Orthographic Projection Linear Regression for Single Image 3D Human Pose Estimation

Yahui Zhang
Shaodi You
Theo Gevers
Introduction

Motivation

Current 3D human pose datasets are collected in indoor environment, limiting the generalization of learning-based approaches for 3D human pose estimation.

- Challenge
  - 2D in-the-wild images are extremely complex.
  - In-the-wild images do not have corresponding 3D ground truth.

Fig. 1. Indoor image and corresponding 3D ground truth\(^1\)

Fig. 2. In-the-wild image with no 3D ground truth\(^2\)

Introduction

Goal

The goal of this paper is to regress 3D human joint locations in camera coordinates from a single image.

Solution

Propose an orthographic projection linear regression module.

Fig. 3. General procedure of our method to connect in-the-wild images and 3D predictions.
Our Approach

Definition

• Human pose representations: a set of joints, $p_{3D}^{abs} = [J_1^{abs}, J_2^{abs}, ..., J_n^{abs}]$, where $J_i^{abs} = [X_i^{abs}, Y_i^{abs}, Z_i^{abs}, 1]^T$

• 2D projections $p_{2D}$, a 3 by $n$ matrix with $J_i^{abs} = [x_i^{abs}, y_i^{abs}, 1]^T$

Camera Model

Given intrinsic ($K$) and extrinsic ($R$ and $T$) parameters, 2D projections are obtained by:

$$p_{2D} = K[R|T]p_{3D}^{abs} \quad (1)$$

Small angle problem arises:
resulting in overfitting in the depth dimension.

Fig. 4. Perspective projection from (a) the 3D pose to (b) the 2D pose with illustration of small angle problem (c).
Our Approach

**Orthographic Projection Linear Regression**

*Step 1: Orthographic Projection.*

$$p_{2D} = \Pi p_{3D},$$

$$\Pi = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad (2)$$

where $p_{3D}$ is root-relative 3D joint locations.

*Step 2: Constrained Linear Regression.*

$$p_{2D} = [S|t]\Pi p_{3D}, \quad (3)$$

where $S$ and $t$ indicate scale and translation parameters.

**Optimization**

The linear regression computes the scaling and translation by minimizing:

$$\arg \min_{S,t} \|[S|t]\Pi p_{3D} - p_{2D}\|^2_2. \quad (4)$$

Fig. 5. The general idea of matching 3D with 2D poses by the orthographic projection linear regression method.
Our Approach

- **Architecture**

![Diagram of the proposed framework](image)

Fig. 6. The overview of the proposed framework.

- **Loss Function**

\[
\mathcal{L}_{\text{pose}} = \lambda_{hm}\mathcal{L}_{\text{Heatmap}} + \mathcal{L}_{3D} + \lambda_{OPLR}\mathcal{L}_{OPLR} \tag{5}
\]

Specifically, 
\[
\mathcal{L}_{\text{Heatmap}} = \|\text{HM} - \text{HM}^{GT}\|_2, \quad \mathcal{L}_{3D} = \|\mathbf{P}_{3D} - \mathbf{P}_{3D}^{GT}\|_2, \quad \mathcal{L}_{OPLR} = \|S[t]\Pi\mathbf{P}_{3D} - \mathbf{p}_{2D}^{GT}\|_2
\]

where \(\text{HM}\) denotes heatmap.
## Experiments

### Datasets

Current public datasets: Human3.6m\(^1\) and MPI-INF-3DHP\(^2\)

#### Metric

- **Human3.6m**
  - Protocol #1: Mean Per Joint Position Error (MPJPE).
  - Protocol #2: Mean Per Joint Position Error after a rigid transformation (PA MPJPE).

  *The smaller, the better.*

- **MPI-INF-3DHP**
  - Percentage of Correct Keypoints (PCK). The threshold is set to 150\(mm\).
  - Aera under the Curve (AUC)

  *The larger, the better.*

---

Experiments

- Evaluation on Human3.6m

Table 1. The quantitative results compared to state-of-the-art 3D human pose estimation methods on Human3.6m.

**Table 1.** The quantitative results compared to state-of-the-art 3D human pose estimation methods on Human3.6m.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhou et al. (CVPR’16) [41]</td>
<td>87.4</td>
<td>109.3</td>
<td>87.1</td>
<td>103.2</td>
<td>116.2</td>
<td>143.3</td>
<td>106.9</td>
<td>99.8</td>
<td>124.5</td>
<td>199.2</td>
<td>107.4</td>
<td>118.1</td>
<td>114.2</td>
<td>79.4</td>
<td>97.7</td>
<td>113.0</td>
</tr>
<tr>
<td>Chen et al. (CVPR’17) [21]</td>
<td>89.9</td>
<td>97.6</td>
<td>90.0</td>
<td>107.9</td>
<td>107.3</td>
<td>93.6</td>
<td>136.1</td>
<td>133.1</td>
<td>240.1</td>
<td>106.7</td>
<td>139.2</td>
<td>106.2</td>
<td>87.0</td>
<td>114.1</td>
<td>90.6</td>
<td>114.2</td>
</tr>
<tr>
<td>Pavlakos et al. (CVPR’17) [13]</td>
<td>67.4</td>
<td>71.9</td>
<td>66.7</td>
<td>69.1</td>
<td>72.0</td>
<td>77.0</td>
<td>65.0</td>
<td>68.3</td>
<td>83.7</td>
<td>96.5</td>
<td>71.7</td>
<td>65.8</td>
<td>74.9</td>
<td>59.1</td>
<td>63.2</td>
<td>71.9</td>
</tr>
<tr>
<td>Mehta et al. (3DV’17) [42]</td>
<td>57.5</td>
<td>68.6</td>
<td>59.6</td>
<td>67.3</td>
<td>78.1</td>
<td>56.9</td>
<td>69.1</td>
<td>98.0</td>
<td>117.5</td>
<td>69.5</td>
<td>82.4</td>
<td>68.0</td>
<td>55.3</td>
<td>76.5</td>
<td>61.4</td>
<td>72.9</td>
</tr>
<tr>
<td>Zhou et al. (ICCV’17) [28]</td>
<td>54.8</td>
<td>60.7</td>
<td>58.2</td>
<td>71.4</td>
<td>62.0</td>
<td>65.5</td>
<td>53.8</td>
<td>55.6</td>
<td>75.2</td>
<td>111.6</td>
<td>64.1</td>
<td>66.0</td>
<td>51.4</td>
<td>63.2</td>
<td>55.3</td>
<td>64.9</td>
</tr>
<tr>
<td>Sun et al. (ICCV’17) [39]</td>
<td>52.8</td>
<td>54.8</td>
<td>54.2</td>
<td>54.3</td>
<td>61.8</td>
<td>67.2</td>
<td>53.1</td>
<td>53.6</td>
<td>71.7</td>
<td>86.7</td>
<td>61.5</td>
<td>53.4</td>
<td>61.6</td>
<td>47.1</td>
<td>53.4</td>
<td>59.1</td>
</tr>
<tr>
<td>Luo et al. (BMVC’18) [43]</td>
<td>53.5</td>
<td>60.9</td>
<td>56.3</td>
<td>59.1</td>
<td>64.3</td>
<td>74.4</td>
<td>55.4</td>
<td>63.4</td>
<td>74.8</td>
<td>98.0</td>
<td>61.1</td>
<td>58.2</td>
<td>70.6</td>
<td>49.1</td>
<td>55.7</td>
<td>63.7</td>
</tr>
<tr>
<td>Yang et al. (CVPR’18) [30]</td>
<td>51.5</td>
<td>58.9</td>
<td>50.4</td>
<td>57.0</td>
<td>62.1</td>
<td>65.4</td>
<td>49.8</td>
<td>52.7</td>
<td>69.2</td>
<td>85.2</td>
<td>57.4</td>
<td>58.4</td>
<td>43.6</td>
<td>60.1</td>
<td>47.7</td>
<td>58.6</td>
</tr>
<tr>
<td>Zhao et al. (CVPR’19) [38]</td>
<td>47.3</td>
<td>60.7</td>
<td>51.4</td>
<td>60.5</td>
<td>61.1</td>
<td>49.9</td>
<td>47.3</td>
<td>68.1</td>
<td>86.2</td>
<td>55.0</td>
<td>67.8</td>
<td>61.0</td>
<td>42.1</td>
<td>60.6</td>
<td>45.3</td>
<td>57.6</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>46.0</strong></td>
<td><strong>55.3</strong></td>
<td><strong>50.6</strong></td>
<td><strong>53.5</strong></td>
<td><strong>57.5</strong></td>
<td><strong>46.3</strong></td>
<td><strong>49.4</strong></td>
<td><strong>71.7</strong></td>
<td><strong>87.9</strong></td>
<td><strong>56.6</strong></td>
<td><strong>68.4</strong></td>
<td><strong>53.5</strong></td>
<td><strong>41.4</strong></td>
<td><strong>57.9</strong></td>
<td><strong>46.6</strong></td>
<td><strong>56.2</strong></td>
</tr>
</tbody>
</table>

**Table 1.** The quantitative results compared to state-of-the-art 3D human pose estimation methods on Human3.6m.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Moreno-Noguer (CVPR’17) [22]</td>
<td>66.1</td>
<td>61.7</td>
<td>84.5</td>
<td>73.7</td>
<td>65.2</td>
<td>67.2</td>
<td>60.9</td>
<td>67.3</td>
<td>103.5</td>
<td>74.6</td>
<td>92.6</td>
<td>69.6</td>
<td>71.5</td>
<td>78.0</td>
<td>73.2</td>
<td>74.0</td>
</tr>
<tr>
<td>Sun et al. (ICCV’17) [39]</td>
<td>42.1</td>
<td>44.3</td>
<td>45.0</td>
<td>45.4</td>
<td>51.5</td>
<td>53.0</td>
<td>43.2</td>
<td>41.3</td>
<td>59.3</td>
<td>73.3</td>
<td>51.0</td>
<td>44.0</td>
<td>48.0</td>
<td>38.3</td>
<td>44.8</td>
<td>48.3</td>
</tr>
<tr>
<td>Luo et al. (BMVC’18) [43]</td>
<td>40.8</td>
<td>44.6</td>
<td>42.1</td>
<td>45.1</td>
<td>48.3</td>
<td>54.6</td>
<td>44.2</td>
<td>42.9</td>
<td>85.3</td>
<td>69.9</td>
<td>46.7</td>
<td>42.5</td>
<td>48.0</td>
<td>38.0</td>
<td>41.4</td>
<td>46.6</td>
</tr>
<tr>
<td>Yang et al. (CVPR’18) [30]</td>
<td>26.9</td>
<td>30.9</td>
<td>36.3</td>
<td>39.9</td>
<td>43.9</td>
<td>47.4</td>
<td>28.8</td>
<td>29.4</td>
<td>36.9</td>
<td>58.4</td>
<td>41.5</td>
<td>40.5</td>
<td>29.5</td>
<td>42.5</td>
<td>32.2</td>
<td>37.7</td>
</tr>
<tr>
<td>Zhou et al. (TPAMI’18) [44]</td>
<td>47.9</td>
<td>48.8</td>
<td>52.7</td>
<td>55.0</td>
<td>56.8</td>
<td>65.5</td>
<td>49.0</td>
<td>45.5</td>
<td>60.8</td>
<td>81.1</td>
<td>53.7</td>
<td>51.6</td>
<td>54.8</td>
<td>50.4</td>
<td>55.9</td>
<td>55.3</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>35.8</strong></td>
<td><strong>41.0</strong></td>
<td><strong>42.3</strong></td>
<td><strong>42.0</strong></td>
<td><strong>43.4</strong></td>
<td><strong>36.3</strong></td>
<td><strong>36.7</strong></td>
<td><strong>55.1</strong></td>
<td><strong>66.5</strong></td>
<td><strong>45.0</strong></td>
<td><strong>49.6</strong></td>
<td><strong>41.2</strong></td>
<td><strong>32.9</strong></td>
<td><strong>43.9</strong></td>
<td><strong>39.0</strong></td>
<td><strong>43.4</strong></td>
</tr>
</tbody>
</table>

Our method achieves
1) best performance in Protocol #1.
2) better performance than most of existing methods in Protocol #2.
Experiments

Evaluation on MPI-INF-3DHP

Table 2. The quantitative results on MPI-INF-3DHP.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Extra information</th>
<th>PCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mehta et al. (3DV’17) [42]</td>
<td></td>
<td>64.7</td>
<td>31.7</td>
</tr>
<tr>
<td>Zhou et al. (ICCV’17) [28]</td>
<td>Post-processing</td>
<td>68.2</td>
<td>32.5</td>
</tr>
<tr>
<td>Yang et al. (CVPR’18) [30]</td>
<td></td>
<td>69.0</td>
<td>32.0</td>
</tr>
<tr>
<td>Habibie et al. (CVPR’19) [7]</td>
<td>Extra training set</td>
<td>70.4</td>
<td>36.0</td>
</tr>
<tr>
<td>Want et al. (CVPR’19) [8]</td>
<td>Extra training set</td>
<td>81.8</td>
<td>54.8</td>
</tr>
<tr>
<td>Ci et al. (ICCV’19) [50]</td>
<td>2D Pose</td>
<td>74.0</td>
<td>34.7</td>
</tr>
<tr>
<td>Ours (w/o $L_{OPLR}$)</td>
<td></td>
<td>23.9</td>
<td>8.9</td>
</tr>
<tr>
<td>Ours (full)</td>
<td></td>
<td>66.8</td>
<td>31.9</td>
</tr>
</tbody>
</table>

Table 3. The quantitative evaluation with using rigid transformation.

<table>
<thead>
<tr>
<th>Methods (Using rigid transformation)</th>
<th>Extra information</th>
<th>PCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habibie et al. (CVPR’19) [7]</td>
<td>Extra training set</td>
<td>82.9</td>
<td>45.4</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>84.4</td>
<td>46.9</td>
</tr>
</tbody>
</table>

- Our method achieves superior performance even without using extra information.
- Our method outperforms the existing method with using rigid transformation for evaluation.
- Our method significantly performs better than Ours (w/o $L_{OPLR}$) with an improvement from 23.9% to 66.8%.
Fig. 7. The qualitative results on MPII and LSP dataset generated by the proposed method.
We propose a novel orthographic projection and linear regression to constrain the 3D and 2D poses.

A network is proposed which is adaptive to various in-the-wild images without retraining the 3D pose.

Our network achieves state-of-the-art performance on the Human3.6m dataset and generalizes well to in-the-wild datasets.