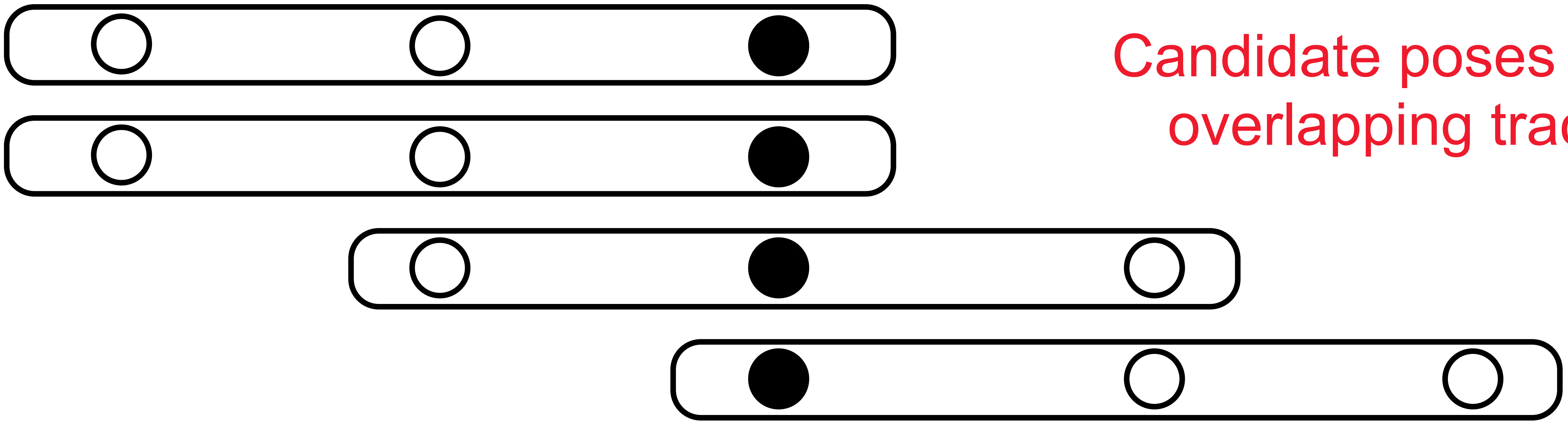
t $t+1$ $t+2$ $t+3$ ...



Candidate poses from all overlapping tracklets

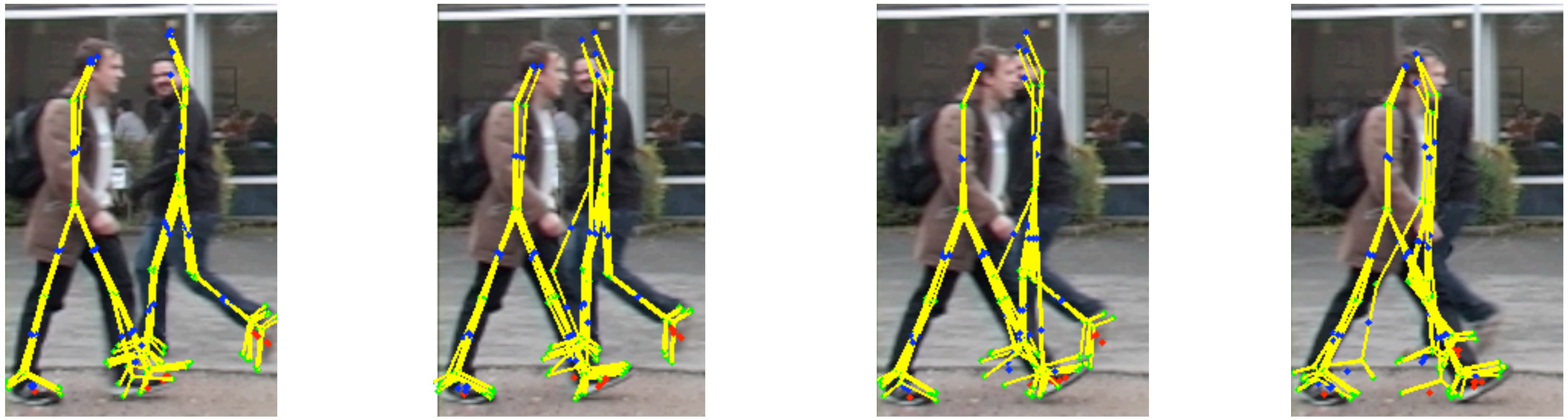
Tracks from Overlapping Tracklets



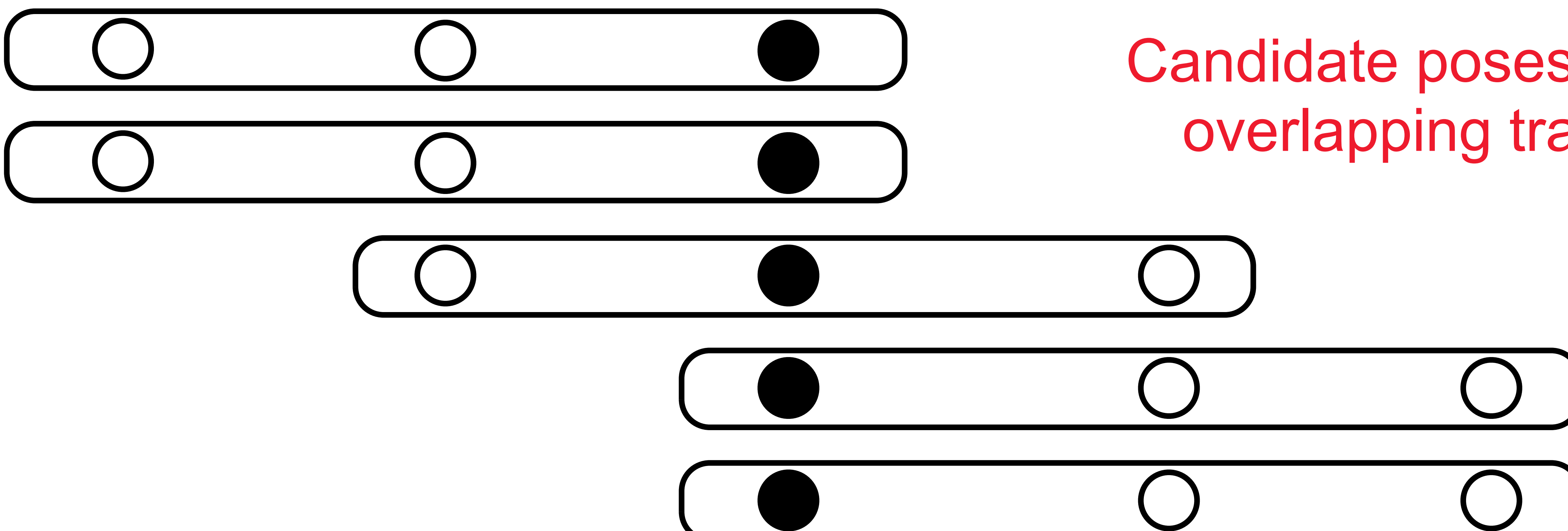


Candidate poses from all overlapping tracklets

Tracks from Overlapping Tracklets

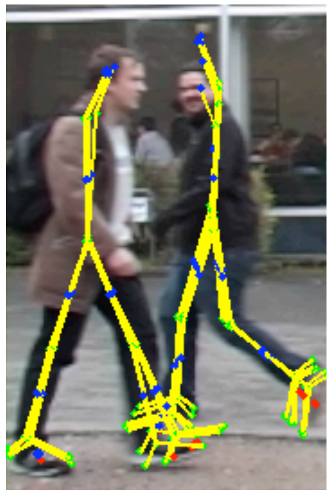


t $t+1$ $t+2$ $t+3$ ...

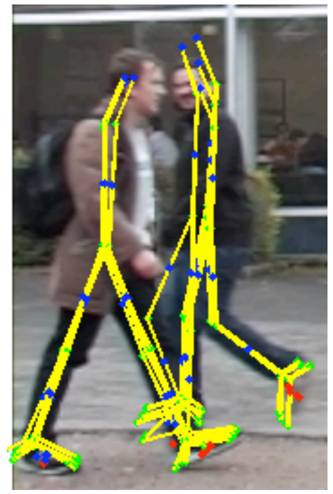


Candidate poses from all overlapping tracklets

Tracks from Overlapping Tracklets

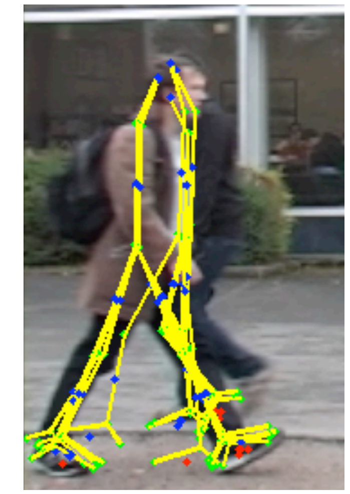


t



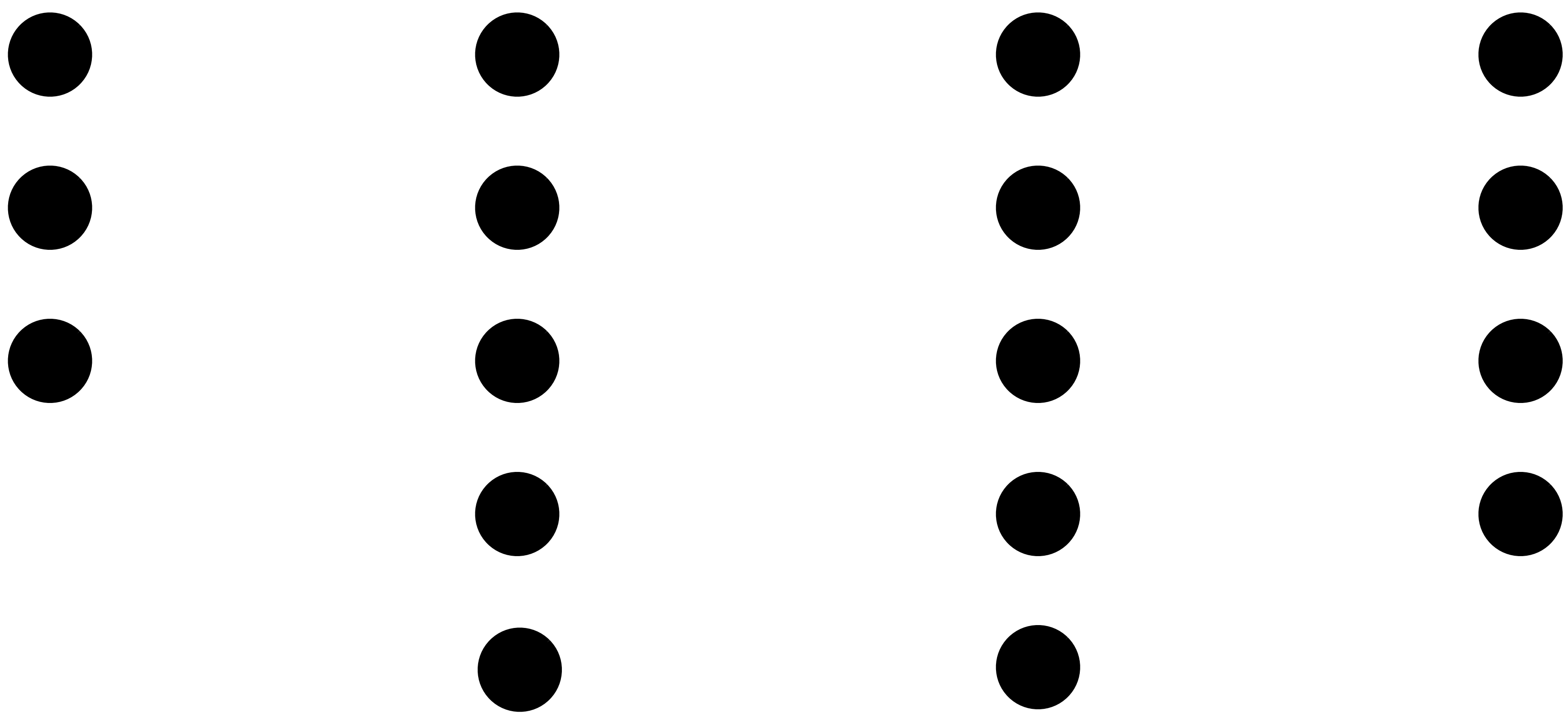
$t+1$

$t+2$

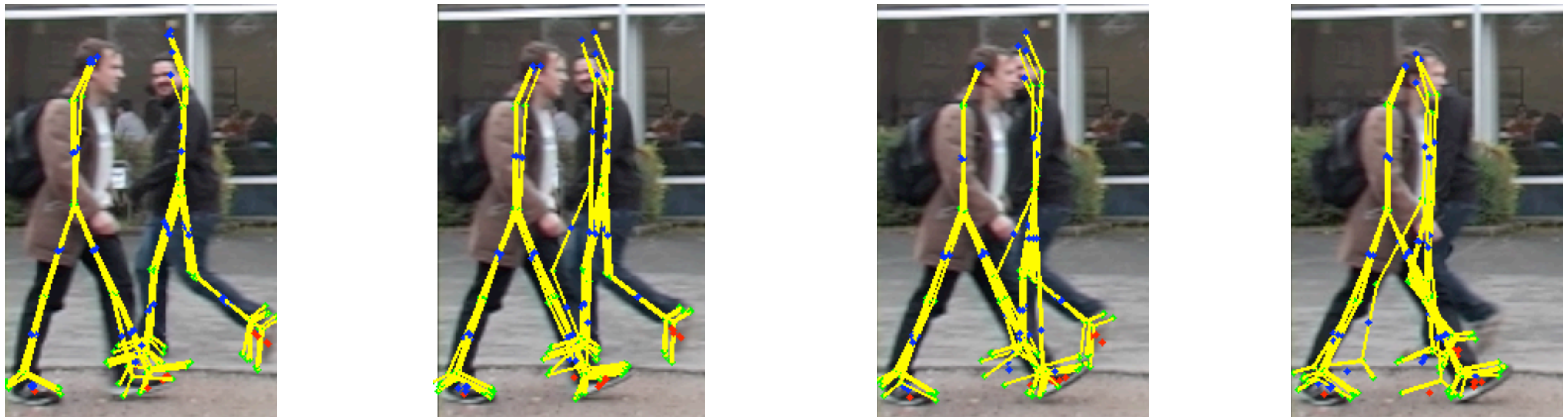


$t+3$

...

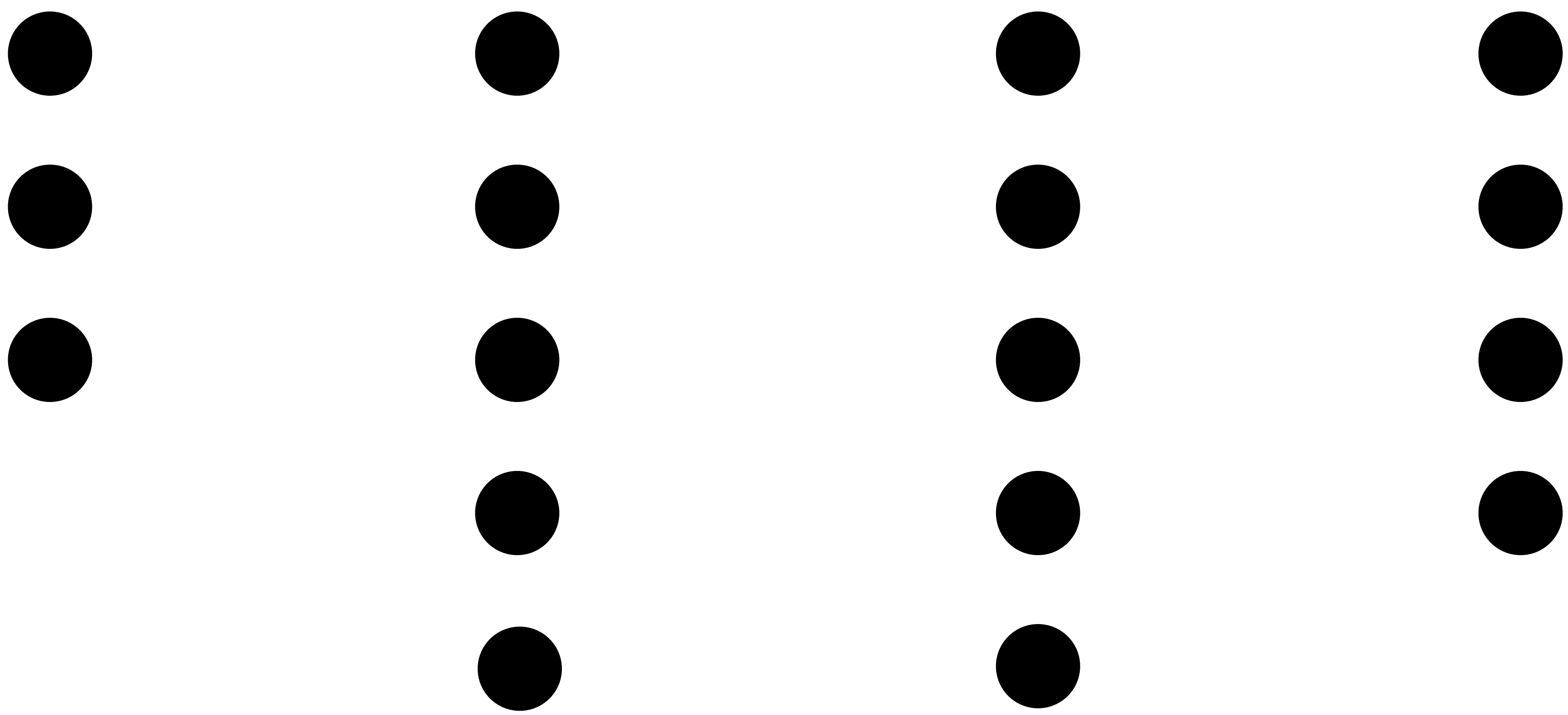


Tracks from Overlapping Tracklets

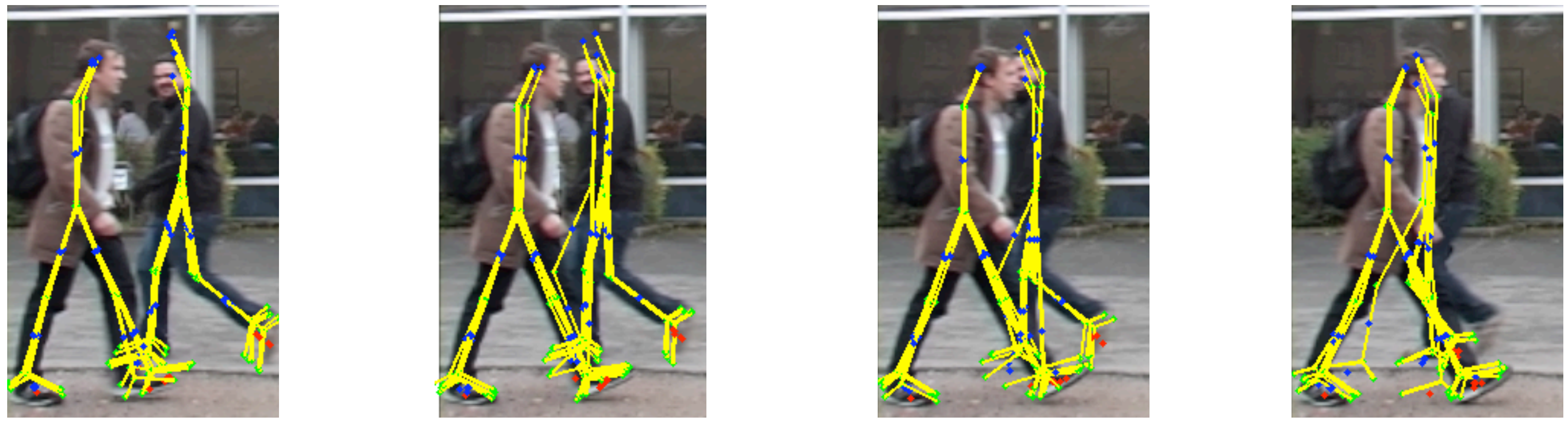


t $t+1$ $t+2$ $t+3$ ...

Viterbi Decoding

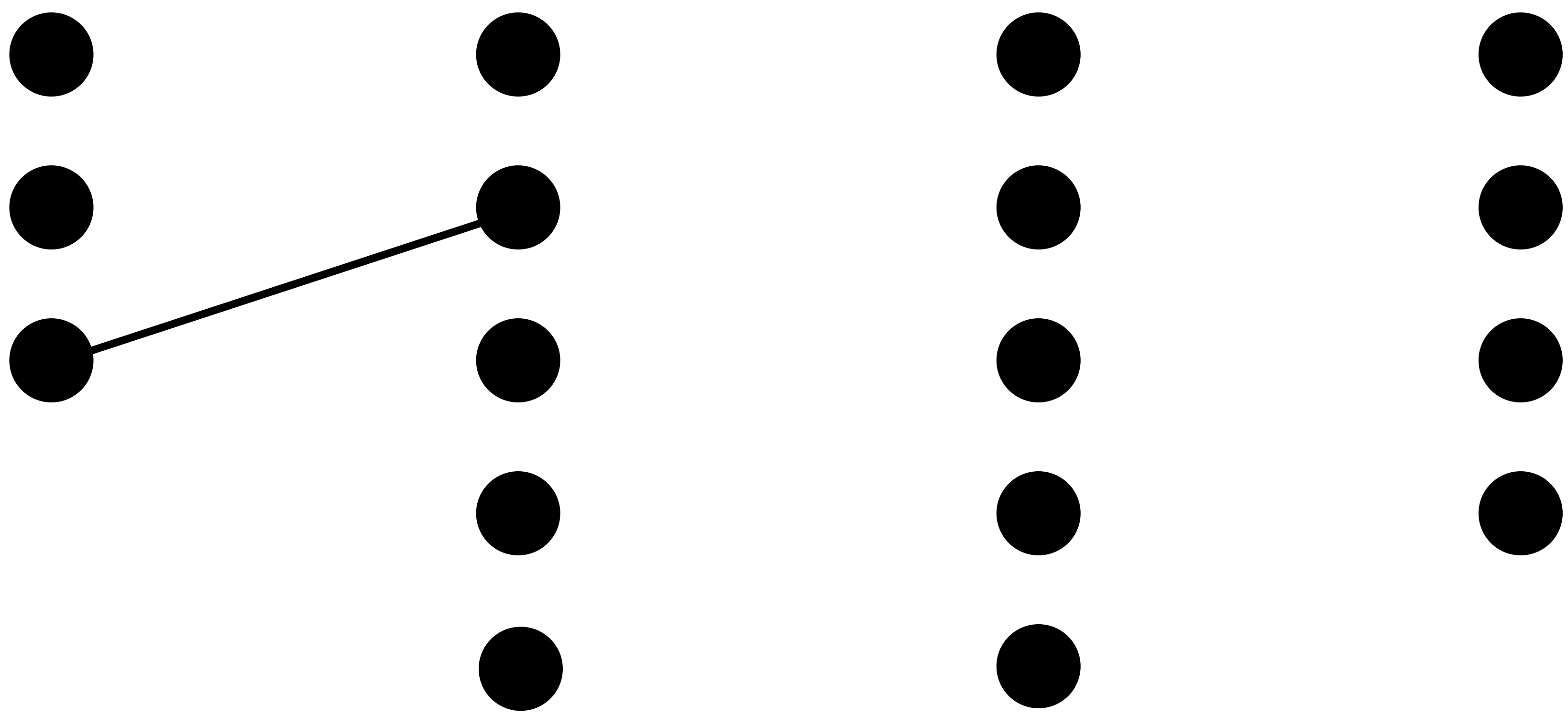


Tracks from Overlapping Tracklets

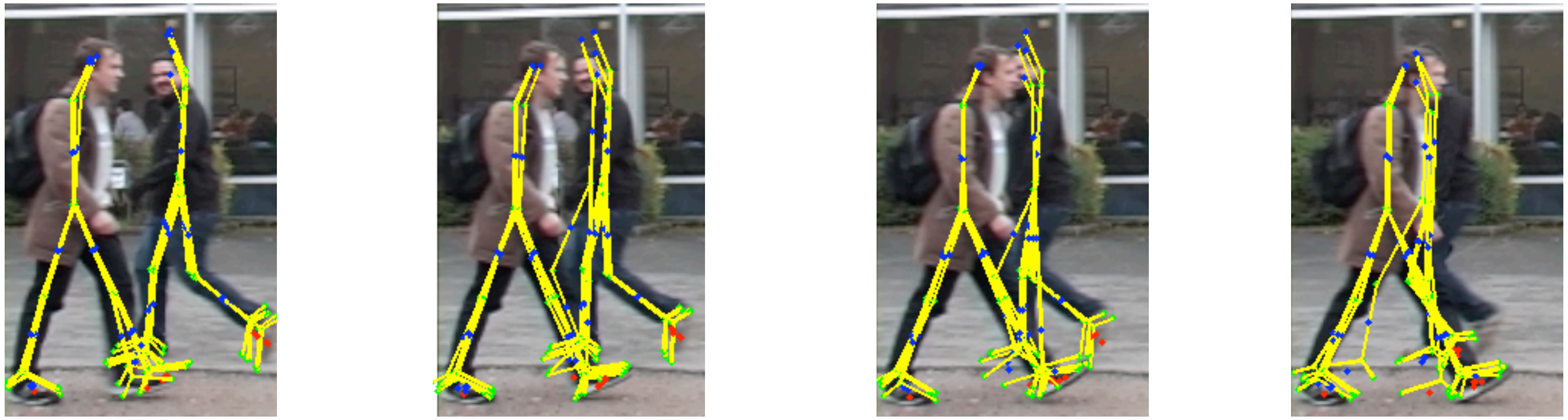


t $t+1$ $t+2$ $t+3$ …

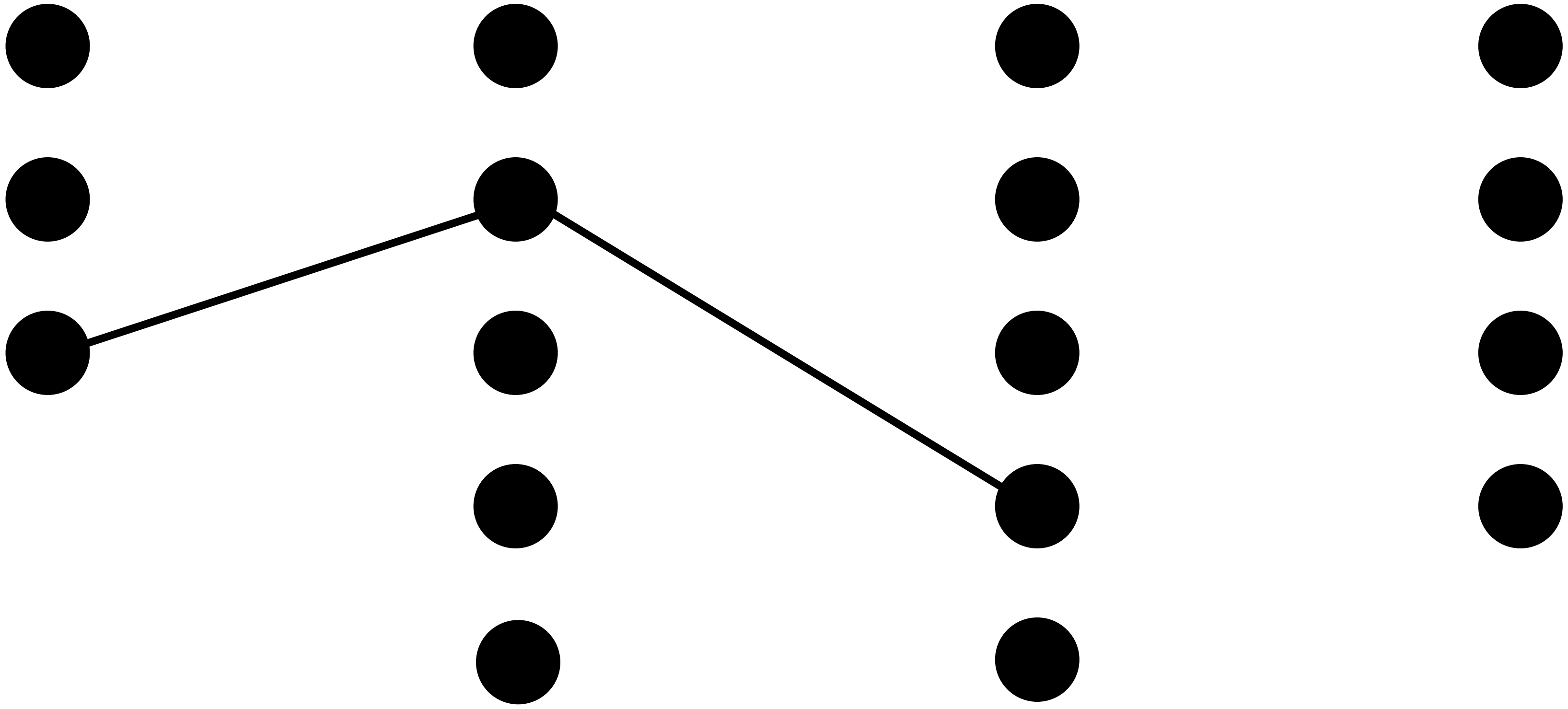
Viterbi Decoding



Tracks from Overlapping Tracklets

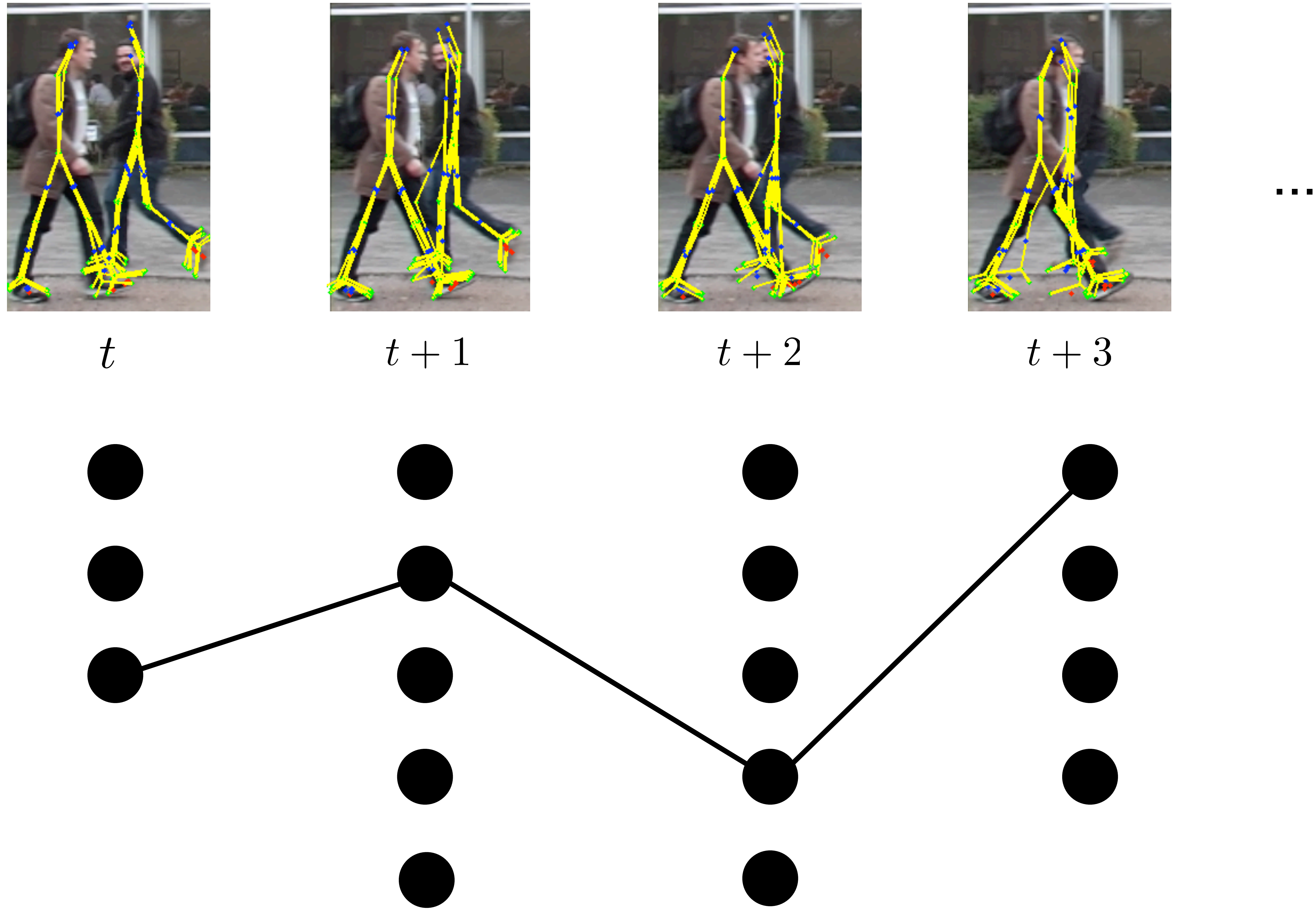


t $t+1$ $t+2$ $t+3$ ...



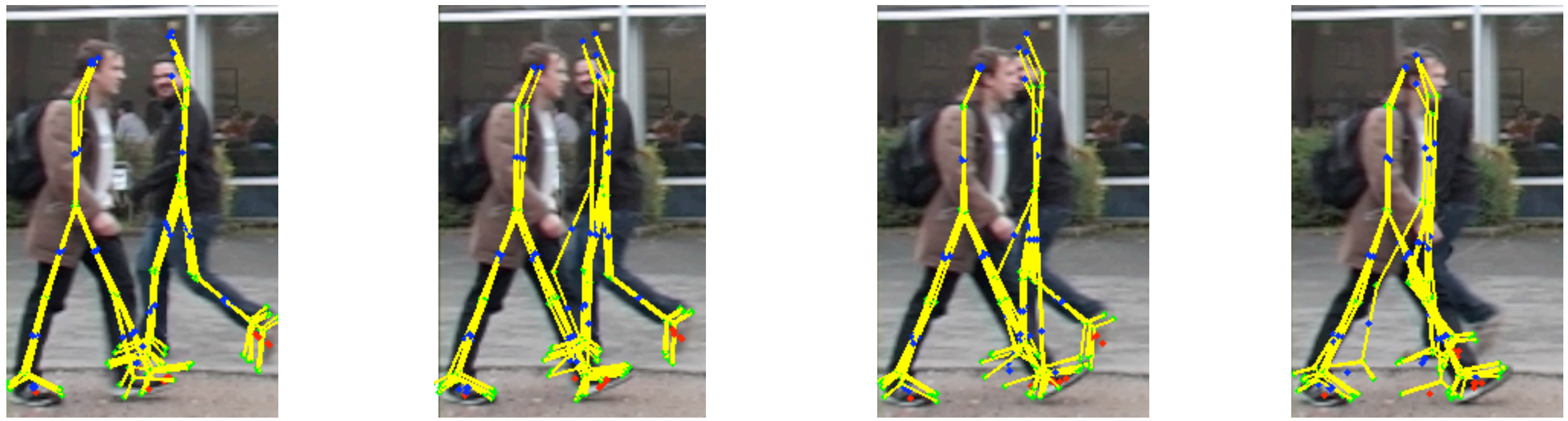
Viterbi Decoding

Tracks from Overlapping Tracklets



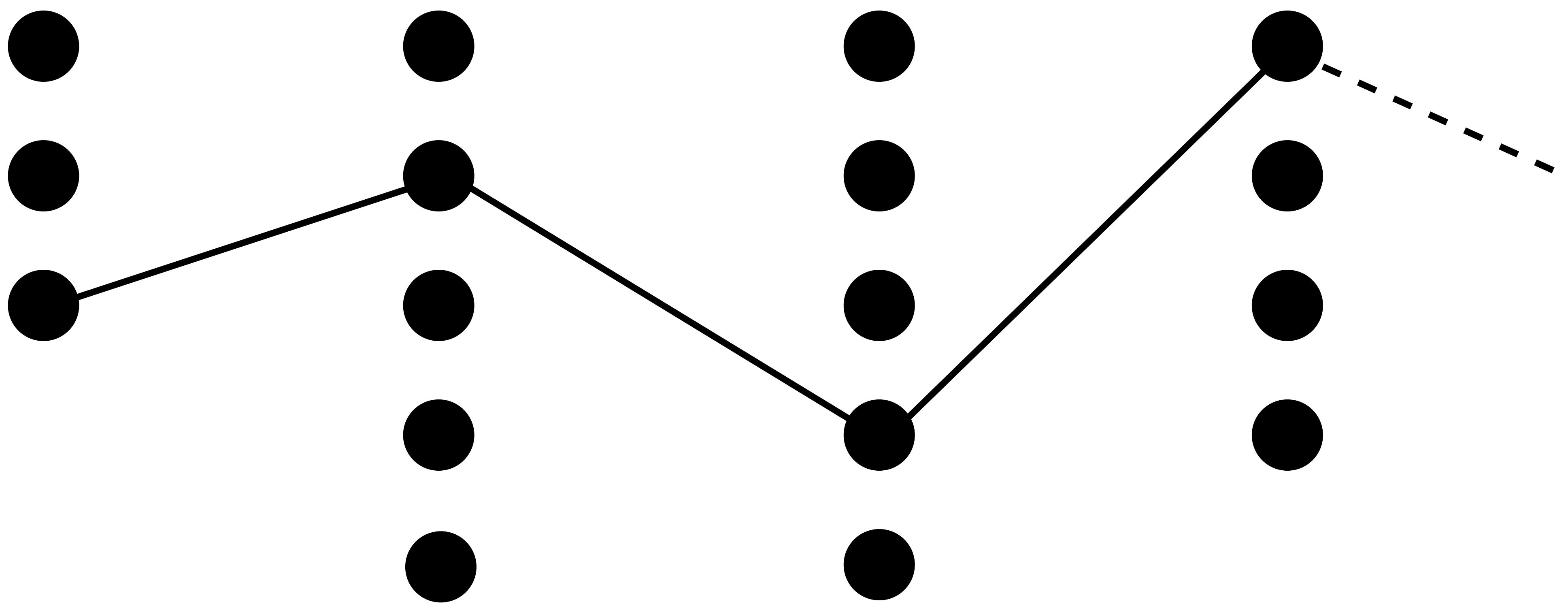
Viterbi Decoding

Tracks from Overlapping Tracklets



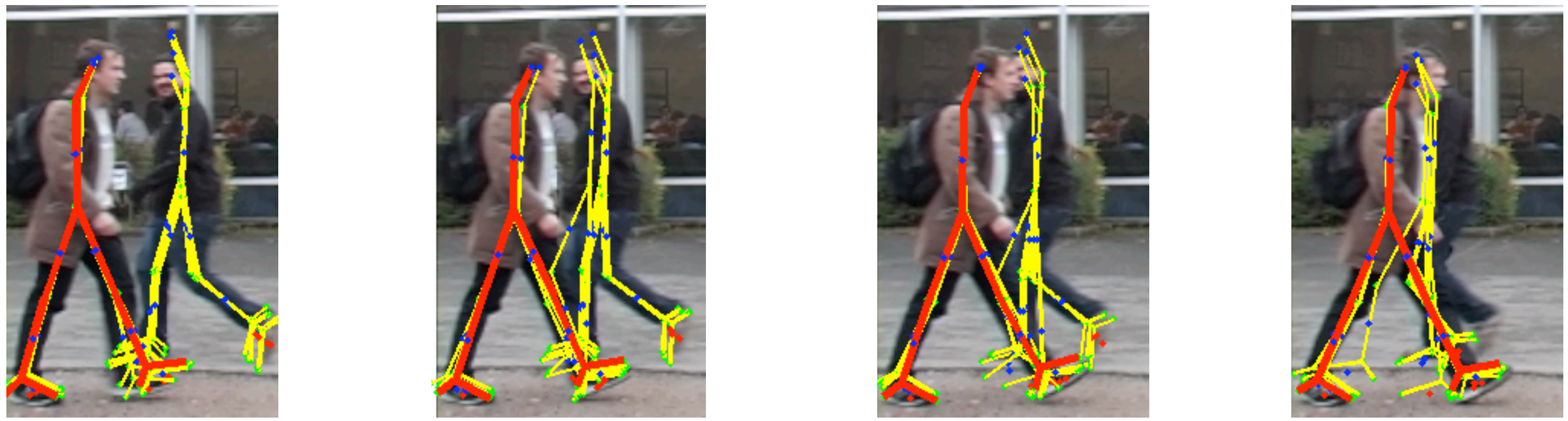
t $t+1$ $t+2$ $t+3$ ...

Viterbi Decoding



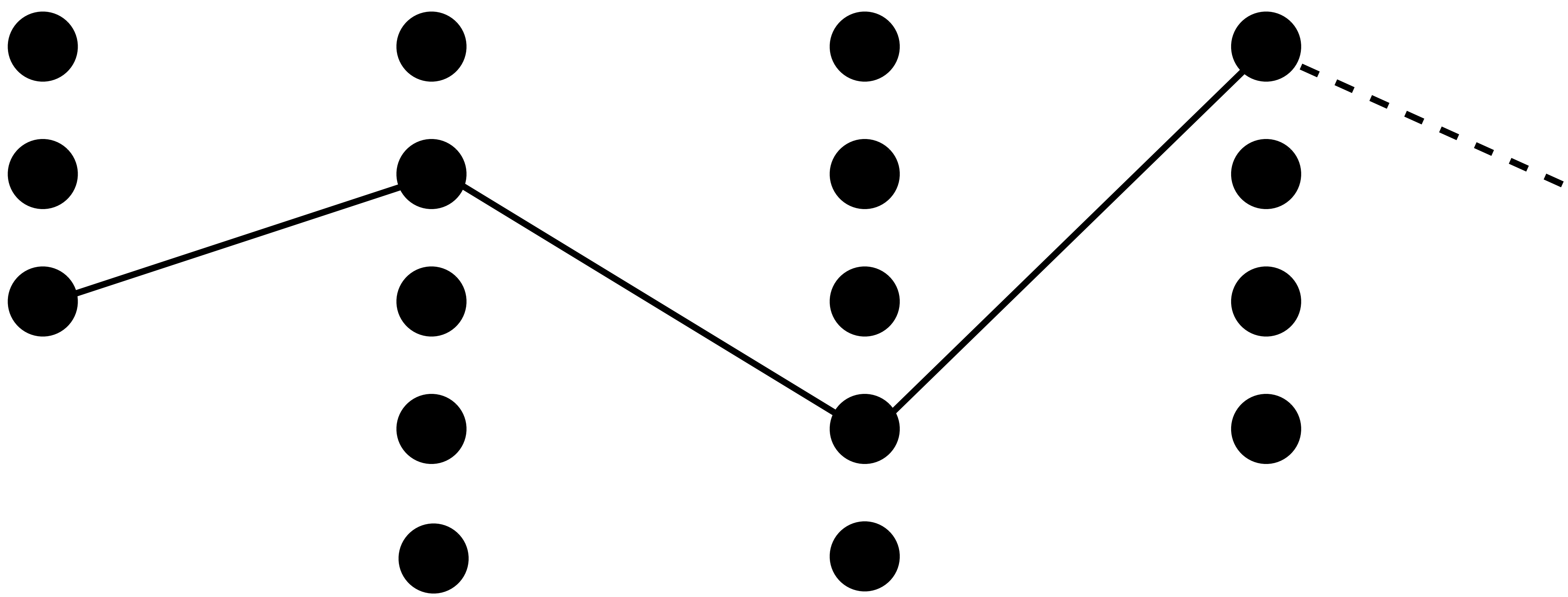
Tracks from Overlapping Tracklets

TECHNISCHE UNIVERSITÄT DARMSTADT

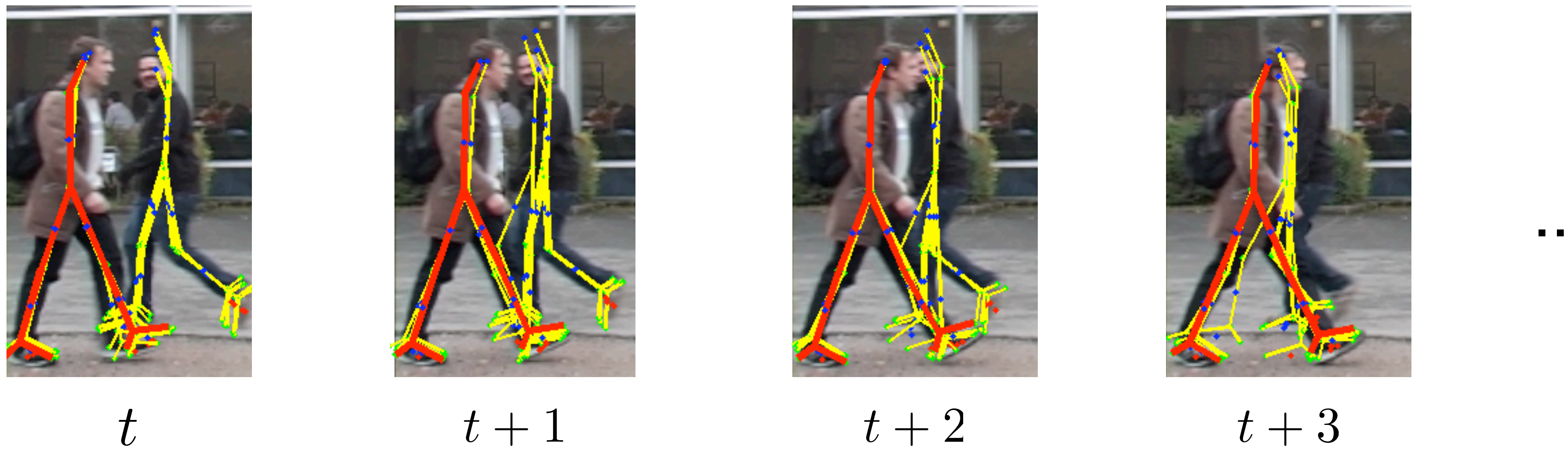


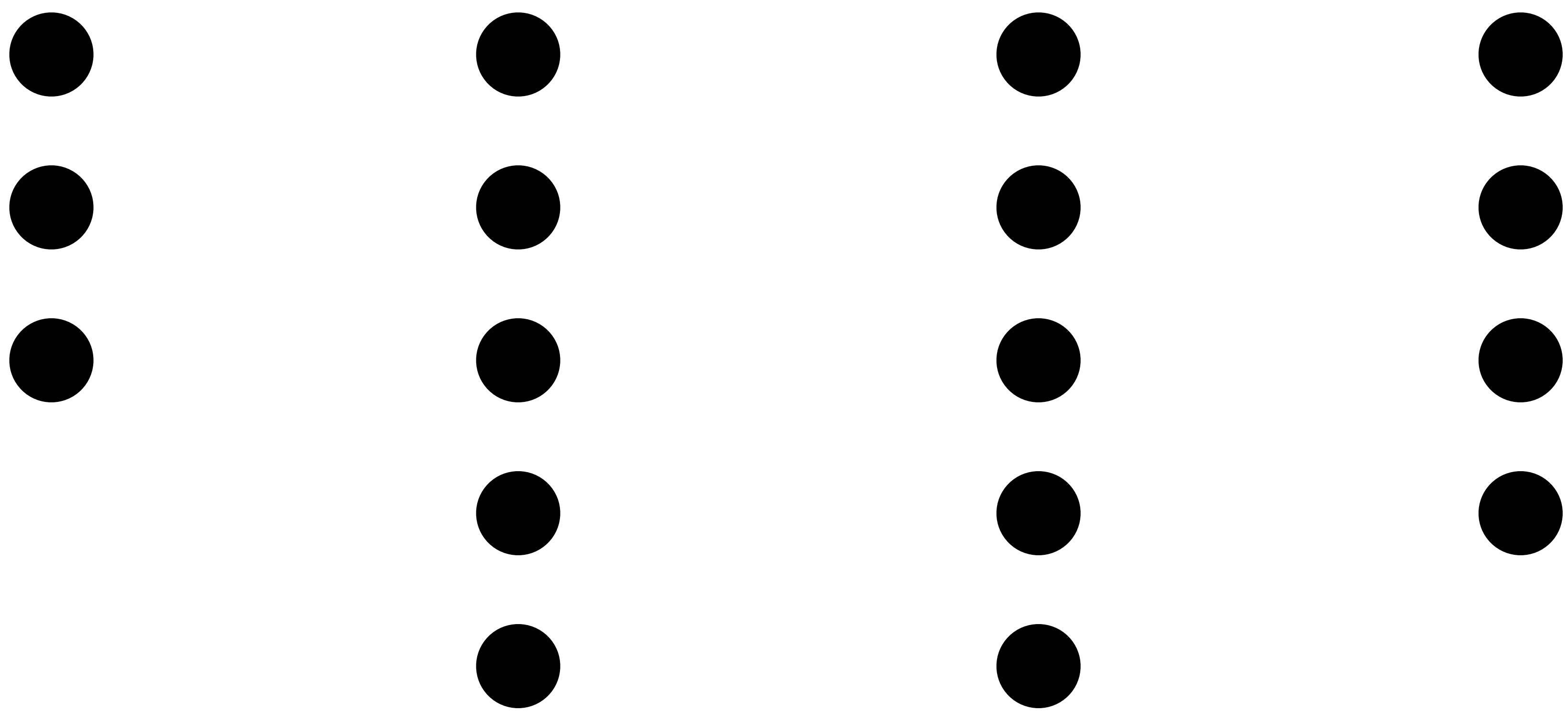
t $t+1$ $t+2$ $t+3$ ...

Viterbi Decoding



Finding Multiple Tracks



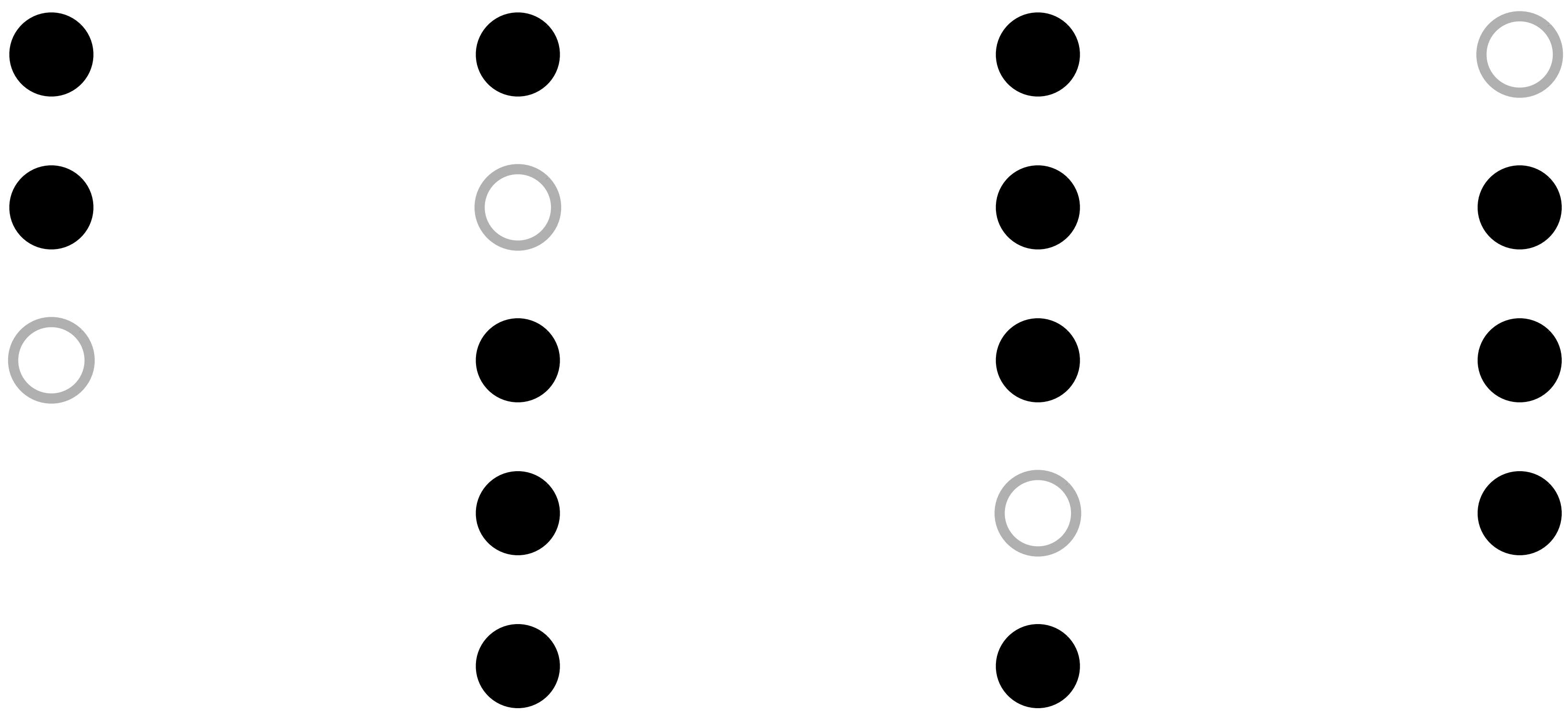


- Find the best track
- Remove its hypotheses
- Repeat

Finding Multiple Tracks

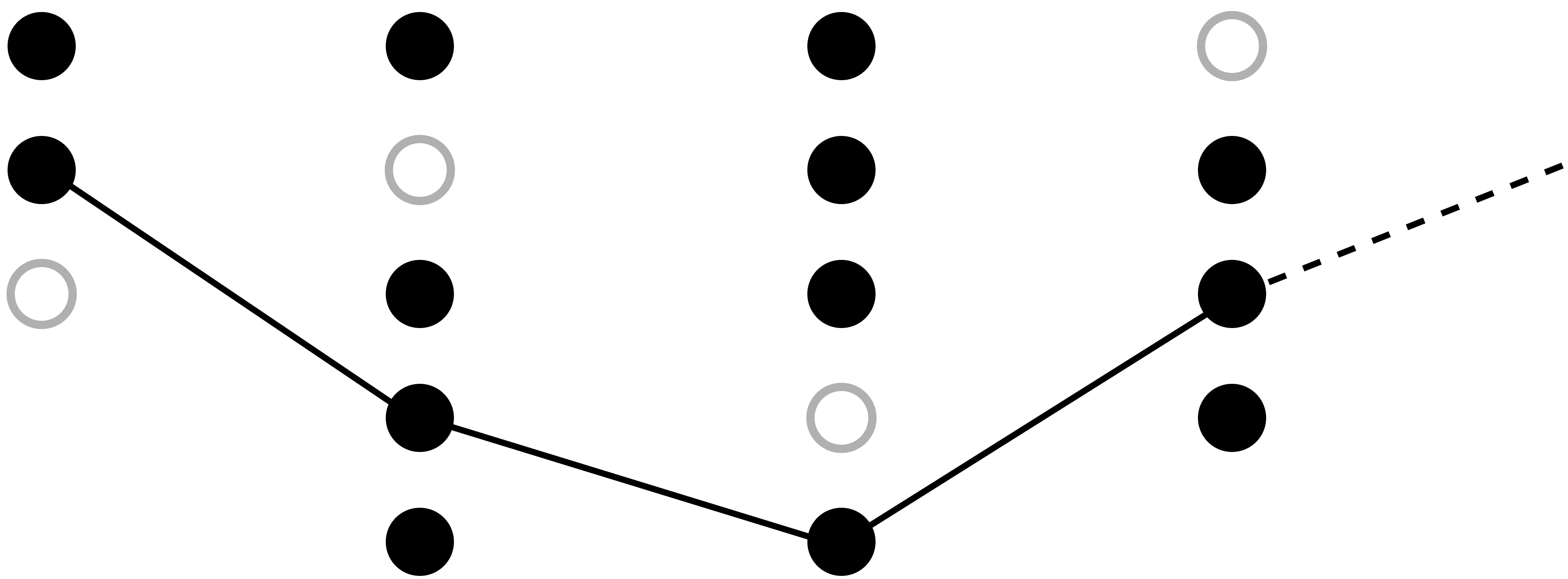
t $t+1$ $t+2$ $t+3$ ...

- Find the best track
- Remove its hypotheses
- Repeat



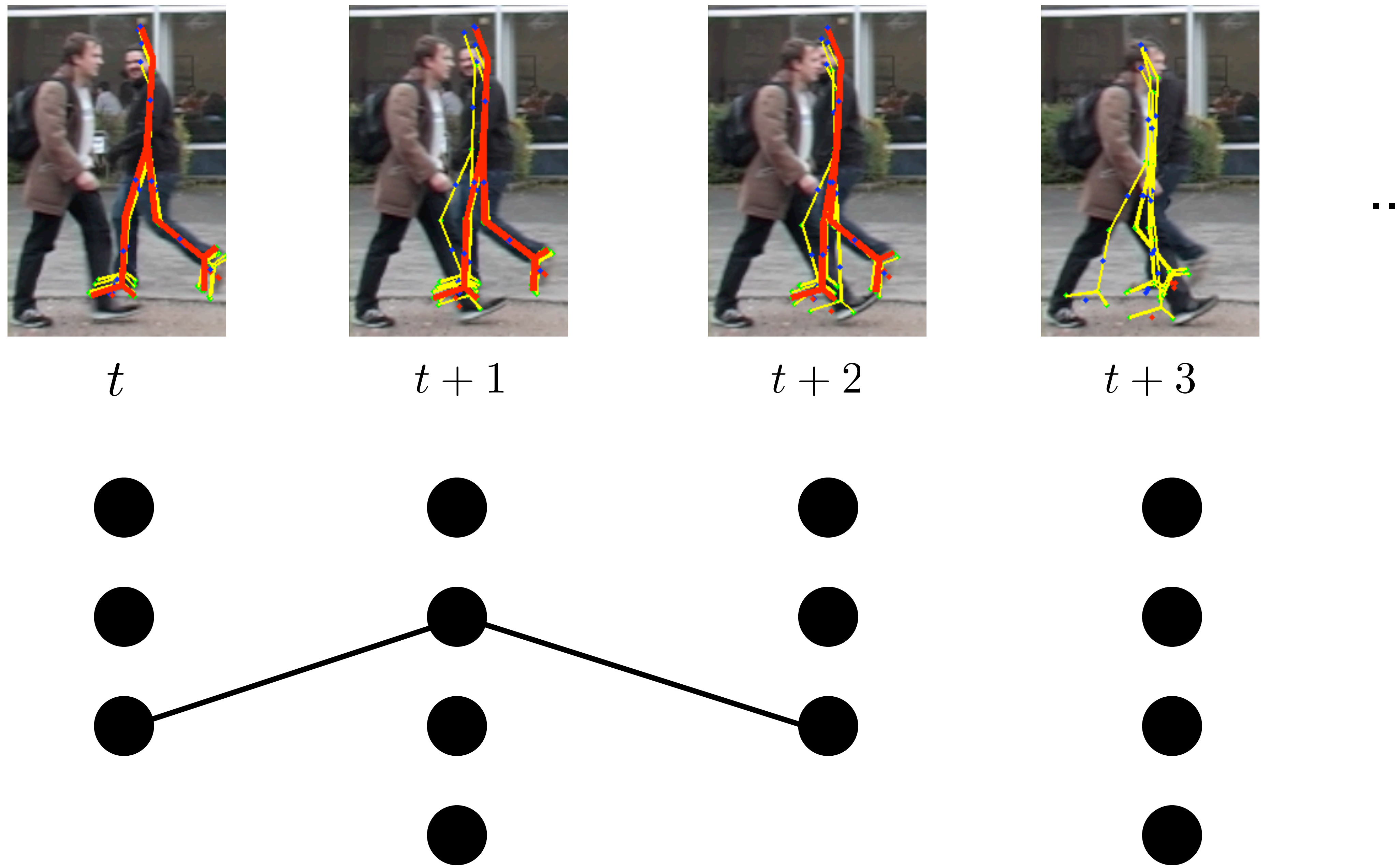
Finding Multiple Tracks

t $t+1$ $t+2$ $t+3$ ...



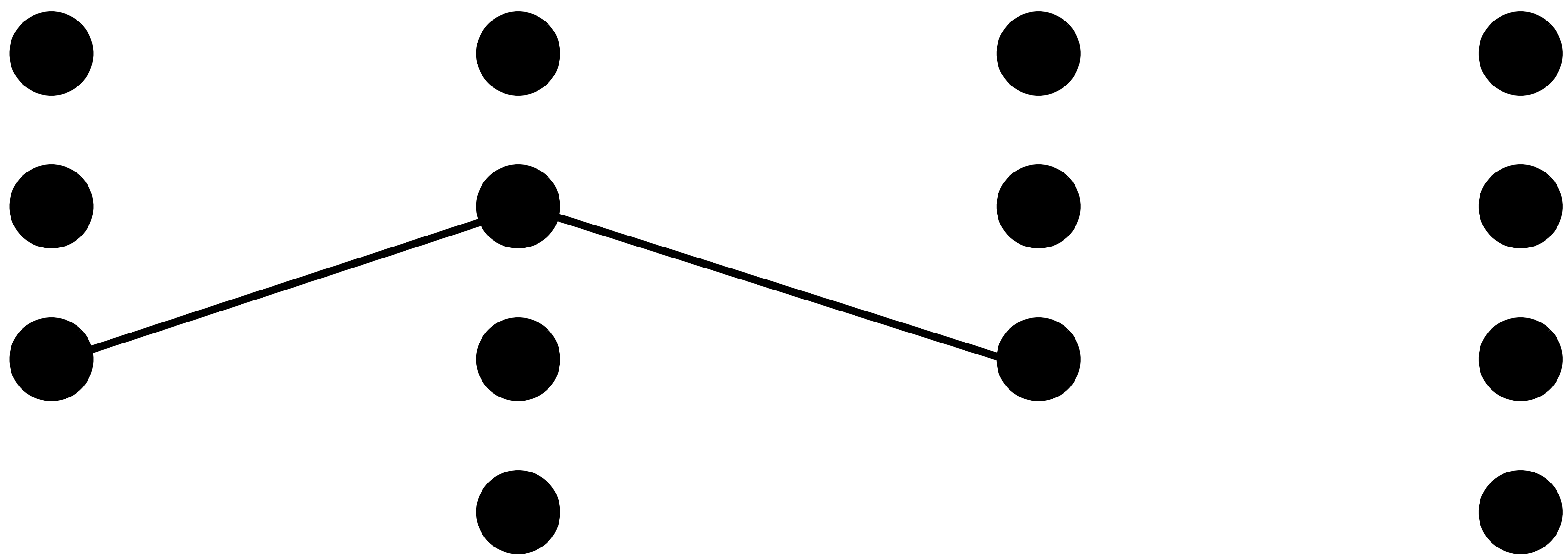
- Find the best track
- Remove its hypotheses
- Repeat

Occlusion Event



Occlusion Event

t $t+1$ $t+2$ $t+3$ …

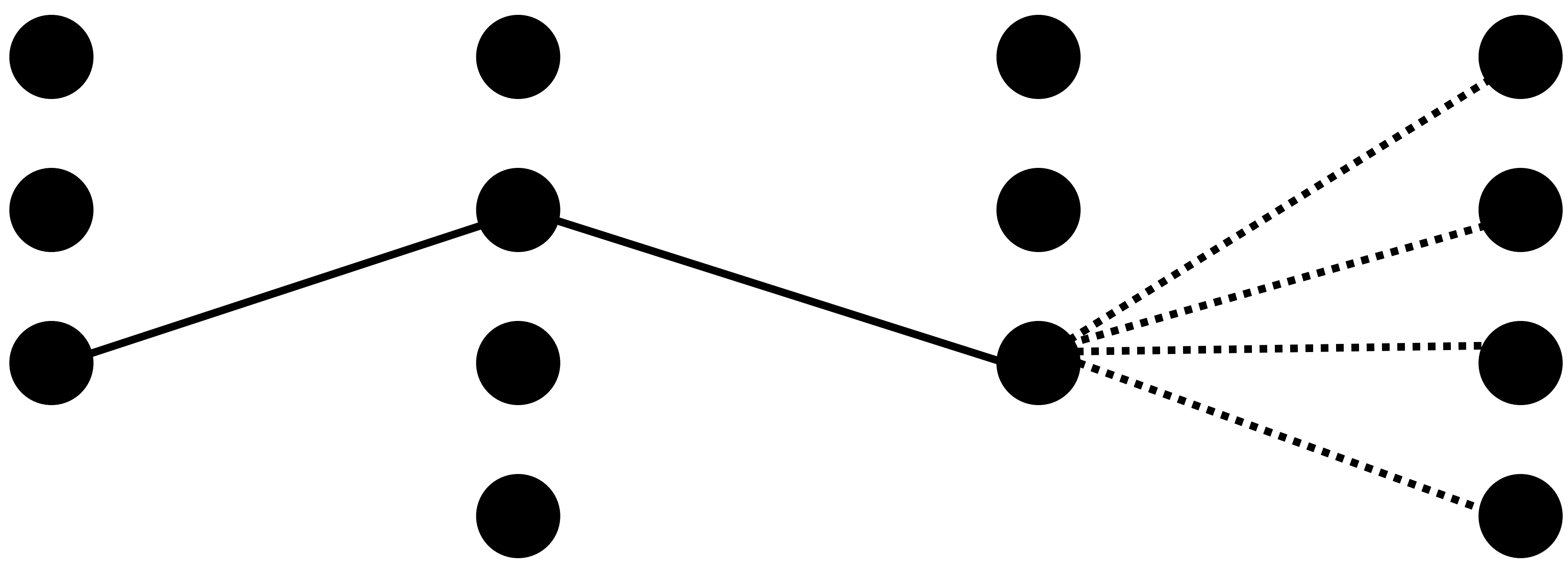


"bad" detections

Occlusion Event

TECHNISCHE UNIVERSITÄT DARMSTADT

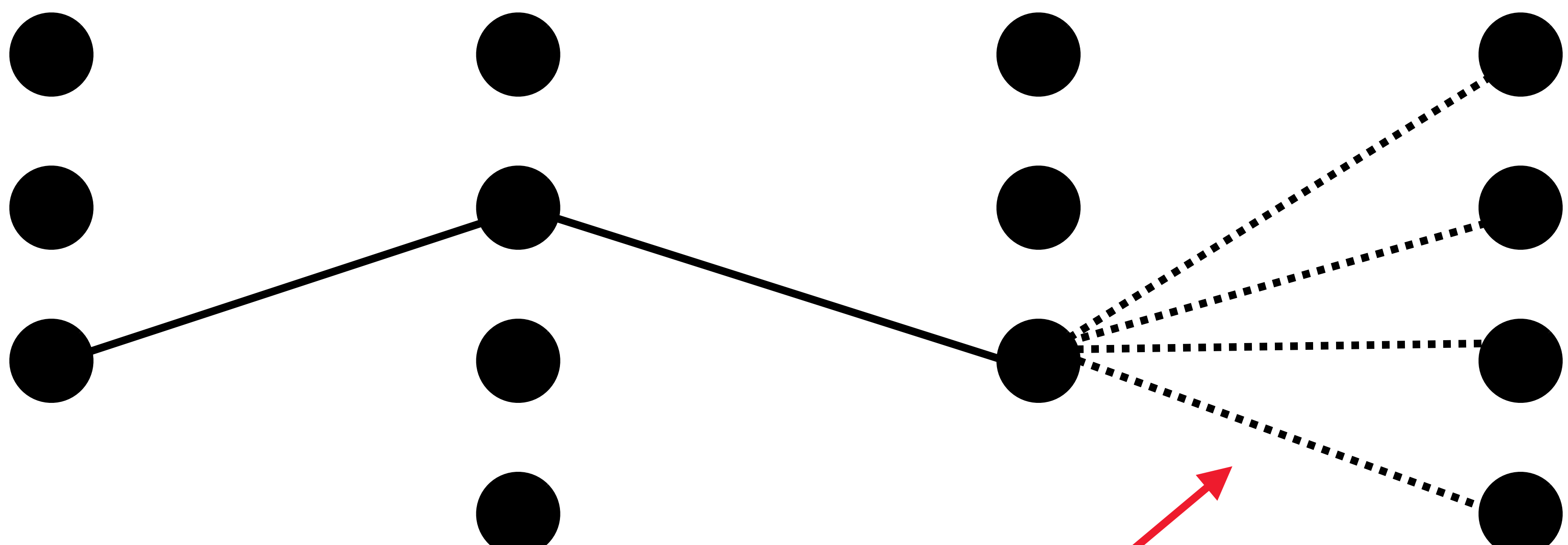
t $t+1$ $t+2$ $t+3$



“bad” detections

Occlusion Event

t $\quad t+1 \quad$ $t+2$ $\quad t+3 \quad$ ...



Appearance Model for Occlusion Recovery

- Extract person-specific appearance model for each limb:
  - Color histogram.
- Require relatively accurate pose estimate:
  - Pose from extracted tracks.
- Appearance comparison measure:
  - Bhattacharyya distance.

Appearance Model for Occlusion Recovery

- Extract person-specific appearance model for each limb:
  - Color histogram.
- Require relatively accurate pose estimate:
  - Pose from extracted tracks.
- Appearance comparison measure:
  - Bhattacharyya distance.

Appearance Model for Occlusion Recovery

- Extract person-specific appearance model for each limb:
  - Color histogram.

- Require relatively accurate pose estimate:
  - Pose from extracted tracks.

Appearance Model for Occlusion Recovery

TECHNISCHE UNIVERSITÄT DARMSTADT

- Extract person-specific appearance model for each limb:
  - Color histogram.
- Require relatively accurate pose estimate:
  - Pose from extracted tracks.
- Appearance comparison measure:
  - Bhattacharyya distance.

Occlusion Recovery

time →

- Greedily link partial tracks based on:
  - Motion & articulation compatibility.
  - Plus appearance compatibility.

Occlusion Recovery

time →

- Greedily link partial tracks based on:
  - Motion & articulation compatibility.
  - Plus appearance compatibility.

Occlusion Recovery

- Greedily link partial tracks based on:
  - Motion & articulation compatibility.
  - Plus appearance compatibility.

Occlusion Recovery

TECHNISCHE UNIVERSITÄT DARMSTADT

- Greedily link partial tracks based on:
  - Motion & articulation compatibility.
  - Plus appearance compatibility.

Occlusion Recovery

- Greedily link partial tracks based on:
  - Motion & articulation compatibility.
  - Plus appearance compatibility.

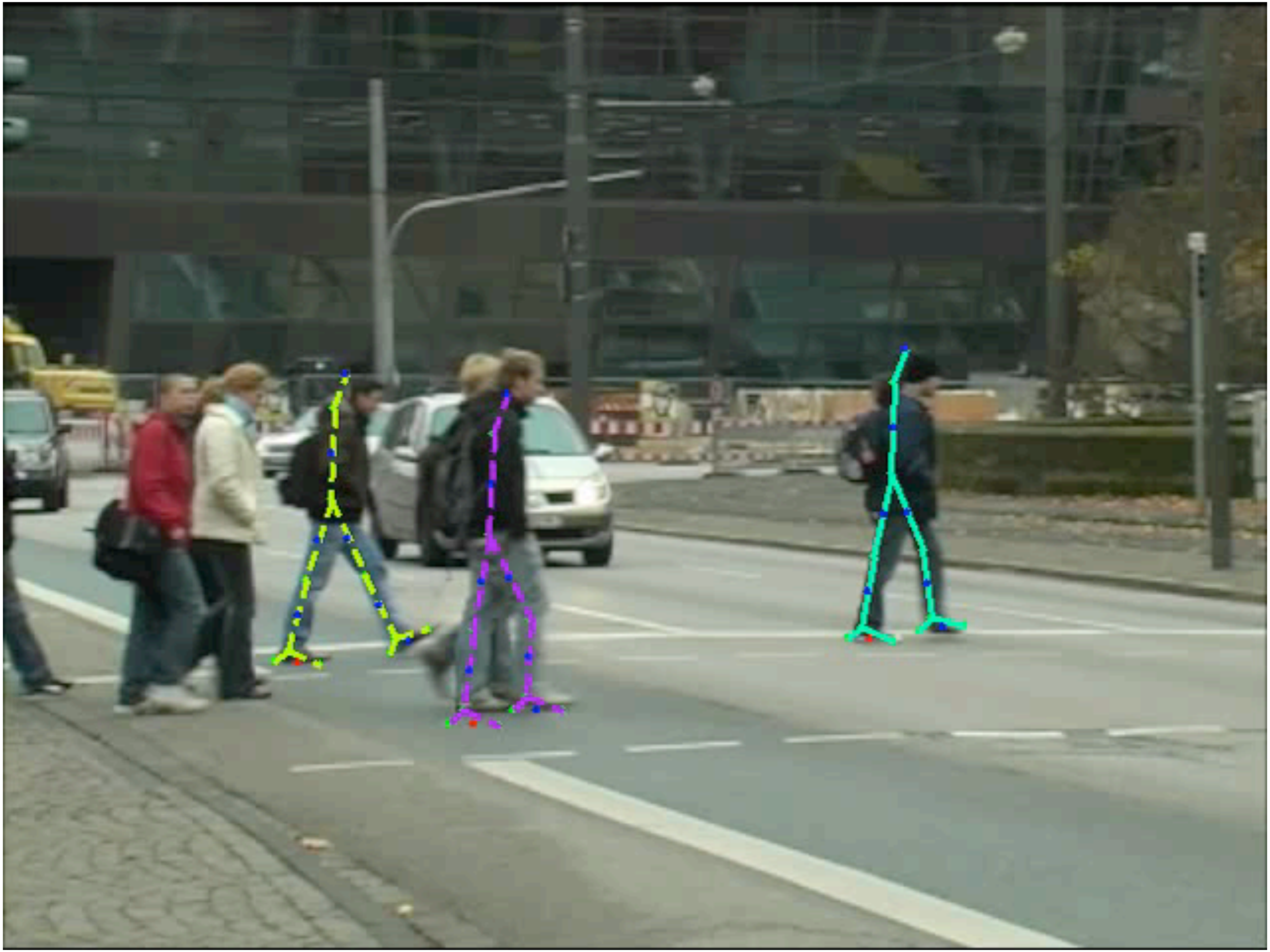
Occlusion Recovery

time →

- Greedily link partial tracks based on:
  - Motion & articulation compatibility.
  - Plus appearance compatibility.

Occlusion Recovery

- Greedily link partial tracks based on:
  - Motion & articulation compatibility.
  - Plus appearance compatibility.



Summary

TECHNISCHE UNIVERSITÄT DARMSTADT

- partISM: Extended the ISM detection framework to part-based detection:
  - Improved detection
  - Basis for incorporating body dynamics.
- Incorporated temporal continuity in a “tracklet” detection framework:
  - hGPLVM dynamics model.
  - Improves occlusion robustness.
  - Reduces false positives.
- Extracted and combined tracks across occlusion events:
  - Person identification throughout entire sequences.

Thanks!

TECHNISCHE UNIVERSITÄT DARMSTADT

- Acknowledgements:
  - Neil Lawrence for his GPLVM code.
  - Mario Fritz for helpful discussions.
  - Partial funding through DFG GRK “Cooperative, Adaptive and Responsive Monitoring in Mixed Mode Environments”
  - Travel funding from DFG.

- Data available at:

  http://www.mis.informatik.tu-darmstadt.de/