Towards 3D Human Pose Estimation in the Wild: a Weakly-supervised Approach

Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, Yichen Wei
UT Austin & MSRA & Fudan
Human Pose Estimation

Pose representation

Joint locations

\[ y = \{p_1, \cdots, p_N\} \]
Current Research on 2D Human Pose

- 2D human pose estimation is a well studied problem

Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR 2017
Is 2D human pose all we need?

- Ambiguous 3D structure

Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR 2017
Why we have such a success on 2D?

• 2D human pose data is easy to annotate and largely available

3D data not easy to annotate
Current 3D human pose data.

- Captured in control-environment with accurate sensors.

Catalin Ionescu, Dragos Papava, Vlad Olaru and Cristian Sminchisescu, Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, PAMI 2014
Supervised Pose Regression on Human3.6M

Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, Yichen Wei. Deep Kinematic Pose Regression, In ECCV Workshop on Geometry Meets Deep Learning, 2016
Kinematic Pose Regression-Problems

- Training data is biased to indoor environment

Fail on in-the-wild images!

Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, Yichen Wei. Deep Kinematic Pose Regression, In ECCV Workshop on Geometry Meets Deep Learning, 2016
Problem setting

Given:

In-the-wild images with 2D annotation

Indoor images with 3D annotation

Goal:

In-the-wild image

3D pose
Previous approaches: 2 Stages

Wei et al. Convolutional Pose Machines

Newell et al. Hourglass Network

Bulat et al. Part Heatmap Regression

2D pose estimation

Zhou et al. Shape Convex

Akhter et al. Pose Conditioned Angle Limits

Chen et al. KNN Matching

3D geometry recovery
Previous approaches: 2 Stages

Wei et al. Convolutional Pose Machines

Newell et al. Hourglass Network

Bulat et al. Part Heatmap Regression

The original in-the-wild 2D image, which contains rich cues for 3D pose recovery, is discarded in the second step.

Zhou et al. Shape Convex

Akhter et al. Pose Conditioned Angle Limits

Chen et al. KNN Matching
Our solution: Weakly-supervised Transfer for 3D Human pose estimation in the wild

- Train a unified neural network using both 2D and 3D annotation.
- 2D and 3D pose are inherently entangled
- 2D-to-3D transfer: provide rich image features
- 3D-to-2D transfer: provide 3D annotation
Weakly-supervised Transfer

\[ S_{2D} = \{\mathcal{I}_{2D}, \mathcal{Y}_{2D}\} \quad S_{3D} = \{\mathcal{I}_{3D}, \mathcal{Y}_{2D}, \mathcal{Y}_{dep}\} \]

- Images from both dataset are fed into the same mini-batch
- First estimate 2D pose and then regress depth from 2D results and lower layer image features
- Geometry constraint is applied for weakly-labeled 2D data
Weakly-supervised Transfer

A batch of images from both datasets

\[ S_{2D} = \{I_{2D}, Y_{2D}\} \quad S_{3D} = \{I_{3D}, Y_{2D}, Y_{dep}\} \]

- Images from both datasets are fed into the same mini-batch
- First estimate 2D pose and then regress depth from 2D results and lower layer image features
- Geometry constraint is applied for weakly-labeled 2D data
2D Human Pose estimation: HourglassNetwork

Weakly-supervised Transfer

A batch of images from both datasets

Conv layers → 2D pose estimation module

2D data: constraint

3D data: regression

\[ \mathcal{S}_{2D} = \{ I_{2D}, Y_{2D} \} \quad \mathcal{S}_{3D} = \{ I_{3D}, Y_{2D}, Y_{dep} \} \]

- Images from both dataset are fed into the same mini-batch
- First estimate 2D pose and then regress depth from 2D results and lower layer image features
- Geometry constraint is applied for weakly-labeled 2D data
Geometry Constraint

Key idea: Ratios between bone lengths remain relative fixed

- $R_i$ : a set of involved bones in a skeleton group
- $l_e$ : length of bone $e$
- $\bar{l}_e$ : length of bone $e$ in canonical skeleton

\[
\left\{ \frac{l_e}{\bar{l}_e} \right\}_{e \in R_i} \text{ should remain fixed, i.e. has zero variance}
\]

\[
L_{geo}(\hat{Y}_{dep}|Y_{2D}) = \sum_i \frac{1}{|R_i|} \sum_{e \in R_i} \left( \frac{l_e}{\bar{l}_e} - \bar{r}_i \right)^2,
\]

where

\[
\bar{r}_i = \frac{1}{|R_i|} \sum_{e \in R_i} \frac{l_e}{\bar{l}_e}.
\]
Weakly-supervised Transfer

\[ S_{2D} = \{ \mathcal{I}_{2D}, \mathcal{Y}_{2D} \} \quad S_{3D} = \{ \mathcal{I}_{3D}, \mathcal{Y}_{2D}, \mathcal{Y}_{dep} \} \]

\[ L_{dep}(\hat{Y}_{dep}|I, Y_{2D}) = \begin{cases} 
\lambda_{reg} \| Y_{dep} - \hat{Y}_{dep} \|^2, & \text{if } I \in \mathcal{I}_{3D} \\
\lambda_{geo} L_{geo}(\hat{Y}_{dep}|Y_{2D}), & \text{if } I \in \mathcal{I}_{2D} 
\end{cases} \]

\[ L(\hat{Y}_{HM}, \hat{Y}_{dep}|I) = L_{2D}(\hat{Y}_{HM}, Y_{2D}) + L_{dep}(\hat{Y}_{dep}|I, Y_{2D}) \]
Evaluation-Datasets

• MPII
  • 2D annotation, in-the-wild images
  • Used for weakly-supervised training

• Human 3.6M
  • MoCap 3D annotation, indoor
  • Used for supervised training

• MPI-INF-3DHP
  • MoCap 3D annotation, indoor & outdoor
  • Used for evaluation

• MPII-Validation
  • Used for evaluation
Evaluation-Baseline setup

A batch of images from both datasets

Conv layers

2D pose estimation module

Depth regression module

supervised 2D heatmap regression

summation

skip-connection

2D data: regression

3D data: constraint

(Y_{dep})

Table 2. Definition of our baselines. Transfer for taking both datasets for training, Geometry for the geometry constraint loss.
Supervised 3D pose estimation on Human3.6M dataset

<table>
<thead>
<tr>
<th></th>
<th>Sitting</th>
<th>SittingDown</th>
<th>Smoking</th>
<th>Waiting</th>
<th>WalkDog</th>
<th>Walking</th>
<th>WalkPair</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen &amp; Ramanan [6]</td>
<td>133.14</td>
<td>240.12</td>
<td>106.65</td>
<td>106.21</td>
<td>87.03</td>
<td>114.05</td>
<td>90.55</td>
<td>114.18</td>
</tr>
<tr>
<td>Tome et al. [26]</td>
<td>110.19</td>
<td>172.91</td>
<td>84.95</td>
<td>85.78</td>
<td>86.26</td>
<td>71.36</td>
<td>73.14</td>
<td>88.39</td>
</tr>
<tr>
<td>Zhou et al. [35]</td>
<td>124.52</td>
<td>199.23</td>
<td>107.42</td>
<td>118.09</td>
<td>114.23</td>
<td>79.39</td>
<td>97.70</td>
<td>79.9</td>
</tr>
<tr>
<td>Metha et al. [16]</td>
<td>96.19</td>
<td>122.92</td>
<td>70.82</td>
<td>68.45</td>
<td>54.41</td>
<td>82.03</td>
<td>59.79</td>
<td>74.14</td>
</tr>
<tr>
<td>Pavlakos et al. [20]</td>
<td>76.84</td>
<td><strong>103.48</strong></td>
<td>65.73</td>
<td><strong>61.56</strong></td>
<td>67.55</td>
<td><strong>56.38</strong></td>
<td>59.47</td>
<td><strong>66.92</strong></td>
</tr>
</tbody>
</table>

- **3D/wo geo** (82.44mm) shows the effectiveness of our architecture.
- **3D/w geo** shows the geo-constraint is consistent with supervision.
- Training with 3D&2D data (**3D+2D/wo geo**) provides great performance gain.
- Weakly supervised constraint **3D+2D/w geo** brings further improvements.
- Only 2-steps methods Chen & Ramanan(114.18mm) and Zhou et al,(79.9mm) can be applied in-the-wild.
Results Analysis

<table>
<thead>
<tr>
<th></th>
<th>Sitting</th>
<th>SittingDown</th>
<th>Smoking</th>
<th>Waiting</th>
<th>WalkDog</th>
<th>Walking</th>
<th>WalkPair</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen &amp; Ramanan [6]</td>
<td>133.14</td>
<td>240.12</td>
<td>106.65</td>
<td>106.21</td>
<td>87.03</td>
<td>114.05</td>
<td>90.55</td>
<td>114.18</td>
</tr>
<tr>
<td>Tome et al. [26]</td>
<td>110.19</td>
<td>172.91</td>
<td>84.95</td>
<td>85.78</td>
<td>86.26</td>
<td>71.36</td>
<td>73.14</td>
<td>88.39</td>
</tr>
<tr>
<td>Zhou et al. [35]</td>
<td>124.52</td>
<td>199.23</td>
<td>107.42</td>
<td>118.09</td>
<td>114.23</td>
<td>79.39</td>
<td>97.70</td>
<td>79.9</td>
</tr>
<tr>
<td>Metha et al. [16]</td>
<td>96.19</td>
<td>122.92</td>
<td>70.82</td>
<td>68.45</td>
<td>54.41</td>
<td>82.03</td>
<td>59.79</td>
<td>74.14</td>
</tr>
<tr>
<td>Pavlakos et al. [20]</td>
<td>76.84</td>
<td>103.48</td>
<td>65.73</td>
<td>61.56</td>
<td>67.55</td>
<td>56.38</td>
<td>59.47</td>
<td>66.92</td>
</tr>
</tbody>
</table>

| 3D/wo geo              | 98.41   | 141.60      | 80.01   | 86.31   | 61.89   | 76.32   | 71.47   | 82.44   |
| 3D/w geo               | 93.52   | 131.75      | 79.61   | 85.10   | 67.49   | 76.95   | 71.99   | 80.98   |
| 3D+2D/wo geo          | **74.79** | **113.99** | **64.34** | **68.78** | **52.22** | **63.97** | **57.31** | **65.69** |
| 3D+2D/w geo           | 75.20   | 111.59      | **64.15** | **66.05** | **51.43** | **63.22** | **55.33** | **64.90** |

- Is the improvement from more accurate 2D position or better depth estimation?
- All baselines have very high 2D pose estimation.
- This indicates that depth estimation are greatly benefit from more 2D data.
- 2-stage approaches can not have such benefit.
### In-the-wild 3D pose estimation on MPII-INF-3DHP Dataset

![Images showing in-the-wild 3D pose estimation results](image_url)

<table>
<thead>
<tr>
<th>Method</th>
<th>Stand/Walk</th>
<th>Exercise</th>
<th>Chair</th>
<th>Reach</th>
<th>Ground</th>
<th>Sport</th>
<th>Misc</th>
<th>Total PCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metha et al. (H36M+MPII) [16]</td>
<td>76.4</td>
<td>62.9</td>
<td>58.1</td>
<td>57.4</td>
<td>27.8</td>
<td>66.9</td>
<td>65.6</td>
<td>61.0</td>
<td>28.3</td>
</tr>
<tr>
<td>3D/wo geo</td>
<td>28.6</td>
<td>41.2</td>
<td>41.4</td>
<td>34.3</td>
<td>19.7</td>
<td>36.4</td>
<td>36.4</td>
<td>31.5</td>
<td>18.0</td>
</tr>
<tr>
<td>3D/w geo</td>
<td>37.0</td>
<td>44.5</td>
<td>45.4</td>
<td>38.8</td>
<td>22.9</td>
<td>50.1</td>
<td>30.8</td>
<td>37.7</td>
<td>20.9</td>
</tr>
<tr>
<td>3D+2D/wo geo</td>
<td>82.3</td>
<td>66.4</td>
<td>60.3</td>
<td>69.2</td>
<td>37.1</td>
<td>65.7</td>
<td>67.8</td>
<td>65.8</td>
<td>32.1</td>
</tr>
<tr>
<td>3D+2D/w geo</td>
<td>85.4</td>
<td>71.0</td>
<td>60.7</td>
<td>71.4</td>
<td>37.8</td>
<td>70.9</td>
<td>74.4</td>
<td><strong>69.2</strong></td>
<td><strong>32.5</strong></td>
</tr>
<tr>
<td>Metha et al. (MPI-INF-3DHP) [16]</td>
<td><strong>85.0</strong></td>
<td><strong>70.1</strong></td>
<td><strong>72.7</strong></td>
<td><strong>65.2</strong></td>
<td><strong>47.0</strong></td>
<td><strong>79.0</strong></td>
<td><strong>70.3</strong></td>
<td><strong>70.8</strong></td>
<td><strong>35.9</strong></td>
</tr>
</tbody>
</table>

**Table 2.** Results of MPI-INF-3DHP Dataset. The results are shown in PCK and AUC.

- **3D data-only methods fail on in-the-wild images.**
- **3D+2D/wo geo** wins its counterpart of Metha et al.
- Geo-constraint provides further improvements, whose results are close to training on the corresponding training set.
In-the-wild 3D pose estimation on MPII-Validation-3D Set

• **3D+2D/w geo** performs better and correct the symmetry invalidity.
• Our framework keeps 2D accuracy.

<table>
<thead>
<tr>
<th></th>
<th>3D+2D/wo geo</th>
<th>3D+2D/w geo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper arm</td>
<td>42.4mm</td>
<td>37.8mm</td>
</tr>
<tr>
<td>Lower arm</td>
<td>60.4mm</td>
<td>50.7mm</td>
</tr>
<tr>
<td>Upper leg</td>
<td>43.5mm</td>
<td>43.4mm</td>
</tr>
<tr>
<td>Lower leg</td>
<td>59.4mm</td>
<td>47.8mm</td>
</tr>
<tr>
<td>Upper arm</td>
<td>6.27px</td>
<td>4.80px</td>
</tr>
<tr>
<td>Lower arm</td>
<td>10.11px</td>
<td>6.64px</td>
</tr>
<tr>
<td>Upper leg</td>
<td>6.89px</td>
<td>4.93px</td>
</tr>
<tr>
<td>Lower leg</td>
<td>8.03px</td>
<td>6.22px</td>
</tr>
</tbody>
</table>
More qualitative results
Failure Cases

inaccurate 2D prediction/ ambiguous depth/ false torso length.
Extension

- An improved weak-supervision for rigid objects.
- The predicted pose of the same object from different viewpoint should be consistent with each other.

Xingyi Zhou, Arjun Karpur, Chuang Gan, Linjie Luo, Qixing Huang, Unsupervised Domain Adaptation for 3D Keypoint Prediction from a Single Depth Scan, arXiv 1712.05765, 2017
Extension

- Add temporal refinement.
- Add angle constraint.

Demo
Q & A

Code & Model Available!

Torch

PyTorch