Light3DPose

Real-time Multi-Person 3D Pose Estimation from Multiple Views

Alessio Elmi - Davide Mazzini - Pietro Tortella
{alessio, davide, pietro}@standard.ai
Introduction

Images credits:
Problem Statement

Inputs
- Multiple calibrated camera views

Output
- 3D skeletons

Our focus
3D reconstruction of a single time instant
Motivation

Detections can be noisy, due to (self-)occlusions or uncommon views. These errors are hard to recover in later stages.

Images credits:
Motivation

Re-identification techniques have been explored. They are usually built on top of 2D pose estimation networks.

Images credits:
Architecture

**Directly** find 3D people poses from multiple calibrated camera views
2D Backbone

- **Input**: 2D image from single view
- **Output**: 2D features map (512 channels)

- Very fast MobileNet V1 Pretrained on COCO [1]

Lightweight OpenPose [1] references:
Reduction

- Input: 2D features map (512 channels)
- Output: 2D features map (64 channels)

Residual layer with 3 depthwise-separable convs
Unprojection Layer

- **Input:**
  - 2D features maps - for every view
  - Camera parameters (Both Intrinsics and extrinsics)

- **Output:**
  - A single 3D features cube representing the whole scene

- Not learned
- Very fast implementation in GPU (pytorch)
- Lookup table with interpolation
- Differentiable wrt camera params
Volumetric Network

- **Input:**
  - 3D input features

- **Output:**
  - 3D representation of the whole scene. Keypoints + PAFs

- Slightly modified V2V

V2V
Decoding

- **Input**: 3D heatmap + vectormap
- **Output**: List of 3D poses
Datasets

CMU Panoptic [1]
- 30+ HD views
- Hardware-based sync
- Calibration
- 65 sequences (5.5 hours)
- 1.5 millions of 3D skeletons

Shelf [2]
- used to evaluate cross-dataset model generalization
- single scene of four people
- video streams from five calibrated cameras.

Evaluation Metrics

**MPJPE: Mean Per Joint Precision Error**
Average of the square distance of the predicted joints from the corresponding ground-truth joints

**PCP: Percentage of Correctly estimated Parts**

a. Implemented according to [1].
b. A body part is correct if the average distance of the two joints is less than a threshold from the corresponding groundtruth joints locations.
c. The threshold is 50% the length of the groundtruth body part.

Ablation studies

- 3D Augmentations
- Number of volumetric features
- Loss type
  - Different weights on the heatmap / vectormap loss components
- Sub-voxel refinement

<table>
<thead>
<tr>
<th>Cube</th>
<th>Rotation</th>
<th>3D Augmentations</th>
<th>MPJPE (cm)</th>
<th>Head</th>
<th>Torso</th>
<th>Up Arm</th>
<th>Lo Arm</th>
<th>Up Leg</th>
<th>Lo Leg</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td>8.236</td>
<td>99.1</td>
<td>99.3</td>
<td>87.8</td>
<td>96.9</td>
<td>88.3</td>
<td>89.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td>4.598</td>
<td>99.6</td>
<td>99.7</td>
<td>98.5</td>
<td>90.1</td>
<td>99.3</td>
<td>98.5</td>
<td>97.7</td>
</tr>
<tr>
<td></td>
<td>✓ ✓</td>
<td></td>
<td>5.350</td>
<td>99.6</td>
<td>99.7</td>
<td>98.6</td>
<td>91.1</td>
<td>99.0</td>
<td>94.9</td>
<td>97.3</td>
</tr>
<tr>
<td></td>
<td>✓ ✓</td>
<td></td>
<td>3.859</td>
<td>99.7</td>
<td>99.7</td>
<td>99.5</td>
<td>95.6</td>
<td>99.3</td>
<td>98.8</td>
<td>98.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Volumetric Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
</tr>
<tr>
<td>64</td>
</tr>
<tr>
<td>96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Loss Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
</tr>
<tr>
<td>L2</td>
</tr>
<tr>
<td>SmoothL1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Heatmap / Vectormap Loss Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sub-voxel refinement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>✓</td>
</tr>
</tbody>
</table>
Study on the number of input views

- Assessing cross-view generalization
- Good results even with 1 view

Inference time vs Number of views
Sub-modules decomposition
Assessing generalization

Panoptic D2D test set:
- Unseen views
  (new cameras, new angles)
- Unseen scenes

Shelf dataset:
- Completely unseen dataset
- Not yet SOTA results but getting closer
- Probably benefits from variety of pose configurations in training

<table>
<thead>
<tr>
<th>Model</th>
<th>MPHPE (cm)</th>
<th>PCP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>single</td>
<td>multi</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTOR [33] (2 views)*</td>
<td>17.21</td>
<td>50.24</td>
</tr>
<tr>
<td>ACTOR (4 views)*</td>
<td>8.19</td>
<td>20.10</td>
</tr>
<tr>
<td>ACTOR (10 views)*</td>
<td>6.13</td>
<td>12.21</td>
</tr>
<tr>
<td>Oracle [33] (using GT to select cameras)*</td>
<td>4.24</td>
<td>9.19</td>
</tr>
<tr>
<td>Ours (1 unseen view)</td>
<td>10.34</td>
<td>9.32</td>
</tr>
<tr>
<td>Ours (2 to 4 unseen views depending on scene)</td>
<td>5.30</td>
<td>4.09</td>
</tr>
<tr>
<td>Ours (10 views, from training view pool)</td>
<td><strong>3.50</strong></td>
<td><strong>3.56</strong></td>
</tr>
</tbody>
</table>

*ACTOR: number in brackets refers to maximum number of views to choose from. Oracle means: best views to triangulate are selected using groundtruth.
Visual Results - Single View
- Unseen scene
- Unseen camera views
Thank you for your attention

Alessio Elmi  Davide Mazzini  Pietro Tortella

{alessio, davide, pietro}@standard.ai