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

[Images of people with bounding boxes]
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

• **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

- **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

- **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]
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 \hspace{1cm} 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

- \( L = \{x^0, x^1, \ldots, x^8\} \) - 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**.

\[
p(x^i | x^o)
\]

Covariance and mean part positions for \( p(x^i | x^o) \).
Single Frame Detection

- Detections at equal error rate:

HOG

4D-ISM

partISM
Single-frame Detection Results

- 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).

TUD pedestrians data
No occlusions
Overview

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: \( E = [E_1, \ldots, E_m] \)

- Want:

- Posterior over positions and configurations:
Tracklet Detection in Short Subsequences

• Given: \( E = [E_1, \ldots, E_m] \)

• Want:

\[
X^{o*} = [x_1^{o*}, \ldots, x_m^{o*}]
\]
body positions

• Posterior over positions and configurations:
Tracklet Detection in Short Subsequences

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

- Want: \( X^{o*} = [x^{o*}_1, \ldots, x^{o*}_m] \)
  body positions

\[ Y^* = [y^*_1, \ldots, y^*_m] \]
body configurations

- Posterior over positions and configurations:
Tracklet Detection in Short Subsequences

- Given: \( E = [E_1, \ldots, E_m] \)
- Want:

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

body positions

\[
Y^* = [y^*_1, \ldots, y^*_m]
\]

body configurations

- Posterior over positions and configurations:

\[
p(X^{o*}, Y^* | E) \propto p(E | X^{o*}, Y^*) p(X^{o*}) p(Y^*).
\]
Tracklet Detection in Short Subsequences

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

- Want:

  \( X^{o*} = [x_1^{o*}, \ldots, x_m^{o*}] \)  
  body positions

  \( Y^{*} = [y_1^{*}, \ldots, y_m^{*}] \)  
  body configurations

- Posterior over positions and configurations:

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

  Likelihood model  
  (partISM)
Tracklet Detection in Short Subsequences

- **Given:** \( E = [E_1, \ldots, E_m] \)
- **Want:**

\[ X^{o*} = [x_1^{o*}, \ldots, x_m^{o*}] \]

body positions

\[ Y^{*} = [y_1^{*}, \ldots, y_m^{*}] \]

body configurations

- **Posterior over positions and configurations:**

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

Likelihood model (partISM)  
speed prior (Gaussian)
Tracklet Detection in Short Subsequences

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

- Want:

  $X^{o*} = [x_1^{o*}, \ldots, x_m^{o*}]$
  body positions

  $Y^{*} = [y_1^{*}, \ldots, y_m^{*}]$
  body configurations

- Posterior over positions and configurations:

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

  Likelihood model (partISM)
  speed prior (Gaussian)
  dynamical body model (hGPLVM)
Modeling Body Dynamics

- $Y^*$ 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

- $Y^*$ 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

\[
Y = [y_i \in \mathbb{R}^D]
\]
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]

\[ \mathbf{Y}^* \in \mathbb{R}^{D \times m} \]

Latent space \( \mathbf{Z} = [z_i \in \mathbb{R}^q] \)

Configuration \( \mathbf{Y} = [y_i \in \mathbb{R}^D] \)
Modeling Body Dynamics

- $Y^*$ 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]

$$Y^* = \{y_i \in \mathbb{R}^D\}$$

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

$$Z = [z_i \in \mathbb{R}^q]$$

$$Z = \mathbb{Z}$$

$$Y_i = \mathbb{Y}$$

Time (frame #) - Latent space - Configuration

- $Y$ is a configuration of $y_i$.
- $Z$ is the latent space of $Z$.
- $T$ is the time (frame #).

**Diagram:**
- The diagram illustrates the relationship between the configuration $Y$, the latent space $Z$, and the time $T$.
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]

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

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

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

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

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

- Tracklets are local maxima of:

\[ p(X^o*, Y^* | E) \propto p(E | X^o*, Y^*)p(X^o*)p(Y^*). \]

- 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

propagate detection

\[ \text{Tracklet Detection} \]
Tracklet Detection

propagate detection

hGPLVM mean prediction
Tracklet Detection

propagate detection

hGPLVM mean prediction
Tracklet Detection

propagate detection

hGPLVM mean prediction
Tracklet Detection

propagate detection

hGPLVM mean prediction
Tracklet Detection

propagate detection

hGPLVM mean prediction

pose optimization
Single-Frame Detector vs. Tracklet Detector

- At equal error rate:
  - Fewer false positives.
  - More robust detection of partially occluded people.
Single-Frame Detector vs. Tracklet Detector

- At equal error rate:
  - Fewer false positives.
  - More robust detection of partially occluded people.
Single-Frame Detector vs. Tracklet Detector

- At equal error rate:
  - Fewer false positives.
  - More robust detection of partially occluded people.
Single-Frame Detector vs. Tracklet Detector

- At equal error rate:
  - Fewer false positives.
  - More robust detection of partially occluded people.
Single-Frame Detector vs. Tracklet Detector

- At equal error rate:
  - Fewer false positives.
  - More robust detection of partially occluded people.
Detection Performance

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

TUD campus data
With occlusions (up to 50%)
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

Candidate poses from all overlapping tracklets
Tracks from Overlapping Tracklets

Candidate poses from all overlapping tracklets
Tracks from Overlapping Tracklets

Candidate poses from all overlapping tracklets
Tracks from Overlapping Tracklets

Candidate poses from all overlapping tracklets
Tracks from Overlapping Tracklets

Candidate poses from all overlapping tracklets
Tracks from Overlapping Tracklets

Candidate poses from all overlapping tracklets
Tracks from Overlapping Tracklets

$t$

$t + 1$

$t + 2$

$t + 3$

...
Tracks from Overlapping Tracklets

Viterbi Decoding
Tracks from Overlapping Tracklets

Viterbi Decoding
Tracks from Overlapping Tracklets

Viterbi Decoding
Tracks from Overlapping Tracklets

Viterbi Decoding

$V_t$ $V_{t+1}$ $V_{t+2}$ $V_{t+3}$
Tracks from Overlapping Tracklets

Viterbi Decoding
Tracks from Overlapping Tracklets

Viterbi Decoding

\[ t \quad t+1 \quad t+2 \quad t+3 \]
Finding Multiple Tracks

- Find the best track
- Remove its hypotheses
- Repeat

\[
\begin{align*}
\text{at} & \quad \text{at } t+1 & \quad \text{at } t+2 & \quad \text{at } t+3 \\
\bullet & \quad \bullet & \quad \bullet & \quad \bullet \\
\bullet & \quad \bullet & \quad \bullet & \quad \bullet \\
\bullet & \quad \bullet & \quad \bullet & \quad \bullet \\
\bullet & \quad \bullet & \quad \bullet & \quad \bullet \\
\bullet & \quad \bullet & \quad \bullet & \quad \bullet \\
\bullet & \quad \bullet & \quad \bullet & \quad \bullet \\
\bullet & \quad \bullet & \quad \bullet & \quad \bullet
\end{align*}
\]
Finding Multiple Tracks

- Find the best track
- Remove its hypotheses
- Repeat

• Find the best track
• Remove its hypotheses
• Repeat
Finding Multiple Tracks

- Find the best track
- Remove its hypotheses
- Repeat

\[ t \]
\[ t + 1 \]
\[ t + 2 \]
\[ t + 3 \]
Occlusion Event

$t$

$t + 1$

$t + 2$

$t + 3$

...
Occlusion Event

\[ t \quad t + 1 \quad t + 2 \quad t + 3 \]

“bad” detections
Occlusion Event

$t$

$t + 1$

$t + 2$

$t + 3$

“bad” detections
Occlusion Event

"bad" detections

terminate if low-probability for any transition
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

- 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

- 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

- 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

- 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

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

• **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!

- **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:**
  