Human Pose Estimation and Segmentation in Videos

Karteek Alahari
Inria Grenoble – Rhône-Alpes

Joint work with Ivan Laptev, Guillaume Seguin, Josef Sivic (WILLOW team)
Anoop Cherian, Julien Mairal, Cordelia Schmid (LEAR team)

informatiques mathématiques Inria

Pose Estimation and Segmentation
of Multiple People in (Stereoscopic) Videos

+

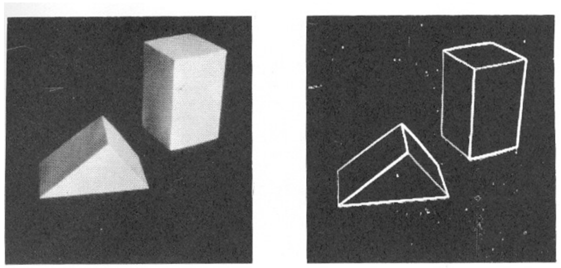
Human Pose Estimation in Videos

Pose Estimation and Segmentation of Multiple People in (Stereoscopic) Videos

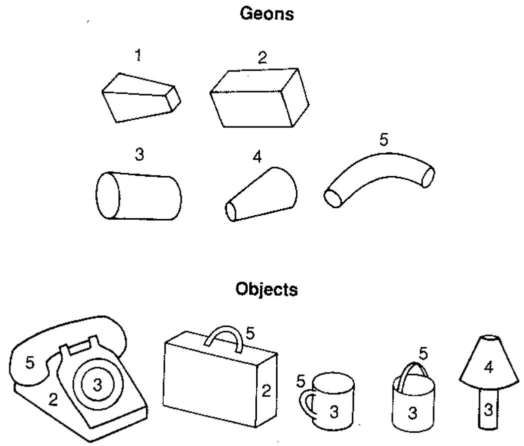
+

Human Pose Estimation in Videos

3D Reasoning



L. G. Roberts, *Machine Perception of Three Dimensional Solids,* Ph.D. Thesis, MIT, 1963



I. Biederman, *Geons*, 1985-87

- … has recently made it out of the lab

Microsoft Kinect, 2010

3D in the Wild : Stereo Movies

- **Inria 3DMovie Dataset**: Annotated stereo pairs
- 440 training stereo pairs, 36 test video sequences
- Labelling: 686 person segmentation, 587 poses, 1158 person bounding boxes

Available at: http://www.di.ens.fr/willow/research/stereoseg

3D in the Wild: The Goal

- Layered segmentation of people in stereoscopic videos

StreetDance 3D (2010)

3D in the Wild: The Goal

StreetDance 3D (2010)

- Pixel-wise segmentation, pose estimation
- Relative depth ordering

Why is this task important?

- A mid-level representation for subsequent recognition tasks

- Annotated data for learning to segment people in monocular videos
  - e.g., 90min movie → 150000 annotated frames

- Interactive annotation/editing tools

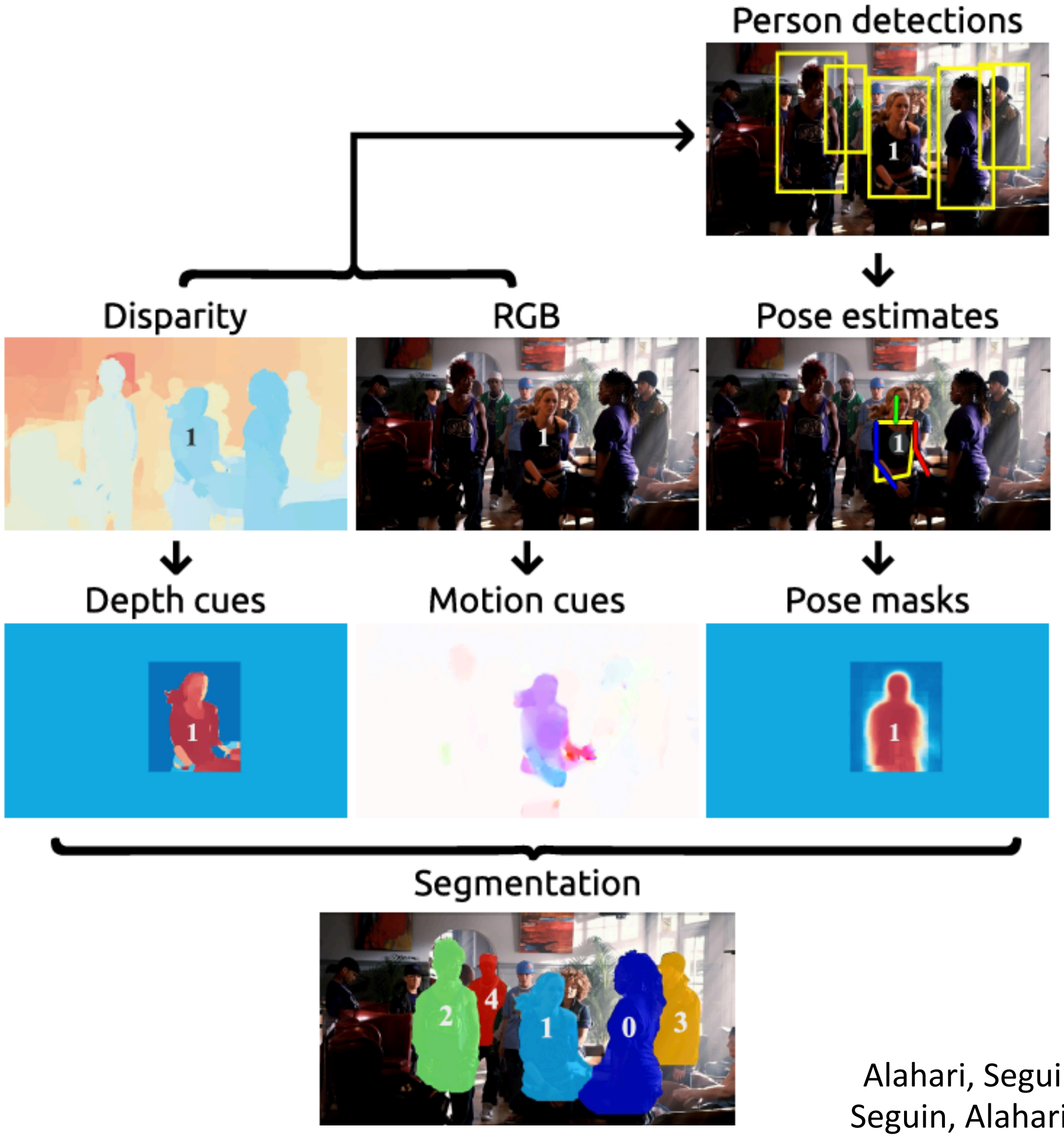
How challenging is it?

- Noisy signal
- Unrestricted indoor/outdoor settings

How challenging is it?

- Noisy signal
- Unrestricted indoor/outdoor settings

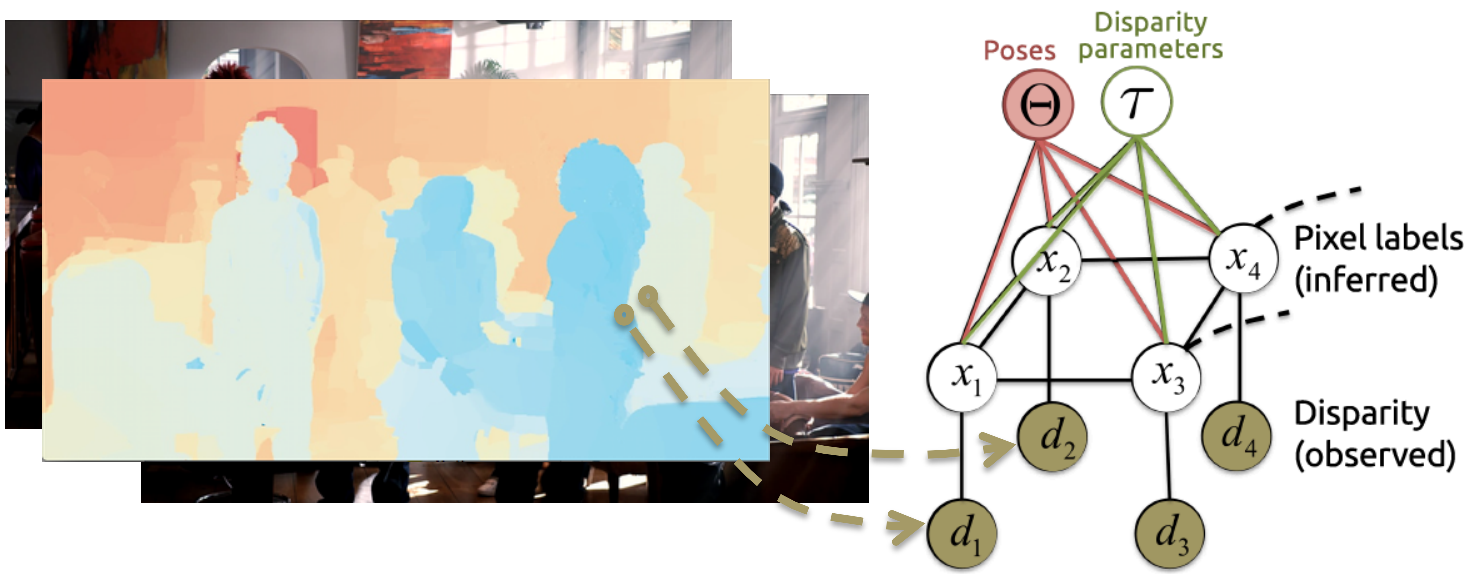
Overview



Alahari, Seguin, Sivic, Laptev, ICCV 2013
Seguin, Alahari, Sivic, Laptev, PAMI 2015

Overview

- Given the disparity (d), estimate
  - Pixel labels (x_i): Denotes the person
  - Poses (Θ): The pose of each person
  - Layers (τ): The layered ordering of people



Overview

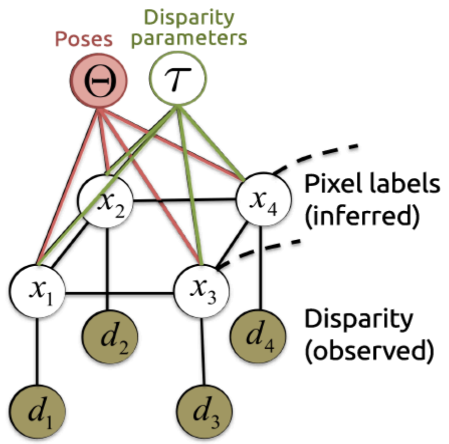
- Define the estimation as:

$$\{\mathbf{x}^*, \mathbf{\Theta}^*, \tau^*\} = \arg\min_{\mathbf{x}, \mathbf{\Theta}, \tau} E(\mathbf{x}, \mathbf{\Theta}, \tau)$$

- NP-hard to solve [Boros and Hammer, 2002]

- Approximate it as:

$$\{\mathbf{x}^*, \tau^*\} = \arg\min_{\mathbf{x}, \tau} E(\mathbf{x}, \tau; \mathbf{\Theta})$$



Overview

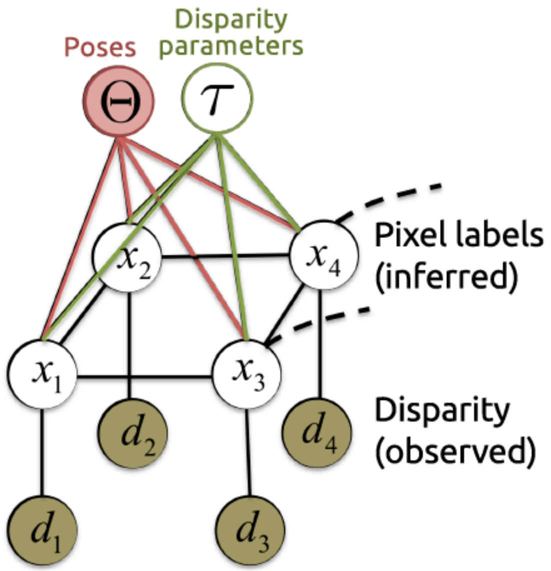
$$\{\mathbf{x}^*, \tau^*\} = \arg\min_{\mathbf{x},\tau} E(\mathbf{x}, \tau; \mathbf{\Theta})$$

- A 2-step approach
  - Estimate disparity parameters $\tau^* = \arg\min_{\{\tau\}} \tilde{E}(\tilde{\mathbf{x}}; \mathbf{\Theta}, \tau)$
  - Minimize $E(\mathbf{x}; \mathbf{\Theta}, \tau^*)$

[Boykov et al. 2001]

Energy function

$$E(\boxed{\mathbf{x}};\boxed{\boldsymbol{\Theta}},\boxed{\tau}) = \boxed{\sum_{i \in \mathcal{V}}}\boxed{\phi_i(x_i;\boldsymbol{\Theta},\tau)} + \boxed{\sum_{(i,j)\in\mathcal{E}}}\boxed{\phi_{ij}(x_i,x_j)} + \boxed{\sum_{(i,k)\in\mathcal{E}^t}}\boxed{\phi^t_{ij}(x_i,x_k)}$$

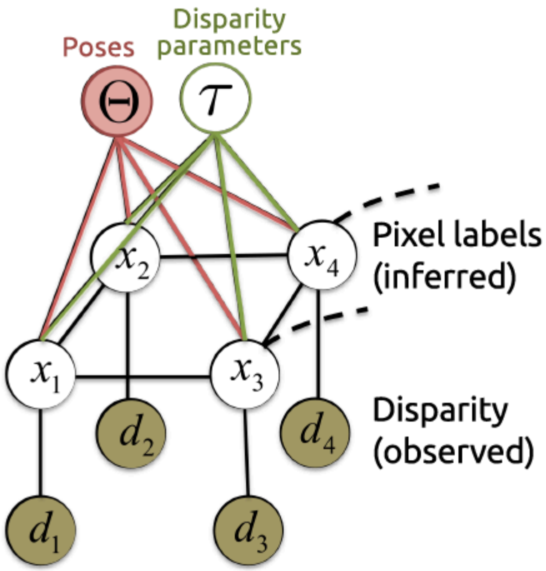


Energy function

sum over temporal edges

$$E(\boxed{\mathbf{x}};\boxed{\boldsymbol{\Theta}},\boxed{\tau}) = \sum_{i\in\mathcal{V}} \phi_i(x_i;\boldsymbol{\Theta},\tau) + \sum_{(i,j)\in\mathcal{E}} \phi_{ij}(x_i,x_j) + \sum_{(i,k)\in\mathcal{E}^t} \phi^t_{ij}(x_i,x_k)$$

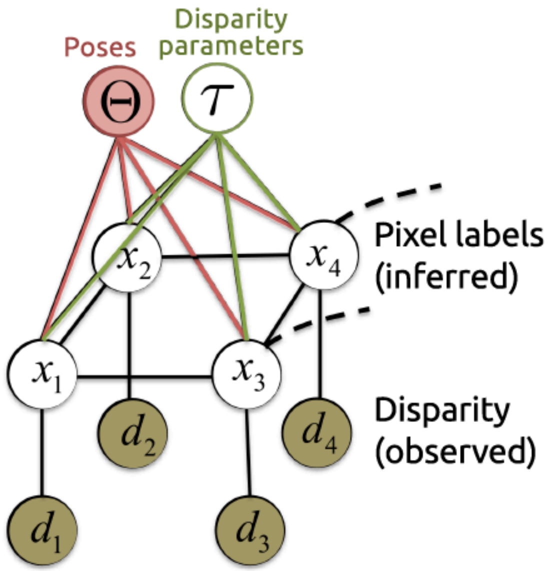
Temporal smoothness:
similar to spatial smoothness



Energy function

sum over spatial edges

$$E(\boxed{\mathbf{x}};\boxed{\boldsymbol{\Theta}},\boxed{\tau}) = \sum_{i\in\mathcal{V}} \phi_i(x_i;\boldsymbol{\Theta},\tau) + \underbrace{\sum_{(i,j)\in\mathcal{E}} \phi_{ij}(x_i,x_j)}_{\text{Spatial smoothness}} + \sum_{(i,k)\in\mathcal{E}^t} \phi^t_{ij}(x_i,x_k)$$



$$\phi_{ij}(x_i,x_j)$$

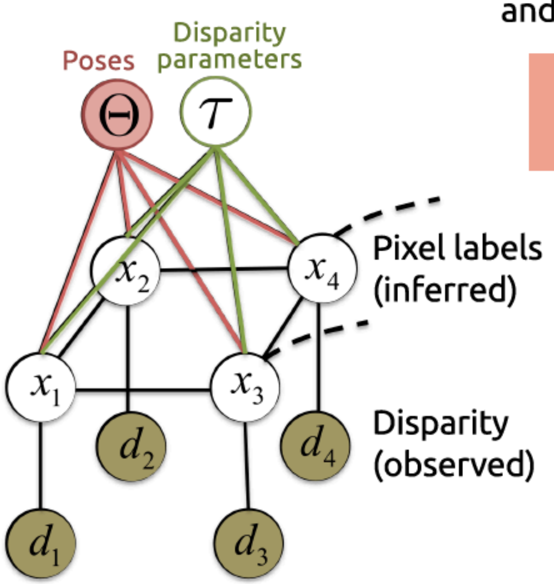
$$\lambda_1 \underbrace{\exp\left(\frac{-(d_i-d_j)^2}{2\sigma_c^2}\right)}_{\text{disparity smoothness}} + \lambda_2 \underbrace{\exp\left(\frac{-||\mathbf{v}_i-\mathbf{v}_j||_2^2}{2\sigma_v^2}\right)}_{\text{motion smoothness}} + \lambda_3 \underbrace{\exp\left(\frac{-(pb_i-pb_j)^2}{2\sigma_p^2}\right)}_{\text{colour smoothness}}$$

Energy function

sum over pixels

$$E(\boxed{\mathbf{x}};\boxed{\Theta},\boxed{\tau}) = \sum_{i\in\mathcal{V}} \phi_i(x_i;\boldsymbol{\Theta},\tau) + \sum_{(i,j)\in\mathcal{E}} \phi_{ij}(x_i,x_j) + \sum_{(i,k)\in\mathcal{E}^t} \phi^t_{ij}(x_i,x_k)$$

Articulated pose masks
and inferred depth cues

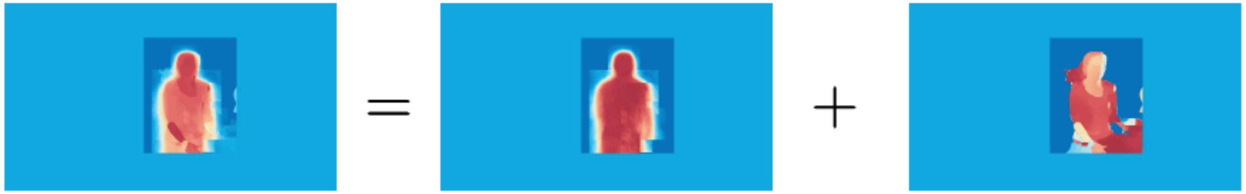


product over persons in front of person p

$$\phi_i(\boxed{x_i=p};\boxed{\Theta},\boxed{\tau}) = -\log\Big(\underbrace{\beta_i^p}_{\text{Positive evidence from current person}} \prod_{0\le m<\pi(p)} \underbrace{(1-\beta_i^m)}_{\text{Negative evidences from occluding people}}\Big)$$

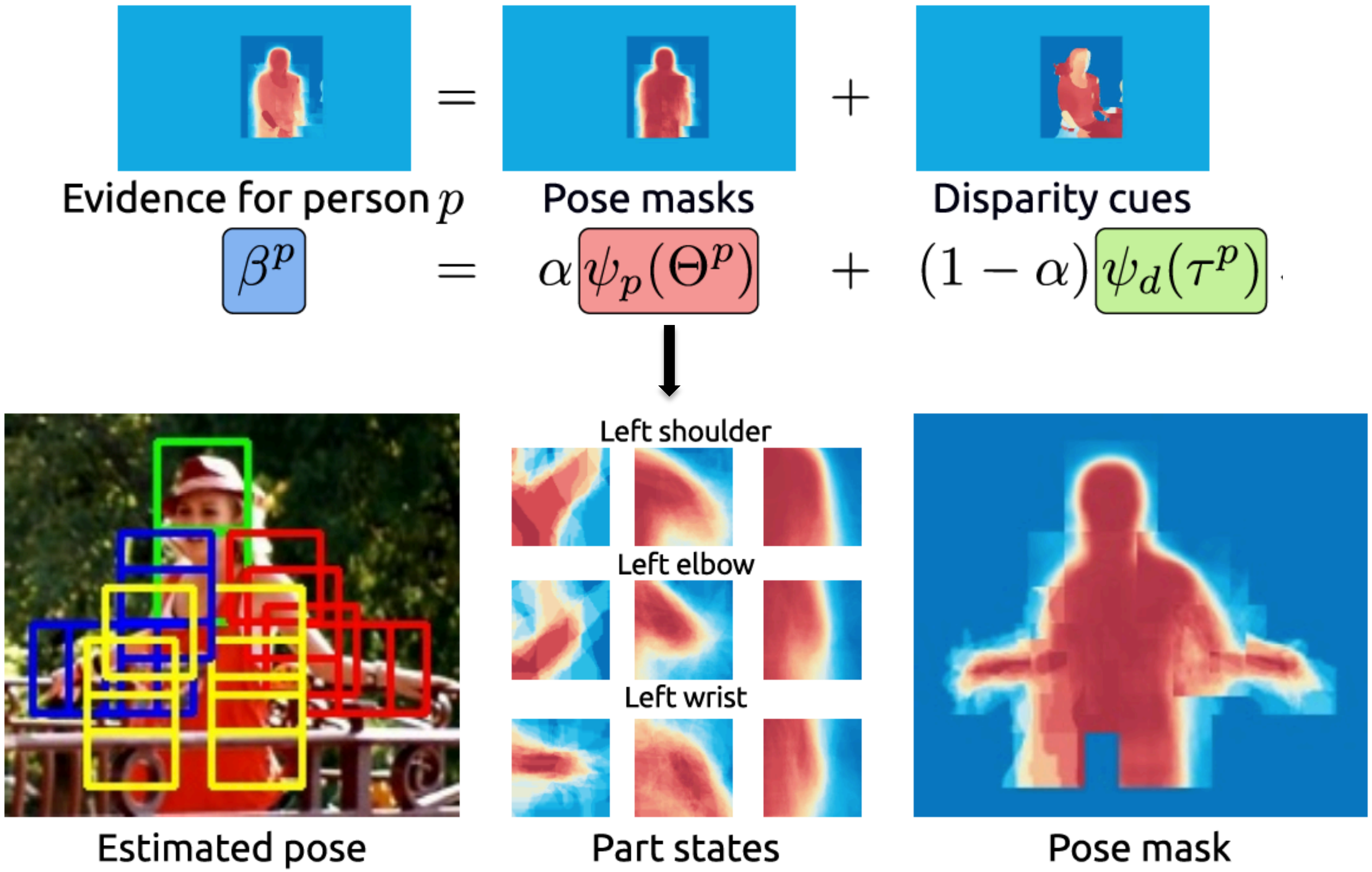
[Wang & Adelson 1994, Torr et al. 2001, Sun et al. 2010, Yang et al. 2011]

Energy function: Unary potential

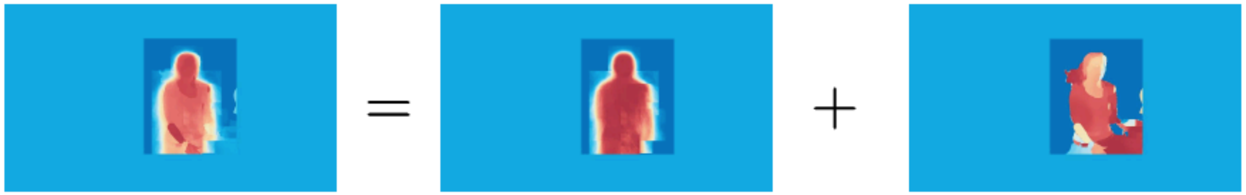


$$\boxed{\beta^p} = \alpha \boxed{\psi_p(\Theta^p)} + (1-\alpha)\boxed{\psi_d(\tau^p)}.$$

Energy function: Unary potential

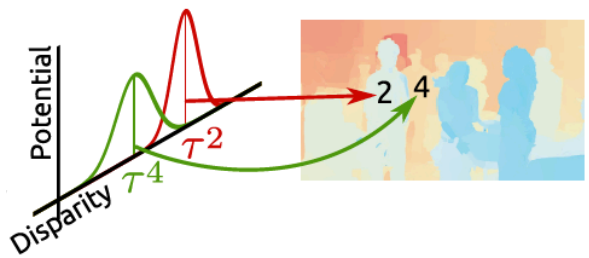


Energy function: Unary potential



Evidence for person p = Pose masks + Disparity cues

$$\boxed{\beta^p} = \alpha\boxed{\psi_p(\Theta^p)} + (1-\alpha)\boxed{\psi_d(\tau^p)}.$$



Energy function: Unary potential

Person
at the front

Back

Background

Energy function: Minimization

- Recall: 2-step approach
  - Estimate disparity parameters $\tau^* = \arg\min_{\{\tau\}} \tilde{E}(\tilde{\mathbf{x}}; \mathbf{\Theta}, \tau)$

$$E(\boxed{\mathbf{x}};\boxed{\mathbf{\Theta}},\boxed{\tau}) = \boxed{\sum_{i \in \mathcal{V}}} \boxed{\phi_i(x_i; \mathbf{\Theta}, \tau)} + \boxed{\sum_{(i,j) \in \mathcal{E}}} \boxed{\phi_{ij}(x_i, x_j)} + \boxed{\sum_{(i,k) \in \mathcal{E}^t}} \boxed{\phi^t_{ij}(x_i, x_k)}$$

Energy function: Minimization

- Recall: 2-step approach
  - Estimate disparity parameters $\tau^* = \arg\min_{\{\tau\}} \tilde{E}(\tilde{\mathbf{x}}; \boldsymbol{\Theta}, \tau)$

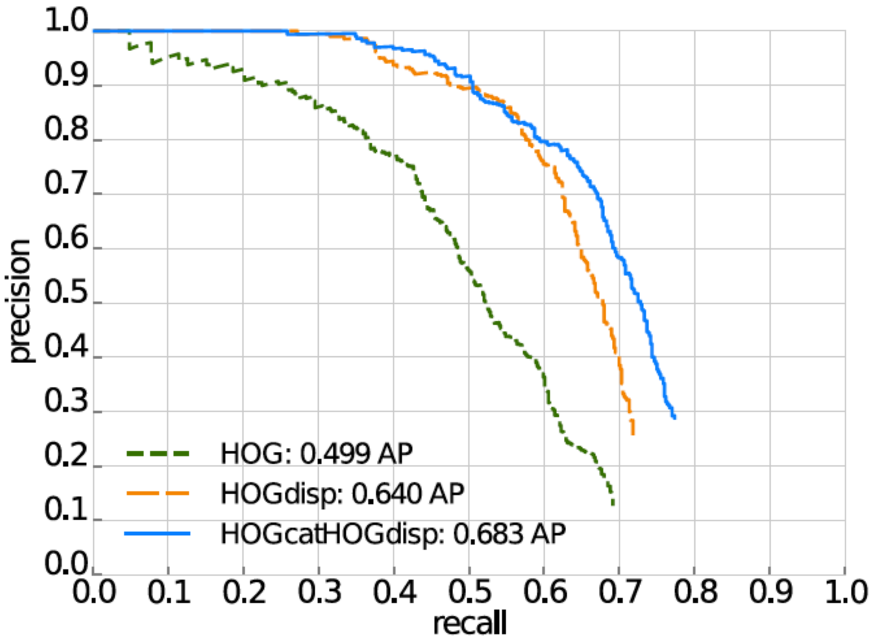
$$\tilde{E}(\boxed{\mathbf{x}};\boxed{\boldsymbol{\Theta}},\boxed{\tau}) = \boxed{\sum_{i \in \mathcal{V}}} \boxed{\phi_i(x_i; \boldsymbol{\Theta}, \tau)}$$

Energy function: Minimization

- Recall: 2-step approach
  - Estimate disparity parameters $\tau^* = \arg\min_{\{\tau\}} \tilde{E}(\tilde{\mathbf{x}}; \mathbf{\Theta}, \tau)$
  - Minimize $E(\mathbf{x}; \mathbf{\Theta}, \tau^*)$ [Boykov et al. 2001]

Detection & Pose Estimation Results

Person detection



- HOG: HOG on RGB only
- HOGdisp: HOG on disparity only
- HOGcatHOGdisp: concatenation of both

Pose estimation

	Yang	HOG	HOGdisp	HOGcomb
Head	0.976	0.983	**0.993**	0.986
Shoulders	0.935	0.931	0.947	**0.969**
Elbows	0.658	0.665	0.759	**0.784**
Wrists	0.298	0.294	0.297	**0.400**
Hips	0.563	0.705	0.714	**0.757**
Global	0.686	0.716	0.742	**0.779**

APK measure

Segmentation Results

Results on Inria 3DMovie Dataset

Method	Precision	Recall	Int. vs Union
Proposed	**0.869**	**0.915**	**0.804**
Variants of our method:			
No mask, single frame	0.525	0.371	0.278
Uni mask, single frame	0.783	0.641	0.544
Pose mask, single frame	0.849	0.905	0.779
Baselines:			
Colours only	0.778	0.769	0.630
Eichner, 2012	0.762	0.853	0.662

Segmentation Results

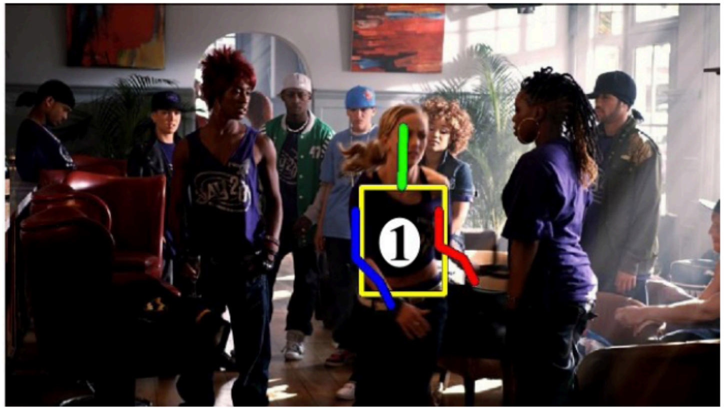
- H2view dataset [Sheasby, Valentin, Crook, Torr, 2012]

Method	Int. vs Union
Upper body segmentation:	
Sheasby, 2012	0.735
Proposed	**0.825**
Full body segmentation:	
Sheasby, 2012	0.692
Proposed	**0.706**

Segmentation Results

Summary: Part I

Video sequence



Pose

Segmentation
+
Layered order

Pose Estimation and Segmentation
of Multiple People in (Stereoscopic) Videos

+

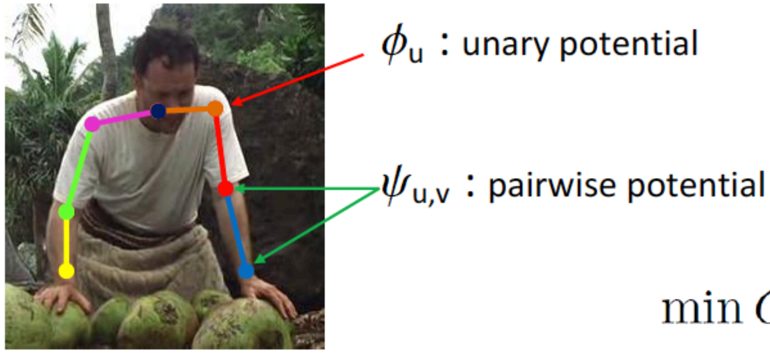
Human Pose Estimation in Videos

Human Pose Estimation

Poses in the Wild dataset: Cherian et al., CVPR 2014

Human Pose Estimation (Image)

- Formulated as a graph optimization problem



For an image *I*, pose model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and

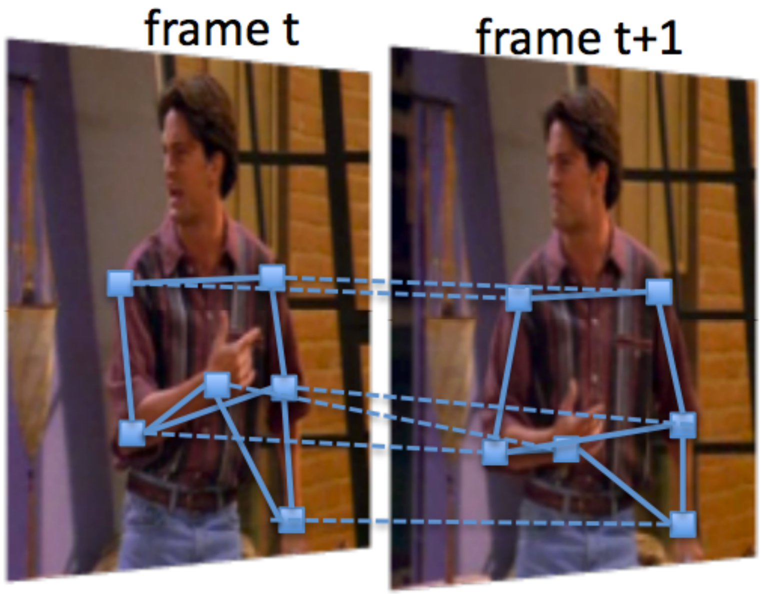
$$p = \{p^u = (x^u, y^u) \in \mathbb{R}^2 : \forall u \in \mathcal{V}\}$$

$$\min C(I, p) := \sum_{u \in \mathcal{V}} \phi_u(I, p^u) + \sum_{(u,v) \in \mathcal{E}} \psi_{u,v}(p^u - p^v)$$

Yang and Ramanan, CVPR 2011

Human Pose Estimation (Video)

- Extension to videos: introduce temporal links
- Inference is now computationally intensive – requires approximate methods



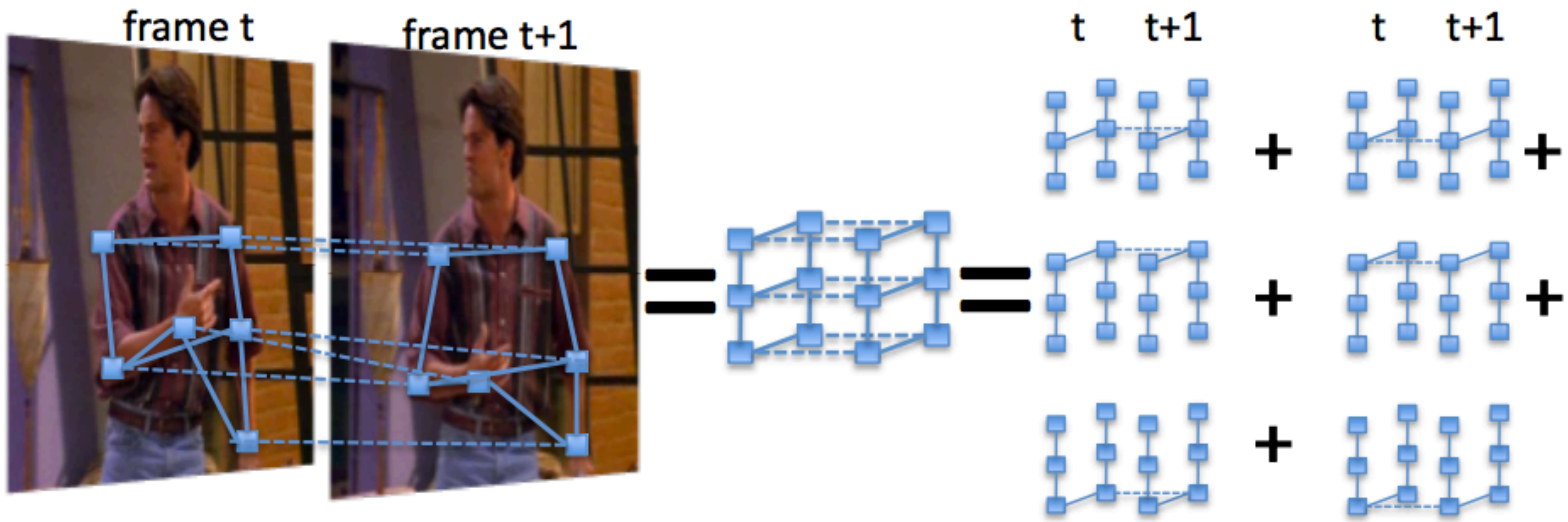
e.g., Sapp et al., '11, Tokola et al., '13

Human Pose Estimation (Video)

- Extension to videos: introduce temporal links
- Inference is now computationally intensive – requires approximate methods
- e.g.,
  - Change graph structure [Sapp et al. ’11, Weiss et al. ’11]
  - Use approximate inference [Ferrari et al. ’08, Wang et al. ’08, Park & Ramanan ’11, Tokola et al. ’13]

Human Pose Estimation (Video)

- Approximate the graph as combination of trees

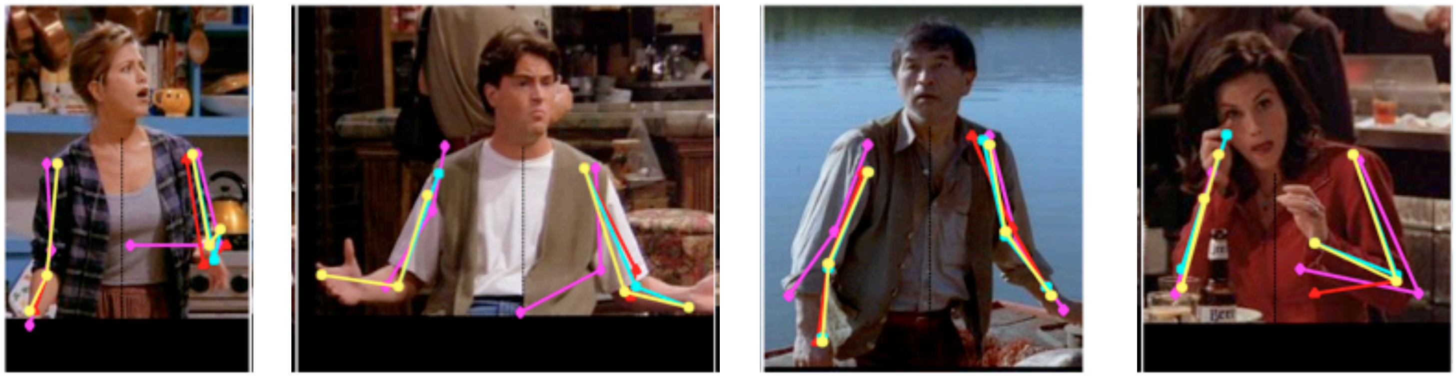


- Computationally expensive for long sequences

Sapp, Weiss, Taskar, CVPR 2011

Human Pose Estimation (Video)

- Compute a candidate set of poses in each frame
- Then, track (entire pose or pose-parts) over time



- Limited by the no. of candidates or regularization

Park and Ramanan, ICCV 2011; Ramakrishna, Kanade, Sheikh, CVPR 2013

Our Pose Estimation Approach

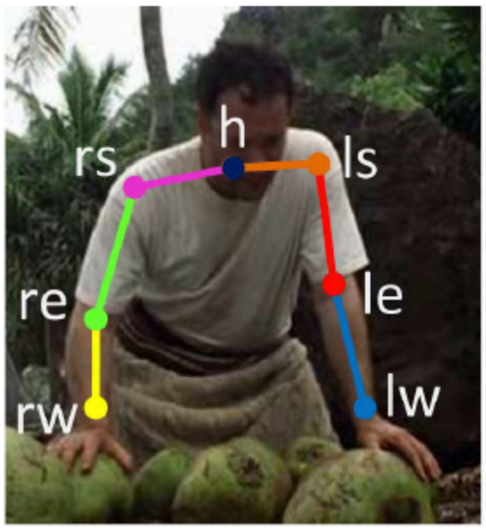
- Combines
  1. Candidate pose set
     - Generate better candidates

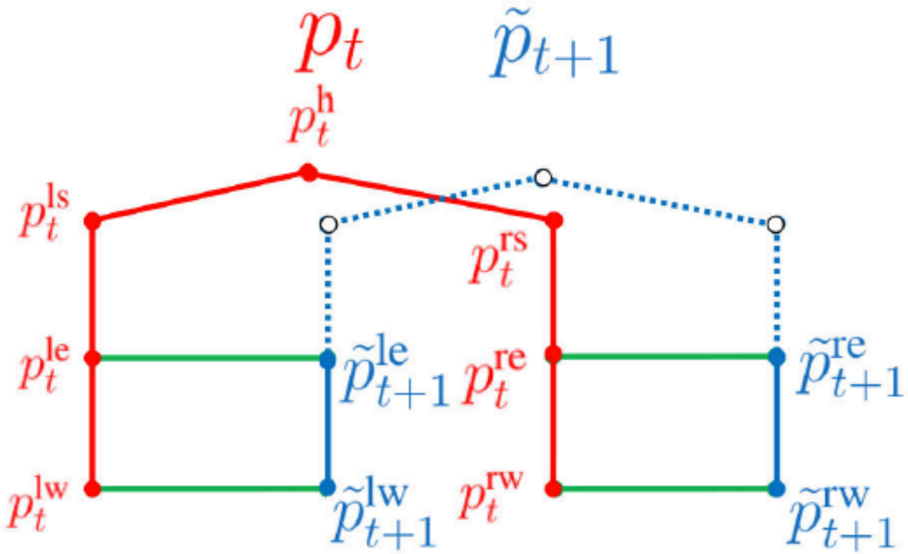
  2. Decomposition strategy
     - Generate limb sequences and recompose the pose

Cherian, Mairal, Alahari, Schmid, CVPR 2014

Better Candidate Poses

- Stabilize the lower-limb pose estimates

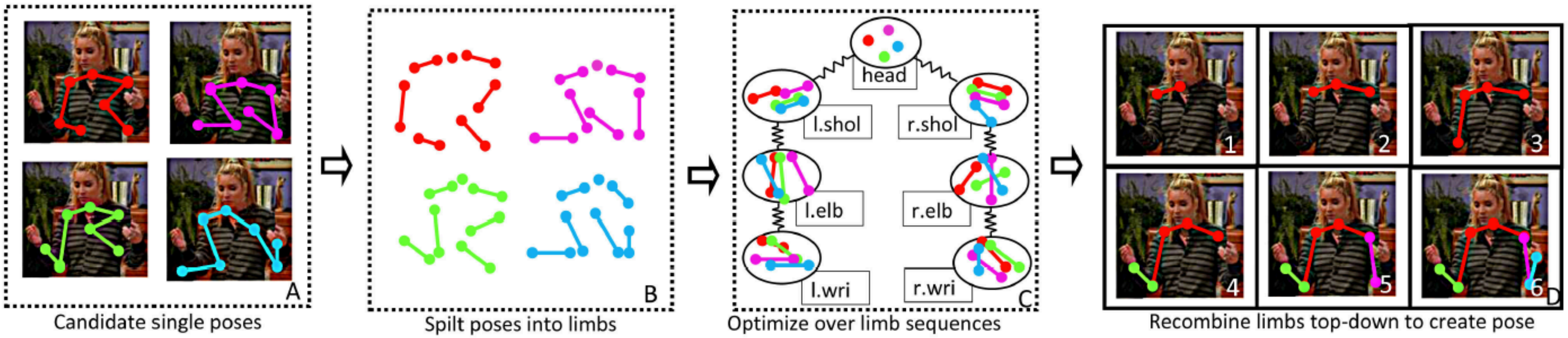




$$C(I_t, p_t) + \tilde{C}(I_{t+1}, \tilde{p}_{t+1}) + \tilde{\lambda}_1 \sum_{u \in \mathcal{W}} \|\tilde{p}_{t+1}^u - p_t^u - f_t(p_t^u)\|_2^2$$

De/Re- composition

- Decompose poses and perform limb-tracking



Pose Estimation Video Datasets

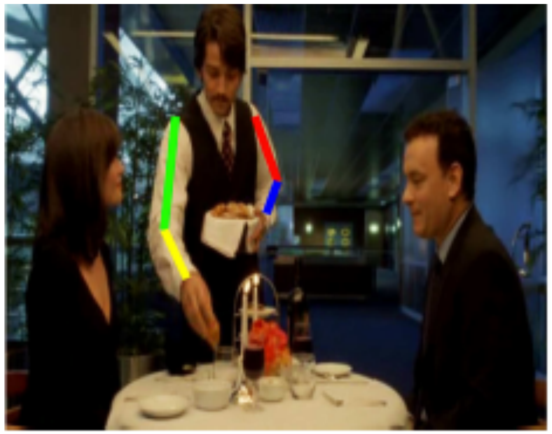
- VideoPose [Sapp et al. '11]

- MPII Cooking Activities [Rohrbach et al. '12]

- Interesting preliminary benchmarks, but
  - Limited occlusion, shot indoors, static camera, pre-processed (head alignment)

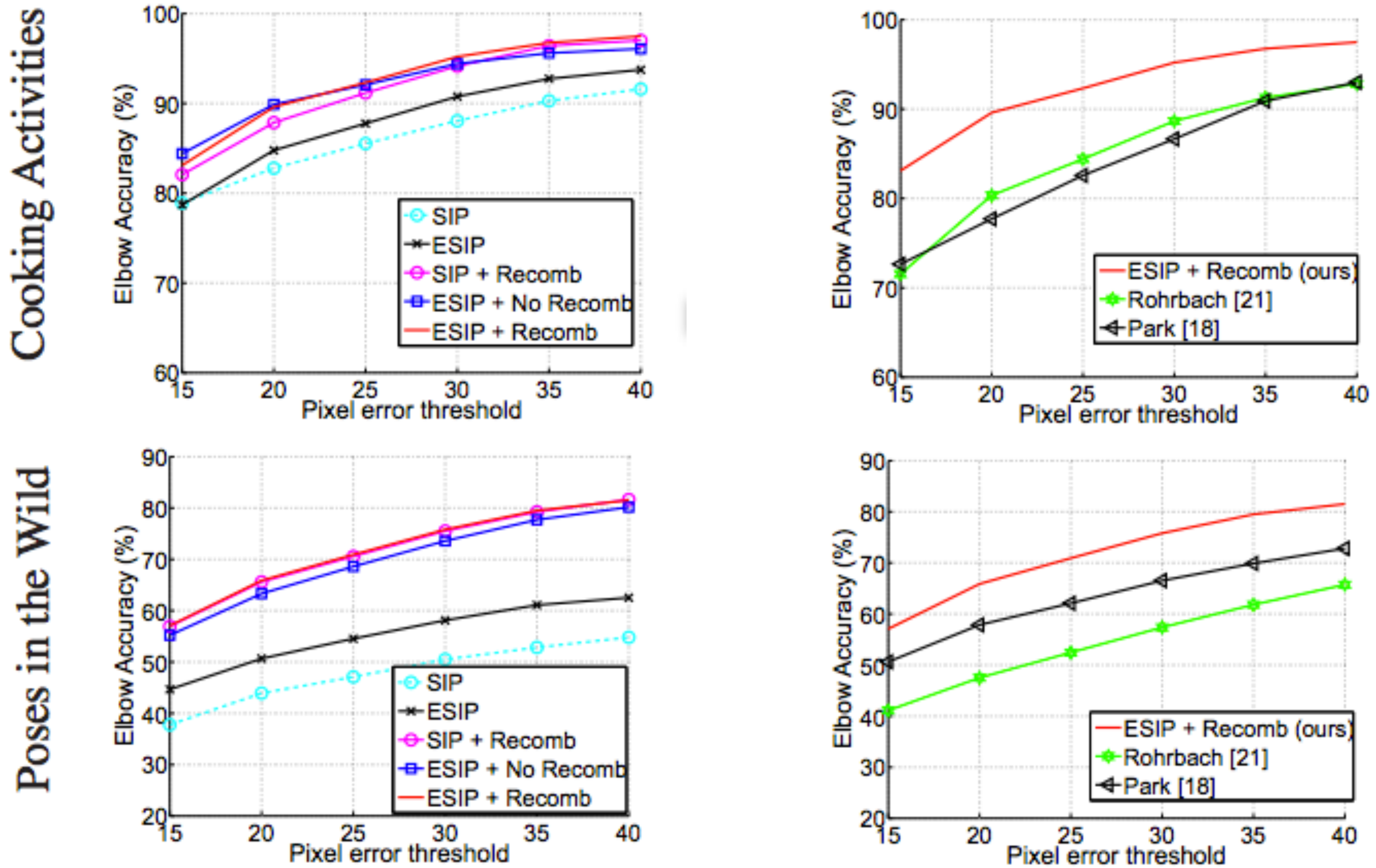
Poses in the Wild Dataset

- 30 (test) sequences from 3 Hollywood movies
- Manually annotated upper-body pose in ~900 frames

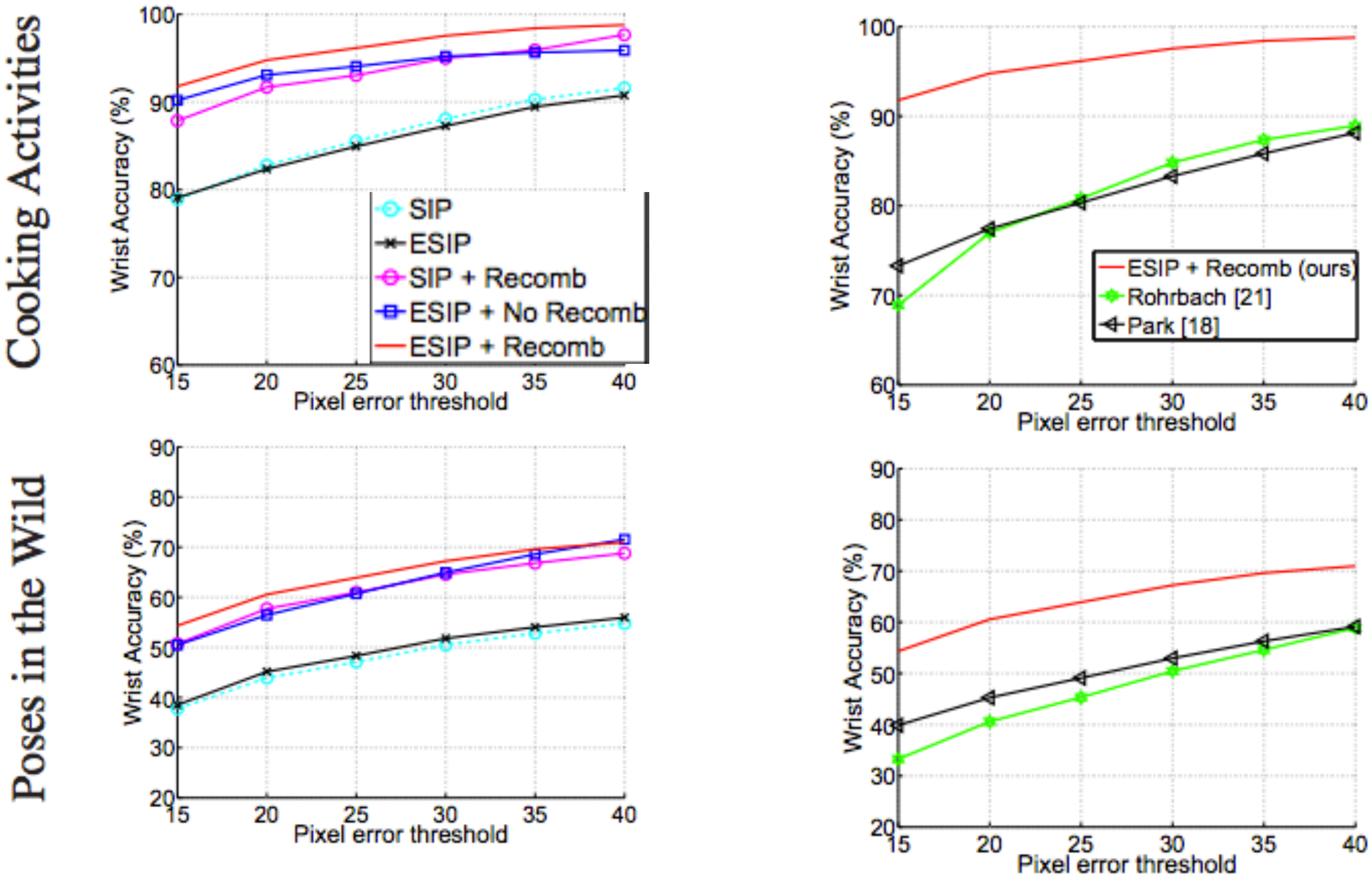
  

Available at: http://lear.inrialpes.fr/research/posesinthewild/

Human Pose Estimation: Elbows

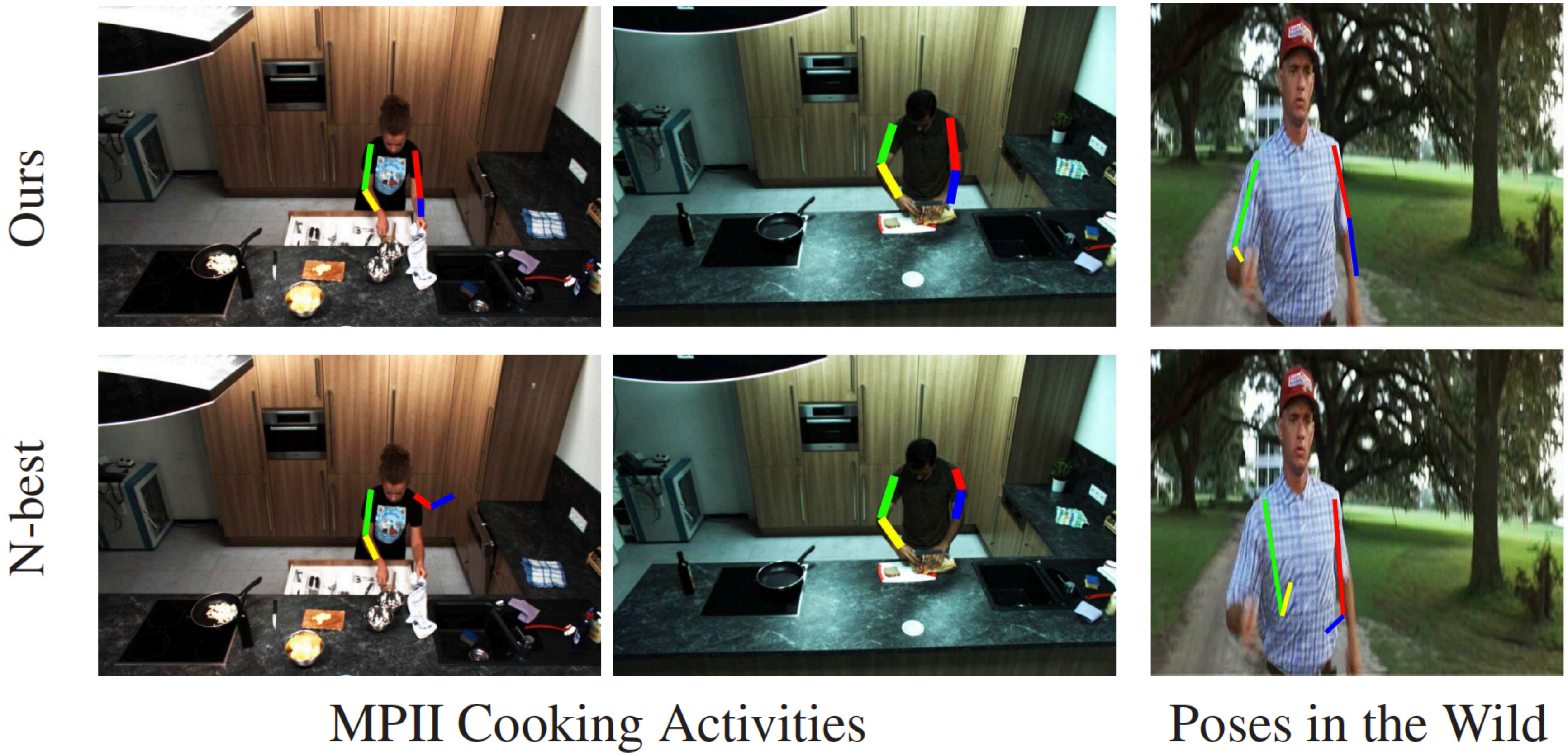


Human Pose Estimation: Wrists

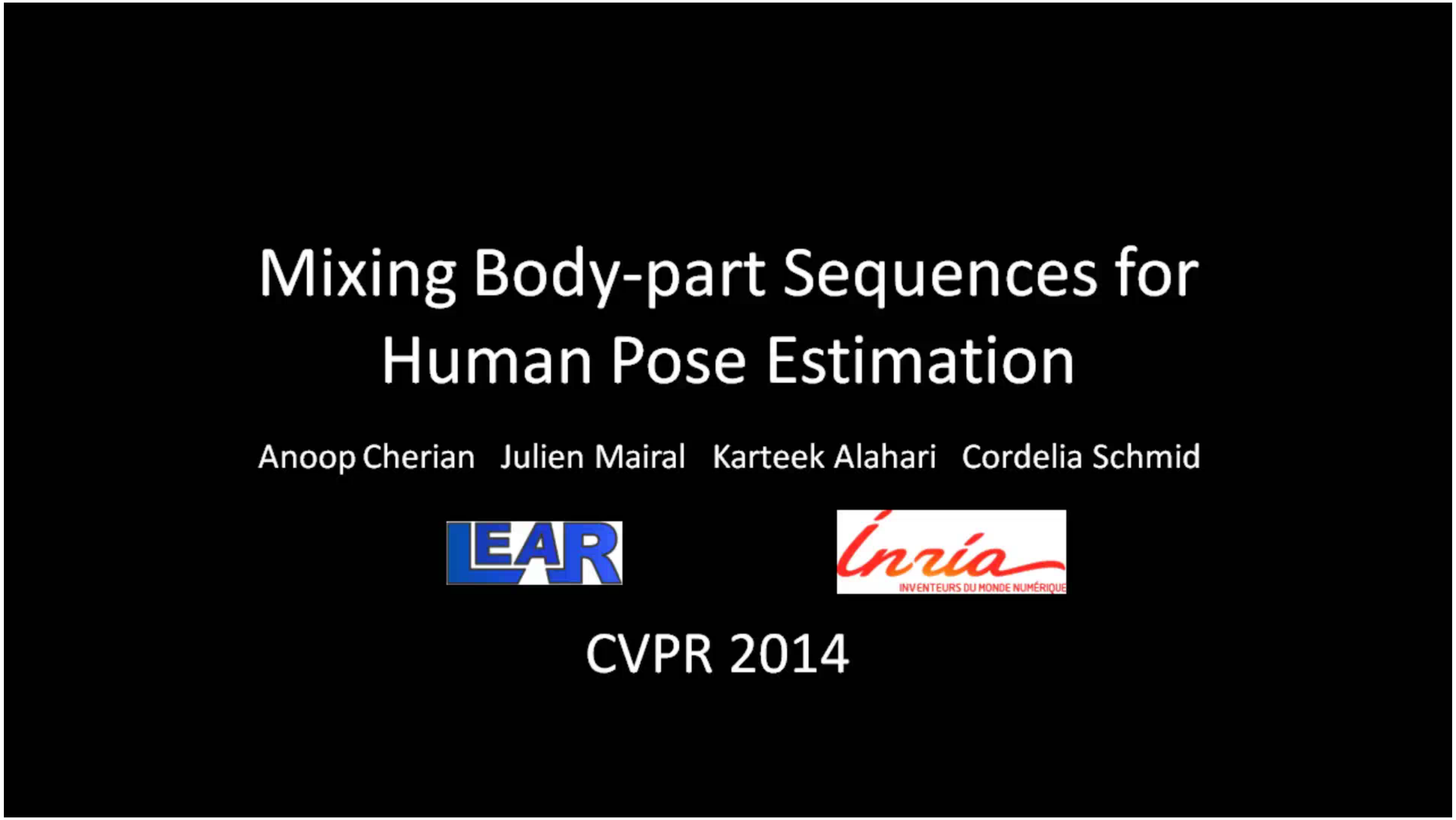


Benefits of decomposition

Benefits of decomposition



Human Pose Estimation



In summary

Pose Estimation and Segmentation
of Multiple People in (Stereoscopic) Videos

+

Human Pose Estimation in Videos