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)
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


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](http://www.di.ens.fr/willow/research/stereoseg)
3D in the Wild: The Goal

- Layered segmentation of people in stereoscopic videos
3D in the Wild: The Goal

- 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

Disparity → RGB → Pose estimates

 Depth cues → Motion cues → Pose masks

Person detections

Segmentation

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 ($\Theta$): The pose of each person
  - Layers ($\tau$): The layered ordering of people
Overview

• Define the estimation as:

\[ \{x^*, \Theta^*, \tau^*\} = \arg \min_{x, \Theta, \tau} E(x, \Theta, \tau) \]

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

• Approximate it as:

\[ \{x^*, \tau^*\} = \arg \min_{x, \tau} E(x, \tau; \Theta) \]
Overview

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

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

[Boykov et al. 2001]
Energy function

\[ E(x; \Theta, \tau) = \sum_{i \in V} \phi_i(x_i; \Theta, \tau) + \sum_{(i,j) \in E} \phi_{ij}(x_i, x_j) + \sum_{(i,k) \in E_t^t} \phi_{ij}^t(x_i, x_k) \]
Energy function

\[
E(x; \Theta, \tau) = \sum_{i \in V} \phi_i(x_i; \Theta, \tau) + \sum_{(i,j) \in E} \phi_{ij}(x_i, x_j) + \sum_{(i,k) \in E_t} \phi^t_{ij}(x_i, x_k)
\]

sum over temporal edges

Temporal smoothness: similar to spatial smoothness
Energy function

\[
E(x; \Theta, \tau) = \sum_{i \in V} \phi_i(x_i; \Theta, \tau) + \sum_{(i,j) \in E} \phi_{ij}(x_i, x_j) + \sum_{(i,k) \in E^t} \phi_{ij}^t(x_i, x_k)
\]

Spatial smoothness

\[
\phi_{ij}(x_i, x_j)
\]

\[
\lambda_1 \exp \left( \frac{-(d_i - d_j)^2}{2\sigma_c^2} \right) + \lambda_2 \exp \left( \frac{-||v_i - v_j||^2}{2\sigma_v^2} \right) + \lambda_3 \exp \left( \frac{-(pb_i - pb_j)^2}{2\sigma_p^2} \right)
\]

disparity smoothness
motion smoothness
colour smoothness
Energy function

\[ E(x; \Theta, \tau) = \sum_{i \in V} \phi_i(x_i; \Theta, \tau) + \sum_{(i,j) \in E} \phi_{ij}(x_i, x_j) + \sum_{(i,k) \in E^t} \phi_{ij}^t(x_i, x_k) \]

Articulated pose masks and inferred depth cues

\[ \phi_i(x_i = p; \Theta, \tau) = -\log \left( \beta_i^p \prod_{0 < m < \pi(p)} (1 - \beta_i^m) \right) \]

Positive evidence from current person

Negative evidence from occluding people

Energy function: Unary potential

\[
\beta^p = \alpha \psi_p(\Theta^p) + (1 - \alpha) \psi_d(\tau^p)
\]
Energy function: Unary potential

\[
\beta^p = \alpha \psi_p(\Theta^p) + (1 - \alpha) \psi_d(\tau^p)
\]
Energy function: Unary potential

\[ \beta_p \quad = \quad \alpha \psi_p(\Theta_p) \quad + \quad (1 - \alpha)\psi_d(\tau^p) \]
Energy function: Unary potential
Energy function: Minimization

• Recall: 2-step approach

  – Estimate disparity parameters \( \tau^* = \arg\min_{\{\tau\}} \tilde{E}(\mathbf{x}; \Theta, \tau) \)

\[
E(\mathbf{x}; \Theta, \tau) = \sum_{i \in \mathcal{V}} \phi_i(x_i; \Theta, \tau) + \sum_{(i,j) \in \mathcal{E}} \phi_{ij}(x_i, x_j) + \sum_{(i,k) \in \mathcal{E}^t} \phi^t_{ij}(x_i, x_k)
\]
Energy function: Minimization

- Recall: 2-step approach

\[
\tau^* = \arg \min_{\tau} \tilde{E}(\tilde{x}; \Theta, \tau)
\]

\[
\tilde{E}(\tilde{x}; \Theta, \tau) = \sum_{i \in V} \phi_i(x_i; \Theta, \tau)
\]
Energy function: Minimization

• Recall: 2-step approach

  – Estimate disparity parameters $\tau^* = \arg \min_{\{\tau\}} \tilde{E}(\tilde{x}; \Theta, \tau)$

  – Minimize $E(x; \Theta, \tau^*)$ [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

<table>
<thead>
<tr>
<th></th>
<th>Yang</th>
<th>HOG</th>
<th>HOGdisp</th>
<th>HOGcomb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>0.976</td>
<td>0.983</td>
<td>0.993</td>
<td>0.986</td>
</tr>
<tr>
<td>Shoulders</td>
<td>0.935</td>
<td>0.931</td>
<td>0.947</td>
<td>0.969</td>
</tr>
<tr>
<td>Elbows</td>
<td>0.658</td>
<td>0.665</td>
<td>0.759</td>
<td>0.784</td>
</tr>
<tr>
<td>Wrists</td>
<td>0.298</td>
<td>0.294</td>
<td>0.297</td>
<td>0.400</td>
</tr>
<tr>
<td>Hips</td>
<td>0.563</td>
<td>0.705</td>
<td>0.714</td>
<td>0.757</td>
</tr>
<tr>
<td>Global</td>
<td>0.686</td>
<td>0.716</td>
<td>0.742</td>
<td>0.779</td>
</tr>
</tbody>
</table>

APK measure
Segmentation Results

Results on Inria 3DMovie Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>Int. vs Union</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.869</td>
<td>0.915</td>
<td>0.804</td>
</tr>
<tr>
<td>Variants of our method:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No mask, single frame</td>
<td>0.525</td>
<td>0.371</td>
<td>0.278</td>
</tr>
<tr>
<td>Uni mask, single frame</td>
<td>0.783</td>
<td>0.641</td>
<td>0.544</td>
</tr>
<tr>
<td>Pose mask, single frame</td>
<td>0.849</td>
<td>0.905</td>
<td>0.779</td>
</tr>
<tr>
<td>Baselines:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colours only</td>
<td>0.778</td>
<td>0.769</td>
<td>0.630</td>
</tr>
<tr>
<td>Eichner, 2012</td>
<td>0.762</td>
<td>0.853</td>
<td>0.662</td>
</tr>
</tbody>
</table>
Segmentation Results

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Int. vs Union</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Upper body segmentation:</strong></td>
<td></td>
</tr>
<tr>
<td>Sheasby, 2012</td>
<td>0.735</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.825</td>
</tr>
<tr>
<td><strong>Full body segmentation:</strong></td>
<td></td>
</tr>
<tr>
<td>Sheasby, 2012</td>
<td>0.692</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.706</td>
</tr>
</tbody>
</table>
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

\[
\begin{align*}
\phi_u &: \text{unary potential} \\
\psi_{u,v} &: \text{pairwise potential} \\
\end{align*}
\]

For an image \( I \), pose model \( G = (\mathcal{V}, \mathcal{E}) \), 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]
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
Better Candidate Poses

• Stabilize the lower-limb pose estimates

\[ C(I_t, p_t) + \tilde{C}(I_{t+1}, \tilde{p}_{t+1}) + \lambda_1 \sum_{u \in \mathcal{W}} \|\tilde{p}_{t+1}^u - p_t^u - f_t(p_t^u)\|_2^2 \]
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

Cooking Activities

Poses in the Wild
Human Pose Estimation: Wrists

Cooking Activities

Poses in the Wild
Benefits of decomposition

Ours

N-best

VideoPose
Benefits of decomposition

MPII Cooking Activities

Poses in the Wild
Human Pose Estimation

Mixing Body-part Sequences for Human Pose Estimation

Anoop Cherian  Julien Mairal  Karteek Alahari  Cordelia Schmid

CVPR 2014
In summary
Pose Estimation and Segmentation of Multiple People in (Stereoscopic) Videos

+ Human Pose Estimation in Videos