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

Mykhaylo Andriluka, Stefan Roth, Bernt Schiele

Computer Science Department, TU Darmstadt, Germany
{andriluka, sroth, schiele}@cs.tu-darmstadt.de

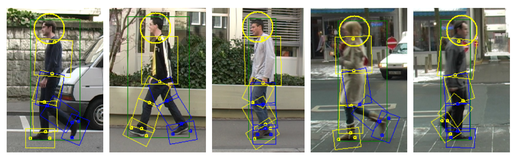
Objective

We consider a problem of *detection* and *tracking* of people in image sequences. The proposed approach is designed to handle long term occlusions which frequently occur in crowded street scenes.

Main steps of our tracking method are:

- People detection combined with estimation of position of body limbs.
- Reconstruction of poses from detections in several subsequence frames, guided by the learned model of the walking motions.
- Long term association of partial tracks based on individualized appearance model and coarse motion model.

Part-based Model for People Detection



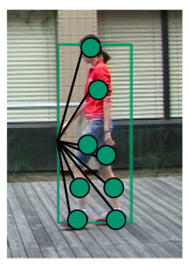
Our person detector is based on combination of recent ideas in object detection:

- *Part representation* is used to cope with high complexity of articulation space.
- Appearance of each part is modeled using codebook local features.
- Correlations between positions of different parts are modelled with additional articulation variable.

- Pictorial structures model [Felzenszwalb & Huttenlocher, IJCV 2005]

$$p(L|E) = \sum_a p(L|a,E)p(a)$$

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



  - Usual approximation: Likelihood can be written as product of part likelihoods

$$p(E|L,a) \approx \prod_{i=1}^{N} p(E|\mathbf{x}^i,a) \propto \prod_{i=1}^{N} p(\mathbf{x}^i|E,a)$$

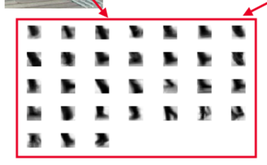
assuming uniform prior over part locations

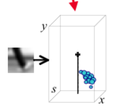
$$p(\mathbf{x}^i|a,E) \approx c_0 + c_1 \sum_{\mathbf{e}_k} p(\mathbf{x}^i|a,\mathbf{e}_k)$$

- To model the part posterior w.r.t. a single feature we introduce a codebook (just as in the ISM):

$$p(\mathbf{x}^i|a,\mathbf{e}_k) = \sum_{\mathbf{c}_j} p(\mathbf{x}^i|a,\mathbf{c}_j,\mathbf{e}_k)p(\mathbf{c}_j|a,\mathbf{e}_k)$$

$$= \sum_{\mathbf{c}_j} p(\mathbf{x}^i|a,\mathbf{c}_j,\mathbf{e}_k^{pos})p(\mathbf{c}_j|\mathbf{e}_k^{app}).$$





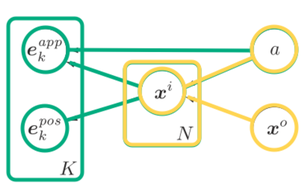
Appearance similarity (Gaussian similarity measure)

Spatial occurrence distribution

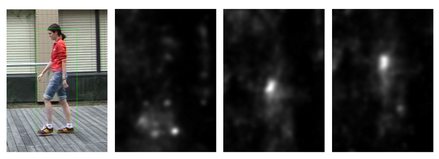
$$p(L|a,E) \approx \prod_i p(\mathbf{x}^i|\mathbf{x}^o,a) \left[\beta + \sum_{\mathbf{e}_k} p(\mathbf{x}^i|a,\mathbf{e}_k)\right]$$

$$p(\mathbf{x}^i|a,\mathbf{e}_k) = \sum_{\mathbf{c}_j} p(\mathbf{x}^i|a,\mathbf{c}_j,\mathbf{e}_k^{pos})p(\mathbf{c}_j|\mathbf{e}_k^{app}).$$

- Graphical model structure:



Reconstruction of Poses in Short Sequences



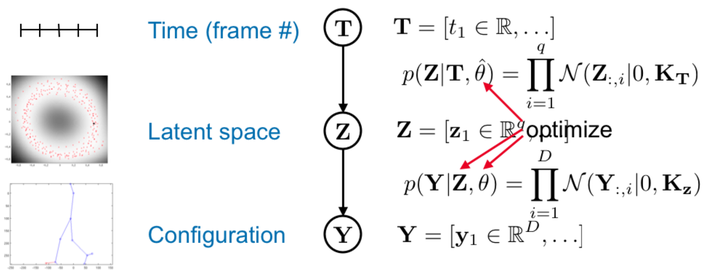
- Sequence of m frames:
  - Given: Image evidence $E = [E_1, \ldots, E_m]$
  - Want: Body positions $\mathbf{X}^{o*} = [\mathbf{x}_1^{o*}, \ldots, \mathbf{x}_m^{o*}]$
  - Want: Body configuration $\mathbf{Y}^* = [\mathbf{y}_1^*, \ldots, \mathbf{y}_m^*]$ relative limb angles
- Posterior over positions and configurations:

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

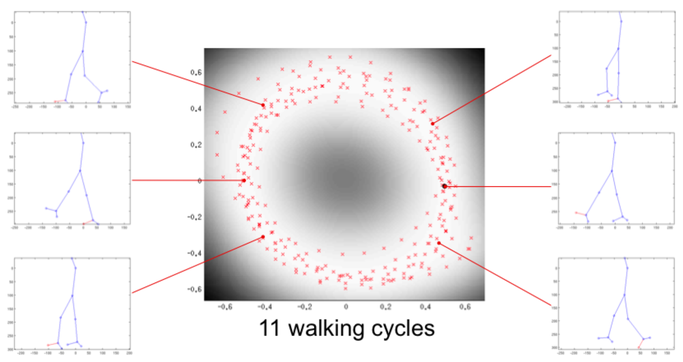
$$\propto p(\mathbf{Y}^*)p(\mathbf{X}^{o*}) \prod_{j=1}^{m} p(E_j|\mathbf{y}_j^*, \mathbf{x}_j^{o*}).$$

dynamical body model (hGPLVM) — simple speed prior (Gaussian) — likelihood model (part ISM)

- Model the body dynamics using a hierarchical Gaussian process latent variable model (hGPLVM) [Lawrence & Moore, ICML 2007]:

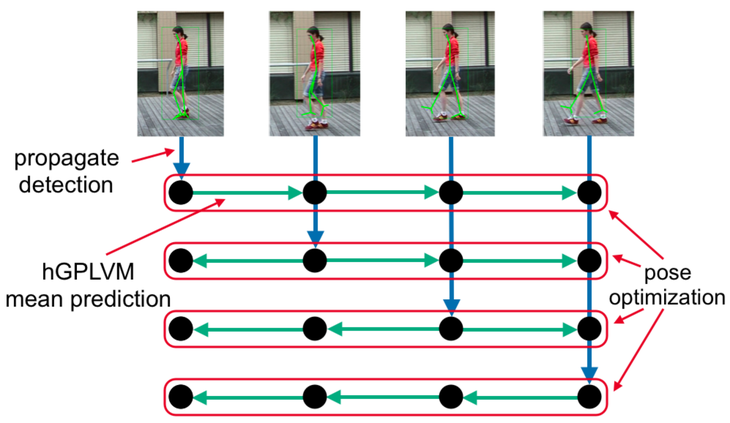


Learning a low-dimensional representation for poses and motions

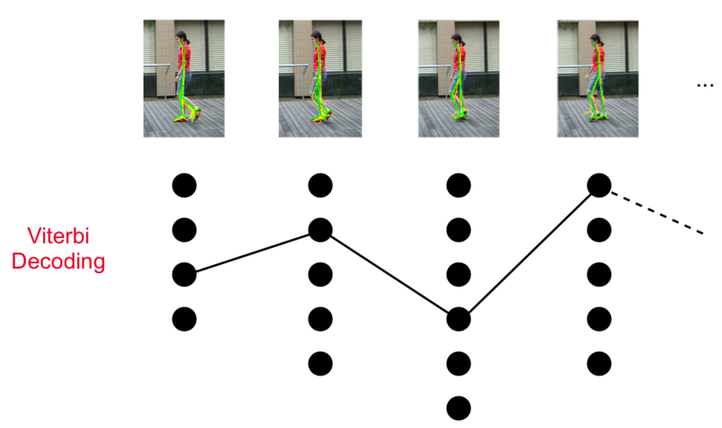


- Similar to GPDM [Urtasun et al, CVPR 2006]
  - But allows for skipped frames (unequal time intervals)
  - Allows for forward **and backward** prediction

Tracklet Detection

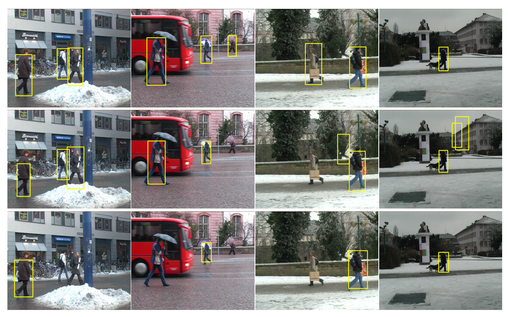


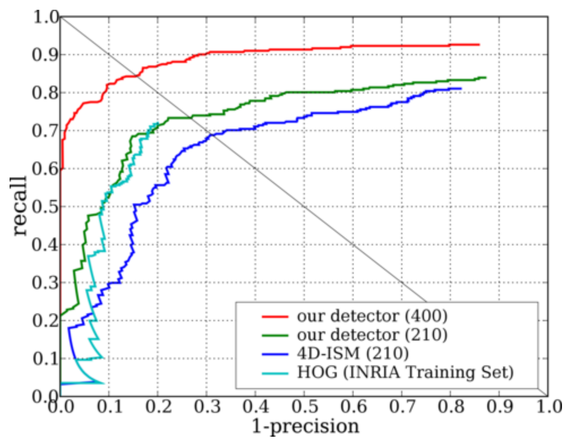
Obtaining Longer Tracks



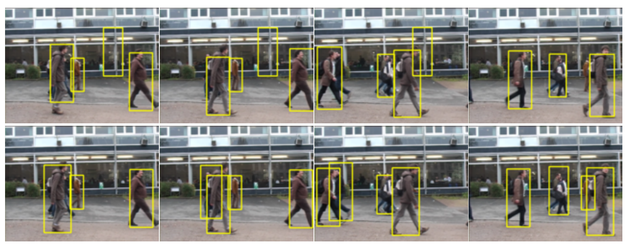
Experimental Results

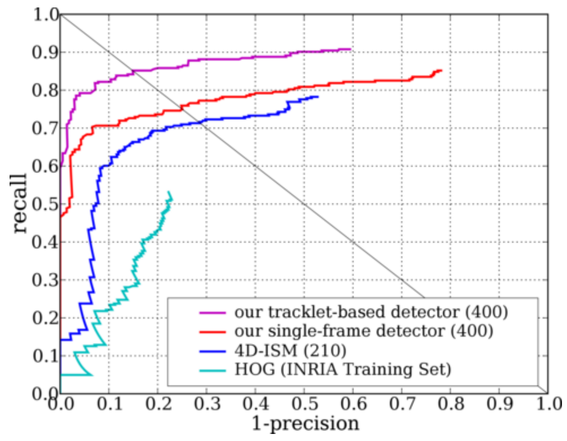
Our part-based person detector compares favorably to other state-of-the-art approaches while additionally being able to compute estimates of limb positions in the image.





Comparison of Tracklet and Single-frame Detectors





Dataset	HOG		4D-ISM		single-frame		tracklets	
TUD-Pedestrians	0.53	-	0.28	0.68	0.81	0.84	-	-
TUD-Campus	0.22	-	0.6	0.71	0.7	0.75	0.82	0.85

References

[1] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *CVPR 2005*.

[2] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. *CVPR 2006*.

[3] N. D. Lawrence and A. J. Moore. Hierarchical Gaussian process latent variable models. *ICML 2007*.

[4] M. Andriluka, S. Roth and B. Schiele. People-Tracking-by-Detection and People-Detection-by-Tracking. *CVPR 2008*.

Acknowledgements: The authors are very thankful to Neil Lawrence for making his GPLVM toolbox available online and to Mario Fritz for helpful suggestions concerning the color model. Mykhaylo Andriluka gratefully acknowledges a scholarship from DFG GRK "Cooperative, Adaptive and Responsive Monitoring in Mixed Mode Environments".