Learning 3D Human Pose Estimation

Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision

Seminar on Current Works in Computer Vision

Presenter: Ikrima Bin Saeed

July 26, 2017
The Focus of My Presentation

- Introduction
- Common Approaches
- Current Approach
  - Bounding Box
  - 3D Pose Prediction
  - Global 3D Pose Computation
- Data Set Contribution
- Evaluation
  - Perspective Correction
  - Multi-level Corrective Skip Connections
  - New Data Set
  - Transfer Learning
  - Multi-modal Prediction
  - State of the Art
- Summary
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
3D Human Pose Estimation Using a Single Image
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
Common Approaches

- Fit 3D models to 2D key points
  - Computationally expensive
  - Unstable under depth ambiguities

- Direct CNN-based 3D regression
  - Performance far below than that of CNN based methods for 2D human pose
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
Approach Overview

- 3 main steps from 2D input image to global 3D human pose
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
Bounding Box Computation

Input image → 2D keypoints $K$
Bounding Box Computation

Input Image

Input image source: ESPN Cricinfo
Bounding Box Computation

Input Image $I$

Heat Map $H$

Input image source: ESPN Cricinfo
Bounding Box Computation

Input Image \( I \)  
Heat Map \( H \)  
2D Joint Locations \( K \)

Input image source: ESPN Cricinfo
Bounding Box Computation

Input Image $I$

Heat Map $H$

2D Joint Locations $K$

Bounding Box $BB$

Input image source: ESPN Cricinfo
2DposeNet Architecture

- Training data: MPII + LSP
- Architecture based on Resnet-101
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
(1) Bounding Box Computation

Input image $\rightarrow$ 2D keypoints $K$ $\rightarrow$ Bounding Box $BB$ $\frac{K}{K}$

(2) 3D Pose Prediction

Cropped input $\rightarrow$ 3D PoseNet $\rightarrow$ 3D pose $P$ $\frac{P}{K}$

(3) Global 3D Pose Computation

3D-2D alignment $\rightarrow$ Global 3D pose $P[G]$
3D Pose Regression

Cropped input → 3D pose $P$
Main Components:
- CNN based on ResNet-101 with Transfer Learning
- Multi-level Corrective Skip Connections
- Multi-modal Prediction
- 3D Pose Fusion
- 2D Pose Estimation
Main Components:

- CNN based on ResNet-101 with Transfer Learning
- Multi-level Corrective Skip Connections
- Multi-modal Prediction
- 3D Pose Fusion
- 2D Pose Estimation
Transfer the knowledge acquired from one task to help solve a new task.

**Traditional ML**

- Task / domain A
- Task / domain B

- Training and evaluation on the same task or domain.

**Transfer learning**

- Source task / domain
- Target task / domain

- Storing knowledge gained solving one problem and applying it to a different but related problem.

Image Source: http://ruder.io/transfer-learning/
Transfer Learning in Our Case

ResNet-101

ImageNet

2DPoseNet

3DPoseNet

Multimodal Prediction

3D Pose Fusion

2D Auxiliary Task
Transfer Learning: Train ResNet-101

ResNet-101

ImageNet

2D PoseNet

3D PoseNet

3D Dataset

Transfer Learning

Corrective Skip Connections

Multimodal Prediction

3D Pose Fusion

2D Auxiliary Task
Transfer Learning: Identify Similar Structure

ResNet-101

ImageNet

2D PoseNet

3D PoseNet
Transfer Learning: Weight Initialization Step 1

ResNet-101

ImageNet

2DPoseNet

3DPoseNet
Transfer Learning: Train 2DposeNet

ImageNet

ResNet-101

2DposeNet

3DposeNet

3D Dataset

Transfer Learning

Multimodal Prediction

3D Pose Fusion

Corrective Skip Connections

2D Auxiliary Task
Transfer Learning: Weight Initialization Step 2

ResNet-101

ImageNet

2DPoseNet

3DPoseNet

3D Dataset

Transfer Learning

Corrective Skip Connections

Multimodal Prediction

3D Pose Fusion

2D Auxiliary Task
Transfer Learning: Train 3D PoseNet

ImageNet

ResNet-101

2D PoseNet

3D PoseNet

Multimodal Prediction

3D Pose Fusion

2D Auxiliary Task

Corrective Skip Connections

University of Freiburg
Seminar on Current Works in CV
28
Main Components:

- CNN based on ResNet-101 with Transfer Learning
- Multi-level Corrective Skip Connections
- Multi-modal Prediction
- 3D Pose Fusion
- 2D Pose Estimation
Multi-level Corrective Skip Connections

- Used only during training
- Key insights
  - Information flow from different levels help in regression problems
  - Deepest level should contribute more than shallower ones
Multi-level Corrective Skip Connections

- The insight “**Deepest level should contribute more than shallower ones**” is not satisfied if we use regular skip connections.
Multi-level Corrective Skip Connections

- The insight “**Deepest level should contribute more than shallower ones**” is not satisfied if we use regular skip connections.

- By using multi-level skip connections, the loss is given as:
  \[ P_{sum} = P_{deep} + \text{loss contribution from shallower terms} \]

- This structure enforces the above insight.
3D PoseNet Architecture

Main Components:
- CNN based on ResNet-101 with Transfer Learning
- Multi-level Corrective Skip Connections
- Multi-modal Prediction
- 3D Pose Fusion
- 2D Pose Estimation
Representation of Joints
Representation of a Pose

- Select a joint
- Represent other joints relative to this one
• Select a joint
• Represent other joints relative to this one

**Problem!** This representation is not always optimal
S. Li et al. suggested representation using first order parents in kinematic structure improves performance.
• S. Li et al. suggested representation using first order parents in kinematic structure improves performance

• Authors of this paper found out that this is not universally true. Optimal order relationship in kinematic structure depends on pose and visibility.
**Different Representations of a Pose**

- **P:** joints relative to root (pelvis here)
- **O1:** joints relative to first order kinematics
- **O2:** joints relative to second order kinematics

*some joint relations omitted for clarity*
Multi-modal Prediction
3D PoseNet Architecture

Main Components:
- CNN based on ResNet-101 with Transfer Learning
- Multi-level Corrective Skip Connections
- Multi-modal Prediction
- 3D Pose Fusion
- 2D Pose Estimation
3D Pose Fusion

Now better constraints are available per joint
Network automatically finds the optimal combination of different pose representations
Main Components:
• CNN based on ResNet-101 with Transfer Learning
• Multi-level Corrective Skip Connections
• Multi-modal Prediction
• 3D Pose Fusion
• 2D Pose Estimation
Accomplishing an additional task without any significant overheads
Network Architecture

MPII+LSP

2D PoseNet

3D Dataset

Transfer Learning

Multimodal Prediction

3D Pose Fusion

3D PoseNet

Corrective Skip Connections

2D Auxiliary Task
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
Approach Overview

(1) Bounding Box Computation
- Input image
- 2D keypoints $K$

(2) 3D Pose Prediction
- Cropped input
- 3D pose $P$

(3) Global 3D Pose Computation
- 3D-2D alignment
- Global 3D pose $P^{[G]}$
Global 3D Pose Computation

- From pelvis-centered pose $P_{fused}$, we need to get a global 3D pose $P^{[G]}$
Global 3D Pose Computation

- From pelvis-centered pose $P_{fused}$, we need to get a global 3D pose $P^{[G]}$
- $P_{fused}$ is obtained from the normalized image that we get from the bounding box BB
From pelvis-centered pose $P_{fused}$, we need to get a global 3D pose $P^{[G]}$

$P_{fused}$ is obtained from the normalized image that we get from the bounding box BB

If we know the joint locations $K$ in the original image, we can know the BB and thus know the normalization parameters used for $P_{fused}$
If we imagine a virtual camera that only sees the cropped part of the image, then $P_{fused}$ can be viewed as the pose inside the camera frame
Now, we need to find $\mathbf{R}$ and $\mathbf{T}$ such the view frame of this virtual camera equal the cropped image, then we can get the global 3D pose $\mathbf{P}^{[G]}$

$$\mathbf{P}^{[G]} = (\mathbf{R} \mid \mathbf{T}) \mathbf{P}_{\text{fused}}$$
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
Another Contribution of this Paper: New Dataset

Stand/Walk

Exercise

On the Ground

Crouch/Reach

Sports

Sit On a Chair

Misc.
### Existing 3D Pose Datasets

<table>
<thead>
<tr>
<th>Marker-based Motion Capture</th>
<th>Marker-less Motion Capture</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Relatively cheaper with fewer cameras</td>
<td>• Hundreds of cameras are required</td>
</tr>
<tr>
<td>• Requires skin-tight clothing to capture 3D data</td>
<td>• Requires very controlled environment</td>
</tr>
<tr>
<td></td>
<td>• Diverse clothing available</td>
</tr>
</tbody>
</table>

A. Baak et al. showed that motion tracking techniques do not work in general scenes
Salient Features

- Multi-camera studio
- Similar to the marker-less motion capture
- Works on everyday apparel in contrast to skin tight clothing
- 8 actors (4m+4f)
- Covered more pose classes than Human3.6M
- Different viewpoints and different elevations
- > 1.3M frames
- Augmentation
- Diverse test set
Salient Features

- Multi-camera studio
- Similar to the marker-less motion capture
- Works on everyday apparel in contrast to skin tight clothing
- 8 actors (4m+4f)
- Covered more pose classes than Human3.6M
- Different viewpoints and different elevations
- > 1.3M frames
  - Augmentation
- Diverse test set
Data Augmentation
## MPI-INF-3DHP Dataset

### Salient Features

- Multi-camera studio
- Similar to the marker-less motion capture
- Works on everyday apparel in contrast to skin tight clothing
- 8 actors (4m+4f)
- Covered more pose classes than Human3.6M
- Different viewpoints and different elevations
- > 1.3M frames
- Augmentation
  - Diverse test set
Diverse Test Set

• Test sets of Human3.6M and HumanEva are recorded indoors

• Test set of MPI-INF-3DHP is more general due to
  • More diverse motions
  • Greater variation in camera viewpoint
  • Outdoor scenes
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
## Metrics for Evaluation

- Mean Per Joint Position Error (MPJPE)
- Percentage of Correct Keypoints (PCK)
- Area Under the Curve (AUC)
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
Evaluating Perspective Correction

- On MPI-INF-3DHP, performance improves by 3 PCK
- On HumanEva, error is decreases by 3 mm MPJPE
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
Evaluating Multi-level Corrective Skip Connections

Mean Per Joint Position Error (MPJPE) in mm on Human3.6M Test Data

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>Discuss</th>
<th>Eating</th>
<th>Greet</th>
<th>Phone</th>
<th>Posing</th>
<th>Purch.</th>
<th>Sitting</th>
<th>Down</th>
<th>Smoke</th>
<th>Take Photo</th>
<th>Wait</th>
<th>Walk</th>
<th>Walk Dog</th>
<th>Walk Pair</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>98.98</td>
<td>100.14</td>
<td>86.07</td>
<td>101.83</td>
<td>101.34</td>
<td>96.74</td>
<td>94.89</td>
<td>125.28</td>
<td>158.31</td>
<td>100.21</td>
<td>112.49</td>
<td>99.57</td>
<td>83.39</td>
<td>109.61</td>
<td>95.79</td>
<td>104.32</td>
</tr>
<tr>
<td>Regular Skip</td>
<td>113.34</td>
<td>112.26</td>
<td>97.40</td>
<td>110.50</td>
<td>108.63</td>
<td>112.09</td>
<td>105.67</td>
<td>125.97</td>
<td>173.41</td>
<td>109.34</td>
<td>120.87</td>
<td>107.75</td>
<td>97.30</td>
<td>126.05</td>
<td>117.45</td>
<td>115.29</td>
</tr>
<tr>
<td>Corr. Skip + Fusion</td>
<td>92.57</td>
<td>99.08</td>
<td>85.46</td>
<td>96.93</td>
<td>95.60</td>
<td>89.56</td>
<td>95.67</td>
<td>123.54</td>
<td>160.98</td>
<td>97.13</td>
<td>107.56</td>
<td>93.86</td>
<td>76.99</td>
<td>110.93</td>
<td>88.73</td>
<td>101.09</td>
</tr>
<tr>
<td>Transfer Learning from 2DPoseNet weights</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr. + Fusion</td>
<td>59.69</td>
<td>69.74</td>
<td>60.55</td>
<td>68.77</td>
<td>76.36</td>
<td>59.05</td>
<td>75.04</td>
<td>96.19</td>
<td>122.92</td>
<td>70.82</td>
<td>85.42</td>
<td>68.45</td>
<td>54.41</td>
<td>82.03</td>
<td>59.79</td>
<td>74.14</td>
</tr>
</tbody>
</table>

Key Observations:
- Error increases if we use regular skip connections
- Error decreases if we use multi level skip connections
Evaluating Multi-level Corrective Skip Connections

Percentage of Correct Keypoints (PCK) and Area Under the Curve (AUC) on MPI-INF-3DHP Test Data. GS = green screen background

<table>
<thead>
<tr>
<th>3D Data</th>
<th>Network Arch.</th>
<th>Studio GS 3DPCK</th>
<th>Studio no GS 3DPCK</th>
<th>Outdoor 3DPCK</th>
<th>All 3DPCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human 3.6m</td>
<td>Base</td>
<td>21.1</td>
<td>32.5</td>
<td>10.8</td>
<td>22.6</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>Corr Skip +Fusion</td>
<td>22.2</td>
<td>33.9</td>
<td>18.5</td>
<td>25.1</td>
<td>8.7</td>
</tr>
<tr>
<td>Ours Unaug.</td>
<td>Base</td>
<td>73.6</td>
<td>42.9</td>
<td>19.5</td>
<td>49.0</td>
<td>23.3</td>
</tr>
<tr>
<td></td>
<td>Corr Skip +Fusion</td>
<td>66.9</td>
<td>38.2</td>
<td>27.9</td>
<td>46.8</td>
<td>20.9</td>
</tr>
<tr>
<td>Ours Aug.</td>
<td>Base</td>
<td>77.2</td>
<td>59.5</td>
<td>48.7</td>
<td>63.7</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td>Corr. Skip +Fusion</td>
<td>71.1</td>
<td>51.7</td>
<td>36.1</td>
<td>55.4</td>
<td>26.0</td>
</tr>
</tbody>
</table>

Key Observations:
- Nearly 8% 3DPCK improvement for the outdoor scenes
- Multi-level skip connections improve generality
- Performance decreases for augmented MPI-INF-3DHP training data
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
Evaluating New Data Set

Percentage of Correct Keypoints (PCK) and Area Under the Curve (AUC) on MPI-INF-3DHP Test Data. GS = green screen background

<table>
<thead>
<tr>
<th>3D Data</th>
<th>Network Arch.</th>
<th>Studio GS 3DPCK</th>
<th>Studio no GS 3DPCK</th>
<th>Outdoor 3DPCK</th>
<th>All 3DPCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human 3.6m</td>
<td>Base</td>
<td>21.1</td>
<td>32.5</td>
<td>10.8</td>
<td>22.6</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>Corr Skip +Fusion</td>
<td>22.2</td>
<td>33.9</td>
<td>18.5</td>
<td>25.1</td>
<td>8.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22.3</td>
<td>34.2</td>
<td>20.0</td>
<td>26.0</td>
<td>9.5</td>
</tr>
<tr>
<td>Ours Unaug.</td>
<td>Base</td>
<td>73.6</td>
<td>42.9</td>
<td>19.5</td>
<td>49.0</td>
<td>23.3</td>
</tr>
<tr>
<td></td>
<td>Corr Skip +Fusion</td>
<td>66.9</td>
<td>38.2</td>
<td>27.9</td>
<td>46.8</td>
<td>20.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>67.6</td>
<td>39.6</td>
<td>28.5</td>
<td>47.8</td>
<td>21.8</td>
</tr>
<tr>
<td>Ours Aug.</td>
<td>Base</td>
<td>77.2</td>
<td>59.5</td>
<td>48.7</td>
<td>63.7</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td>Corr. Skip +Fusion</td>
<td>71.1</td>
<td>51.7</td>
<td>36.1</td>
<td>55.4</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>73.5</td>
<td>53.1</td>
<td>37.9</td>
<td>57.3</td>
<td>28.0</td>
</tr>
</tbody>
</table>
### Evaluating New Data Set

Percentage of Correct Keypoints (PCK) and Area Under the Curve (AUC) on MPI-INF-3DHP Test Data. GS = green screen background

<table>
<thead>
<tr>
<th>3D Dataset</th>
<th>Method</th>
<th>Studio GS 3DPCK</th>
<th>Studio no GS 3DPCK</th>
<th>Outdoor 3DPCK</th>
<th>All 3DPCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human3.6m</td>
<td>Domain adapt.</td>
<td>44.1</td>
<td>42.6</td>
<td>35.2</td>
<td>41.4</td>
<td>17.7</td>
</tr>
<tr>
<td>Human3.6m</td>
<td>Corr. + Fusion</td>
<td>70.8</td>
<td>62.3</td>
<td>58.5</td>
<td>64.7</td>
<td>31.7</td>
</tr>
<tr>
<td>Ours Unaug.</td>
<td>Corr + Fusion</td>
<td>84.1</td>
<td>68.9</td>
<td>59.6</td>
<td>72.5</td>
<td>36.9</td>
</tr>
<tr>
<td>Ours. Aug.</td>
<td>Base + Fusion</td>
<td>82.6</td>
<td>66.7</td>
<td>62.0</td>
<td>71.7</td>
<td>36.4</td>
</tr>
<tr>
<td>Ours. Aug. + Human3.6m</td>
<td>Corr. + Fusion no persp. corr. with pred. BB</td>
<td>84.6</td>
<td>72.4</td>
<td>69.7</td>
<td>76.5</td>
<td>40.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>81.9</td>
<td>68.6</td>
<td>67.4</td>
<td>73.5</td>
<td>37.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80.4</td>
<td>71.2</td>
<td>69.8</td>
<td>74.4</td>
<td>39.6</td>
</tr>
</tbody>
</table>
The Focus of My Presentation

- Introduction
- Common Approaches
- Current Approach
  - Bounding Box
  - 3D Pose Prediction
  - Global 3D Pose Computation
- Data Set Contribution
- Evaluation
  - Perspective Correction
  - Multi-level Corrective Skip Connections
  - New Data Set
  - Transfer Learning
  - Multi-modal Prediction
  - State of the Art
- Summary
Evaluating Transfer Learning from 2DPoseNet to 3DPoseNet

Mean Per Joint Position Error (MPJPE) in mm on Human3.6M Test Data

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>Discuss</th>
<th>Eating</th>
<th>Greet</th>
<th>Phone</th>
<th>Posing</th>
<th>Purch.</th>
<th>Sitting</th>
<th>Sit Down</th>
<th>Smoke</th>
<th>Take Photo</th>
<th>Wait</th>
<th>Walk</th>
<th>Walk Dog</th>
<th>Walk Pair</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialized with Resnet-101 (ImageNet) weights</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>98.98</td>
<td>100.14</td>
<td>86.07</td>
<td>101.83</td>
<td>101.34</td>
<td>96.74</td>
<td>94.89</td>
<td>125.28</td>
<td>158.31</td>
<td>100.21</td>
<td>112.49</td>
<td>99.57</td>
<td>83.39</td>
<td>109.61</td>
<td>95.79</td>
<td>104.32</td>
</tr>
<tr>
<td>Regular Skip</td>
<td>113.34</td>
<td>112.26</td>
<td>97.40</td>
<td>110.50</td>
<td>108.63</td>
<td>112.09</td>
<td>105.67</td>
<td>125.97</td>
<td>173.41</td>
<td>109.34</td>
<td>120.87</td>
<td>107.75</td>
<td>97.30</td>
<td>126.05</td>
<td>117.45</td>
<td>115.29</td>
</tr>
<tr>
<td>Corr. Skip + Fusion</td>
<td>92.57</td>
<td>99.08</td>
<td>85.46</td>
<td>95.60</td>
<td>96.93</td>
<td>95.67</td>
<td>89.56</td>
<td>95.67</td>
<td>123.54</td>
<td>160.98</td>
<td>97.13</td>
<td>107.56</td>
<td>93.86</td>
<td>76.99</td>
<td>88.73</td>
<td>101.09</td>
</tr>
<tr>
<td>Transfer Learning from 2DPoseNet weights</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr. + Fusion</td>
<td>59.69</td>
<td>69.74</td>
<td>60.55</td>
<td>68.77</td>
<td>76.36</td>
<td>59.05</td>
<td>75.04</td>
<td>96.19</td>
<td>122.92</td>
<td>70.82</td>
<td>85.42</td>
<td>68.45</td>
<td>54.41</td>
<td>82.03</td>
<td>59.79</td>
<td>74.14</td>
</tr>
</tbody>
</table>
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
### Evaluating Multi-modal Prediction and 3D Pose Fusion

#### Mean Per Joint Position Error (MPJPE) in mm on Human3.6M Test Data

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>Discuss</th>
<th>Eating</th>
<th>Greet</th>
<th>Phone</th>
<th>Posing</th>
<th>Purch.</th>
<th>Sitting</th>
<th>Sit Down</th>
<th>Smoke</th>
<th>Take Photo</th>
<th>Wait</th>
<th>Walk</th>
<th>Walk Dog</th>
<th>Walk Pair</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initialized with Resnet-101 (ImageNet) weights</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base</td>
<td>98.98</td>
<td>100.14</td>
<td>86.07</td>
<td>101.83</td>
<td>101.34</td>
<td>96.74</td>
<td>94.89</td>
<td>125.28</td>
<td>158.31</td>
<td>100.21</td>
<td>112.49</td>
<td>99.57</td>
<td>83.39</td>
<td>109.61</td>
<td>95.79</td>
<td>104.32</td>
</tr>
<tr>
<td>Regular Skip</td>
<td>113.34</td>
<td>112.26</td>
<td>97.40</td>
<td>110.50</td>
<td>108.63</td>
<td>112.09</td>
<td>105.67</td>
<td>125.97</td>
<td>173.41</td>
<td>109.34</td>
<td>120.87</td>
<td>107.75</td>
<td>97.30</td>
<td>126.05</td>
<td>117.45</td>
<td>115.29</td>
</tr>
<tr>
<td>Corr. Skip + Fusion</td>
<td><strong>92.57</strong></td>
<td>99.08</td>
<td>85.46</td>
<td><strong>95.04</strong></td>
<td>96.93</td>
<td><strong>89.56</strong></td>
<td>95.67</td>
<td>123.54</td>
<td>160.98</td>
<td>97.13</td>
<td><strong>107.56</strong></td>
<td><strong>93.86</strong></td>
<td><strong>76.99</strong></td>
<td>110.93</td>
<td><strong>88.73</strong></td>
<td>101.09</td>
</tr>
<tr>
<td><strong>Transfer Learning from 2DPosenet weights</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr. + Fusion</td>
<td>59.69</td>
<td>69.74</td>
<td>60.55</td>
<td>68.77</td>
<td>76.36</td>
<td>59.05</td>
<td>75.04</td>
<td>96.19</td>
<td>122.92</td>
<td>70.82</td>
<td>85.42</td>
<td>68.45</td>
<td>54.41</td>
<td>82.03</td>
<td>59.79</td>
<td>74.14</td>
</tr>
</tbody>
</table>
Evaluating Multi-modal Prediction and 3D Pose Fusion

Percentage of Correct Keypoints (PCK) and Area Under the Curve (AUC) on MPI-INF-3DHP Test Data. GS = green screen background

<table>
<thead>
<tr>
<th>3D Data</th>
<th>Network Arch.</th>
<th>Studio GS 3DPCK</th>
<th>Studio no GS 3DPCK</th>
<th>Outdoor 3DPCK</th>
<th>All 3DPCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human 3.6m</td>
<td>Base</td>
<td>21.1</td>
<td>32.5</td>
<td>10.8</td>
<td>22.6</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>Corr Skip +Fusion</td>
<td>22.2</td>
<td>33.9</td>
<td>18.5</td>
<td>25.1</td>
<td>8.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22.3</td>
<td>34.2</td>
<td>20.0</td>
<td>26.0</td>
<td>9.5</td>
</tr>
<tr>
<td>Ours Unaug.</td>
<td>Base</td>
<td>73.6</td>
<td>42.9</td>
<td>19.5</td>
<td>49.0</td>
<td>23.3</td>
</tr>
<tr>
<td></td>
<td>Corr Skip +Fusion</td>
<td>66.9</td>
<td>38.2</td>
<td>27.9</td>
<td>46.8</td>
<td>20.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>67.6</td>
<td>39.6</td>
<td>28.5</td>
<td>47.8</td>
<td>21.8</td>
</tr>
<tr>
<td>Ours Aug.</td>
<td>Base</td>
<td>77.2</td>
<td>59.5</td>
<td>48.7</td>
<td>63.7</td>
<td>31.1</td>
</tr>
<tr>
<td></td>
<td>Corr. Skip +Fusion</td>
<td>71.1</td>
<td>51.7</td>
<td>36.1</td>
<td>55.4</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>73.5</td>
<td>53.1</td>
<td>37.9</td>
<td>57.3</td>
<td>28.0</td>
</tr>
</tbody>
</table>
Evaluating Multi-modal Prediction and 3D Pose Fusion

Base (MPII+LSP)  Base + Fusion (MPII+LSP)  Ground Truth

Large Self Occlusion
The Focus of My Presentation

• Introduction
• Common Approaches
• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation
• Data Set Contribution
• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art
• Summary
Comparison With the State of the Art

- Base structure of **Deep Kinematic Pose** consists of ResNet that is pre-trained on ImageNet

- **SMPLify** uses a different technique. It first predicts 2D body joint locations. Then tries to fit 3D shape to these 2D joints

- 25% improvement over the state of the art regression based model

Table 4. Comparison of results on Human3.6m [28] with the state of the art. Human3.6m, Subjects 1,5,6,7,8 used for training, and 9,11 used for testing. $^S$ = Scaled to test subject specific skeleton, computed from T-pose. $^T$ = Uses Temporal Information, $^{14/17}$ = Joint set evaluated, $^A$ = Uses Best Alignment To GT per frame, $^{Act}$ = Activitywise Training, $^{1/10/64}$ = Test Set Frame Sampling

<table>
<thead>
<tr>
<th>Method</th>
<th>Total MPJPE (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Kinematic Pose $^{[78]}$ $^{J17,B}$</td>
<td>107.26</td>
</tr>
<tr>
<td>Sparse. Deep. $^{[80]}$ $^{T,J17,B,10,Act}$</td>
<td>113.01</td>
</tr>
<tr>
<td>Motion Comp. Seq. $^{[65]}$ $^{T,J17,B}$</td>
<td>124.97</td>
</tr>
<tr>
<td>LinKDE $^{[28]}$ $^{J17,B,Act}$</td>
<td>162.14</td>
</tr>
<tr>
<td>Du et al. $^{[75]}$ $^{T,J17,B}$</td>
<td>126.47</td>
</tr>
<tr>
<td>Rogez et al. $^{[51]}$ $^{J13,B,64}$</td>
<td>121.20</td>
</tr>
<tr>
<td>SMPLify $^{[10]}$ $^{J14,B,A,(First cam.)}$</td>
<td>82.3</td>
</tr>
<tr>
<td><strong>Ours (with 2DposeNet Transfer)</strong></td>
<td></td>
</tr>
<tr>
<td>Corr. Skip + Fusion $^{J17,B}$</td>
<td>74.11</td>
</tr>
<tr>
<td>Corr. Skip + Fusion $^{J17,B,S}$</td>
<td>68.61</td>
</tr>
<tr>
<td>Corr. Skip + Fusion $^{J14,B,A}$</td>
<td>54.59</td>
</tr>
</tbody>
</table>
• Training Data Set: Human3.6M
• Evaluation Data Set: MPI-INF-3DHP

• Deep Kinematic Pose attains 13.8% 3DPCK
• This method attains
  • 26% 3DPCK without Transfer Learning
  • 64.7% 3DPCK with Transfer Learning
Percentage of Correct Keypoints (PCK) and Area Under the Curve (AUC) on MPI-INF-3DHP Test Data. GS = green screen background

<table>
<thead>
<tr>
<th>3D Data</th>
<th>Network Arch.</th>
<th>Studio GS 3DPCK</th>
<th>Studio no GS 3DPCK</th>
<th>Outdoor 3DPCK</th>
<th>All 3DPCK</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human</strong></td>
<td><strong>3.6m</strong></td>
<td>Base</td>
<td>21.1</td>
<td>32.5</td>
<td>10.8</td>
<td>22.6</td>
</tr>
<tr>
<td></td>
<td>Corr Skip +Fusion</td>
<td>22.2</td>
<td>33.9</td>
<td>18.5</td>
<td>25.1</td>
<td>8.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22.3</td>
<td>34.2</td>
<td>20.0</td>
<td>26.0</td>
<td>9.5</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>Unaug.</strong></td>
<td>Base</td>
<td>73.6</td>
<td>42.9</td>
<td>19.5</td>
<td>49.0</td>
</tr>
<tr>
<td></td>
<td>Corr Skip +Fusion</td>
<td>66.9</td>
<td>38.2</td>
<td>27.9</td>
<td>46.8</td>
<td>20.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>67.6</td>
<td>39.6</td>
<td>28.5</td>
<td>47.8</td>
<td>21.8</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>Aug.</strong></td>
<td>Base</td>
<td>77.2</td>
<td>59.5</td>
<td>48.7</td>
<td>63.7</td>
</tr>
<tr>
<td></td>
<td>Corr. Skip +Fusion</td>
<td>71.1</td>
<td>51.7</td>
<td>36.1</td>
<td>55.4</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>73.5</td>
<td>53.1</td>
<td>37.9</td>
<td>57.3</td>
<td>28.0</td>
</tr>
</tbody>
</table>
The Focus of My Presentation

• Introduction

• Common Approaches

• Current Approach
  ➢ Bounding Box
  ➢ 3D Pose Prediction
  ➢ Global 3D Pose Computation

• Data Set Contribution

• Evaluation
  ➢ Perspective Correction
  ➢ Multi-level Corrective Skip Connections
  ➢ New Data Set
  ➢ Transfer Learning
  ➢ Multi-modal Prediction
  ➢ State of the Art

• Summary
Summary

• Direct3D Human Pose Estimation Using a Single Image

• Contributions of this paper
  ➢ Transfer Learning
  ➢ Multi-Level Corrective Skip Connections
  ➢ 3D Pose Fusion
  ➢ New Data Set
  ➢ Perspective Correction

• Achieves better performance than the state of the art

• Further work needed for real-time estimation

• Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778

• Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. IEEE CVPR'14

• Sam Johnson and Mark Everingham “Learning Effective Human Pose Estimation from Inaccurate Annotation” In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR2011)


Monocular 3D Human Pose Estimation

Ikrima Bin Saeed