Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images

Authors:
Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, Cristian Sminchisescu

Presentor:
Jack Vice
Problem: Human Localization and Grouping

- e.g. Autonomous vehicle navigation
Objective

Single Input Image

Automatic 3D position and Shape estimation
Challenges:

- Multiple people
- Various activities / orientations
- Occlusions
- Depth ambiguities
- Formulate a single cost function
- Integrated learning process
Integrated Localization and Shape Estimation

- Feed-forward, multi-task bottom up solution for localization
- 3d pose and shape estimator
- Identifies human body structures
- Groups people based on 2D and 3D information
- 2d and 3d data are fused using learned scoring functions
- Aggregates solutions into 3d skeleton hypotheses under kinematic tree constraints
Approach: Find the Skeletal Joints

- Detect body joints and their type
- Solving a global optimal assignment problem
- Find all connected components satisfying kinematic skeletal tree
Ground Truth: Panoptic Studio
Computational Pipeline
Deep Volume Encoding

- Inputs: image features
- Encode 2d and 3d skeletons
- The Human3.6M dataset used for 2d and 3d physiology
- Total of 17 joints defined as $N_J$
- Total of 16 limbs as $N_L$
- Kinematic tree of connected limbs
● Deep volume encoding of multiple ground truth skeletons in a scene
1. Associate a slice (column) in the volume to each joint components
2. Encode each 3d skeleton associated with a person
3. For columns ‘intercepting’ spatial locations of image projection of skeleton
   a. Write each of its components in the corresponding slice
- Image Features

- 2d Prediction from previous stage

- Current 2d representation

- 3d Prediction from previous stage

- Current 3d representation
2d / 3d Encoding Loss

\[ L_I = \sum_{1 \leq x \leq h, 1 \leq y \leq w} \rho_{2d}(M_{2d}(x, y), G_{2d}(x, y)) + \sum_{1 \leq x \leq h, 1 \leq y \leq w} \rho_{3d}(M_{3d}(x, y), G_{3d}(x, y)) \]

\( G_{3d}(x, y) \) is valid
Map all Skeletons

1. Detect potential human body joints and their type in images
2. Scoring joints using trainable functions
3. Find all connected components satisfying strong kinematic tree constraints
4. Each component being a different person
Limb Scoring
Limb Scoring

- Extract Joint proposals
- Elbo, wrist, knee, etc.
- List feasible connections
- Limb scoring network
- Limb map and 2D input to form limb list
- Cross entropy loss using ground truth
Binary Integer Program with constraints
Constraint 1: ensure that connected component fall on a single person
Constraint 2: ensure that connected components select at most one joint
Constraint 3: the connected components are as large as possible
People Grouping as a Binary Integer Programming Problem

$$x^*(c) = \arg \max_x c^\top x, \text{ subject to } Ax \leq b, x \in \{0, 1\}^{NL \times 1}$$

- Estimating the optimal $L^* \subseteq L$ such that graph $G = (J, L^*)$ abides by the three constraints
- Computing $L^*$ is equivalent to finding a binary indicator $x \in \{0, 1\}^{|L| \times 1}$ in the set $L$.
- Can be written as a row by row sparse matrix $A \in \{0, 1\}^{|L| \times |L|}$, that constrains $x$ such that $Ax \leq b$, where $b$ is the all-ones vector $1_{|L|}$. 
3d Pose Decoding and Shape Estimation

Skeletal Grouping

Image Features

\( j_{3d} \rightarrow \) Encoder

\( \theta^p \)  \( \beta^p \)  Decoder

\( j_S^p \)
Given the feature volume of a person and its skeleton

- Learn a function that attends to different regions in the 3D map to decode the position of each joint.
- Sample Multiple Points Along Each Limb
- MultiLayer perceptron assigns a score to each sample and each 3D joint
- Final predicted 3D skeleton is the weighted sum of 3D samples
- Loss is position error from ground truth
SMPL: A Skinned Multi-Person Linear Model
3D pose and Shape Estimation

Joint angle rotations $\theta$

Body Dimensions $B$
Deep Autoencoder

- Input Predicted 3d skeleton
- Predicts Pose parameters $\theta$
- Predicts the shape parameters $B$
- Loss is error between input and output

MLP with ReLU's

$Rvec(V(\theta, \beta))$
$R$: regression Matrix
$V$: All 3D vertices in mesh
Final Solution: 3d Pose and Shape Estimation
Analysis of learning

![Limb Vector Components](chart.png)
Linear growth in time
Quantitative Results

Human3.6M dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
<th>A9</th>
<th>A10</th>
<th>A11</th>
<th>A12</th>
<th>A13</th>
<th>A14</th>
<th>A15</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>[22]</td>
<td>60</td>
<td>56</td>
<td>68</td>
<td>64</td>
<td>78</td>
<td>67</td>
<td>68</td>
<td>106</td>
<td>119</td>
<td>77</td>
<td>85</td>
<td>64</td>
<td>57</td>
<td>78</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>[27]</td>
<td>54</td>
<td>54</td>
<td>63</td>
<td>59</td>
<td>72</td>
<td>61</td>
<td>68</td>
<td>101</td>
<td>109</td>
<td>74</td>
<td>81</td>
<td>62</td>
<td>55</td>
<td>75</td>
<td>60</td>
<td>69</td>
</tr>
<tr>
<td>MubyNet</td>
<td>49</td>
<td>47</td>
<td>51</td>
<td>52</td>
<td>60</td>
<td>56</td>
<td>56</td>
<td>82</td>
<td>94</td>
<td>64</td>
<td>69</td>
<td>61</td>
<td>48</td>
<td>66</td>
<td>49</td>
<td>60</td>
</tr>
</tbody>
</table>

- Mean Per Joint Position Error for actions A1 - A15
- [22] “Deep Multitasking architecture for integrated 2d and 3d sensing”, 2017
- [27] “Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes”, 2018
Human80K dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MPJPE (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[22]</td>
<td>63.35</td>
</tr>
<tr>
<td>MubyNet</td>
<td>59.31</td>
</tr>
<tr>
<td>MubyNet attention</td>
<td>58.40</td>
</tr>
</tbody>
</table>

MPJPE: Mean Per Joint Position Error

- Attention Mechanism improves performance
- [22] “Deep Multitasking architecture for integrated 2d and 3d sensing”, 2017
CMU Panoptic dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Haggling</th>
<th>Mafia</th>
<th>Ultimatum</th>
<th>Pizza</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>[22]</td>
<td>217.9</td>
<td>187.3</td>
<td>193.6</td>
<td>221.3</td>
<td>203.4</td>
</tr>
<tr>
<td>[27]</td>
<td>140.0</td>
<td>165.9</td>
<td>150.7</td>
<td>156.0</td>
<td>153.4</td>
</tr>
<tr>
<td>MubyNet</td>
<td>141.4</td>
<td>152.3</td>
<td>145.0</td>
<td>162.5</td>
<td>150.3</td>
</tr>
<tr>
<td>MubyNet fine-tuned</td>
<td>72.4</td>
<td>78.8</td>
<td>66.8</td>
<td>94.3</td>
<td>72.1</td>
</tr>
</tbody>
</table>

- Fine tuning on the CMU Panoptic dataset (74,936 samples)
- [22] “Deep Multitasking architecture for integrated 2d and 3d sensing”, 2017
- [27] “Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes”, 2018
Qualitative Results
Questions?