Pose Estimation and Segmentation of People in 3D Movies

Karteek Alahari **Guillaume Seguin** **Josef Sivic** **Ivan Laptev**

WILLOW Team - Inria / École normale supérieure / CNRS - Paris, France

Partly supported by EIT ICT Labs, ERC Activia, Google, MSR-Inria, Quaero.

Goal

Obtain
- pixel-wise **segmentation**
- **pose** estimates
- depth **ordering**

for **multiple people** in **stereo movies**

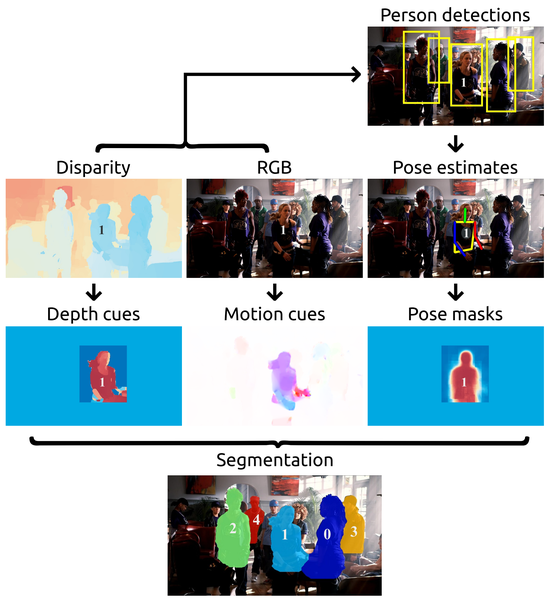
Motivation

- Many movies are now available in 3D (stereo)
- Develop a **mid-level representation** of stereoscopic video for recognition, editing, navigation
- Investigate **benefits of disparity** cues for segmentation and pose estimation
- Collect (noisy) **training data** of segmented people for monocular video

Contributions

1. A **multi-person segmentation model** for stereo videos
   - combines multiple cues: disparity, colour, motion
   - incorporates learnt **pose-specific** segmentation **masks**
   - explicitly represents depth ordering and occlusions
2. A **new annotated Inria 3DMovie Dataset** for person detection, pose estimation and segmentation in indoor and outdoor scenes

Overview



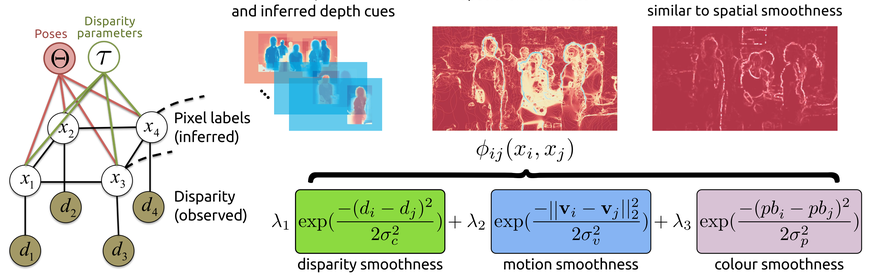
Related work

[1] Shotton *et al.*, CVPR '11
[2] Sheasby, Valentin, Crook and Torr, ACCV '12
[3] Yang, Hallman, Ramanan and Fowlkes, CVPR '10
[4] Yang and Ramanan, CVPR '11
[5] Eichner, Marin-Jimenez, Zisserman and Ferrari, IJCV '12

Model

$$E(\mathbf{x};\Theta;\tau) = \underbrace{\sum_{i\in\mathcal{V}} \phi_i(x_i;\Theta,\tau)}_{\text{sum over pixels}} + \underbrace{\sum_{(i,j)\in\mathcal{E}} \phi_{ij}(x_i,x_j)}_{\text{sum over spatial edges}} + \underbrace{\sum_{(i,k)\in\mathcal{E}^t} \phi^t_{ij}(x_i,x_k)}_{\text{sum over temporal edges}}$$

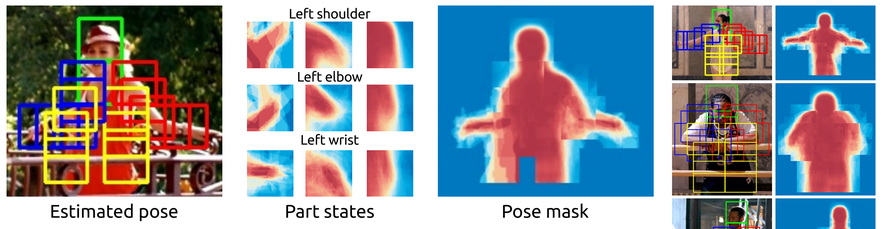
Articulated pose masks and inferred depth cues | Spatial smoothness | Temporal smoothness: similar to spatial smoothness



$$\phi_{ij}(x_i,x_j) = \lambda_1 \exp\left(\frac{-(d_i-d_j)^2}{2\sigma_c^2}\right) + \lambda_2 \exp\left(\frac{-\|\mathbf{v}_i-\mathbf{v}_j\|_2^2}{2\sigma_v^2}\right) + \lambda_3 \exp\left(\frac{-(pb_i-pb_j)^2}{2\sigma_p^2}\right)$$

disparity smoothness | motion smoothness | colour smoothness

Articulated pose masks



Estimated pose | Part states | Pose mask

- Poses are estimated as part positions and part states
- Pose masks are composed from state-specific part masks
- State-specific part masks are learnt from ground truth segmentations

Inferring depth ordering

1. τ determines the ordering of people in the scene
   There is one scalar disparity parameter τ^p for each person p
2. Given a depth ordering:

Person-specific unary: $\phi_i(x_i = p;\Theta,\tau) = -\log\left(\beta_i^p \prod_{0\le m<\pi(p)} (1-\beta_i^m)\right)$

(product over persons in front of person p)

Positive evidence from current person | Negative evidences from occluding people

Example with persons 2 & 4:
Correct ordering: 2 in front of 4
Wrong ordering: 4 in front of 2

3. Evidence for person p = Pose masks + Disparity cues

$$\beta^p = \alpha\,\psi_p(\Theta^p) + (1-\alpha)\,\psi_d(\tau^p)$$

4. To find the best depth ordering (given by τ^*), we solve: $\tau^* = \arg\min_{\{\tau\}} \sum_{i\in\mathcal{V}} \min_{l\in\mathcal{L}} \phi_i(x_i = l;\Theta,\tau)$
5. Final potentials using τ^*:

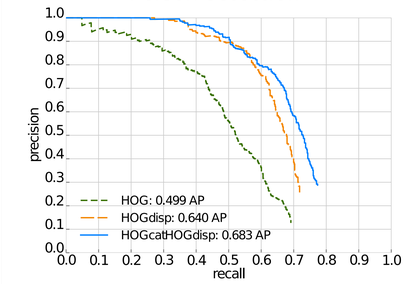
Inria 3DMovie dataset

- Annotated stereo pairs from movies "StreetDance 3D" and "Pina"
- 440 training stereo pairs ; 36 test video sequences — 2727 pairs
- **Labelling:** 686 person segmentations, 587 poses, 1158 person boxes

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

Person detection & Pose estimation

Person detection



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

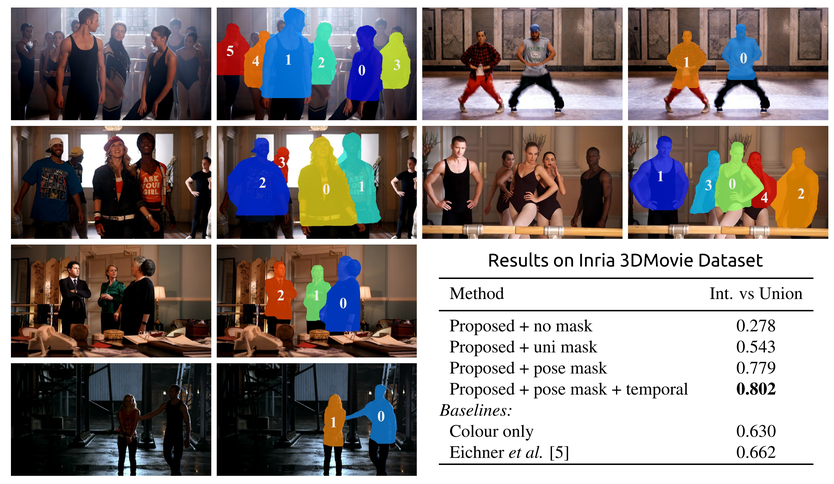
Pose estimation

	[4]*	HOG	HOGdisp	HOGcomb
Head & Torso	0.989	0.989	0.991	**0.998**
Upper arms	0.839	0.856	0.869	**0.889**
Lower arms	0.518	0.559	0.535	**0.594**
Global	0.782	0.802	0.799	**0.827**

* This model was trained on the Buffy dataset

We use the PCP measure from Eichner et al. [5] which evaluates the percentage of correctly estimated body parts

Person segmentation

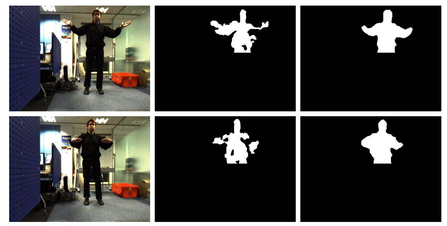


Results on Inria 3DMovie Dataset

Method	Int. vs Union
Proposed + no mask	0.278
Proposed + uni mask	0.543
Proposed + pose mask	0.779
Proposed + pose mask + temporal	**0.802**
Baselines:	
Colour only	0.630
Eichner *et al.* [5]	0.662

Results on H2view [2]

Method	Int. vs Union
Upper body segmentation:	
Sheasby *et al.* [2]	0.735
Proposed + pose mask	0.814
Proposed + pose mask + temporal	**0.825**
Full body segmentation:	
Sheasby *et al.* [2]	0.692
Proposed + upper pose mask	0.706



Evaluation: the intersection vs. union measure between each detected person segmentation and the corresponding ground truth