Human Pose Regression by Combining Indirect Part Detection and Contextual Information

Diogo Luvizon
3rd year PhD student

Supervised by Hedi Tabia and David Picard

diogo.luvizon@ensea.fr

ETIS UMR 8051, Université Paris Seine,
Université de Cergy-Pontoise, ENSEA, CNRS

December 7, 2017
Summary

1. Introduction
2. Proposed regression method
3. Experiments
4. Conclusions
Summary

Introduction

Context

Existing approaches

Motivation

Proposed regression method

Experiments

Conclusions
Human pose estimation is a key step to understand people in images and to perform action recognition.

However, the task involves many difficulties:
- The human body is strongly articulated.
- Some parts may not be visible or are ambiguous.
- The visual appearance of body parts can change significantly.
Summary

1. Introduction
   ▶ Context
   ▶ Existing approaches
   ▶ Motivation

2. Proposed regression method

3. Experiments

4. Conclusions
Existing approaches for human pose estimation

- In the literature, we can find two types of methods:
  - **detection** based, and
  - **regression** based methods.
- Methods based on **detection** try to detect keypoints individually.
  - Classical approaches use handcrafted features + classification.
- **Regression** methods map directly the input image $I$ to poses:
  - $y \leftarrow f(I)$, where $f$ can be a CNN and $y = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- State-of-the-art results are mostly achieved by **detection** based methods.
The most recent methods are based on Convolutional Neural Networks (CNN) e.g., Newell et al., ECCV’16.

In that case, a CNN is used to produce heat maps, which correspond to a map of detection scores for each joint.

Joint coordinates are given by the 2D argmax on heat maps.
Summary

1. Introduction
   - Context
   - Existing approaches
   - Motivation

2. Proposed regression method

3. Experiments

4. Conclusions
Motivation and Contributions

Detection based approaches use the argmax function, which is non-differentiable:

Regression approaches can be end-to-end differentiable:

Our contribution: a regression method that performs as good as detection based approaches.
Summary

1 Introduction

2 Proposed regression method
   ▶ The Soft-argmax layer
   ▶ Network architecture
   ▶ Training details

3 Experiments

4 Conclusions
Proposed approach: the Soft-argmax

- For each body joint, the feature map $h$ is normalized by the spatial Softmax:

$$\Phi(h)_{i,j} = \frac{e^{h_{i,j}}}{\sum_{k=1}^{W} \sum_{l=1}^{H} e^{h_{k,l}}}$$  \hspace{1cm} (1)

- Then, the maximum value is approximated by the expectation on normalized maps:

$$\Psi_d(h) = \sum_{i=1}^{W} \sum_{j=1}^{H} W_{i,j,d} \Phi(h)_{i,j}$$  \hspace{1cm} (2)

$$W_{i,j,x} = \frac{i}{W}, \quad W_{i,j,y} = \frac{j}{H}. \hspace{1cm} (3)$$
The Soft-argmax layer

- The feature maps $h$ converge to “heat map” representations indirectly, but the outputs are the $x$ and $y$ coordinates.
- Since heat maps receive indirect supervision, we can combine many to produce one output $x$-$y \rightarrow$ contextual maps.

![Diagram of Soft-argmax layer with heat maps and probability maps](image)
Summary

Introduction

Proposed regression method
  ▶ The Soft-argmax layer
  ▶ Network architecture
  ▶ Training details

Experiments

Conclusions
Overview of the network architecture

- The entry flow (Stem) is based on Inception-v4
- Prediction blocks (blocks A and B) are based on depth-wise separable convolutions.

Proposed regression method | Network architecture

Human Pose Regression | Diogo Luvizon | December 7, 2017 14 / 21
The probability $p_n$ of the $n^{th}$ joint being detected in the image is given by the Sigmoid function on the maximum value of $h$.

We provide multi-source information to the final prediction by including specialized and contextual maps (e.g., 4 additionnal maps per joints).

The final joint location is given by the aggregated prediction:

$$y_n = \alpha y^d_n + (1 - \alpha) \sum_{i=1}^{N_c} p^c_{i,n} y^c_{i,n} \frac{\sum_{i=1}^{N_c} p^c_{i,n}}{\sum_{i=1}^{N_c} p^c_{i,n}}, \tag{4}$$
Summary

1. Introduction

2. Proposed regression method
   - The Soft-argmax layer
   - Network architecture
   - Training details

3. Experiments

4. Conclusions
Proposed regression method

Training and implementation details

Body joint regression

We use the elastic net loss function (L1 + L2):

\[
L_y = \frac{1}{N_J} \sum_{n=1}^{N_J} \| y_n - \hat{y}_n \|_1 + \| y_n - \hat{y}_n \|_2^2,
\]  

(5)

Body joint probability

We use the binary cross entropy loss:

\[
L_p = \frac{1}{N_J} \sum_{n=1}^{N_J} [(p_n - 1) \log (1 - \hat{p}_n) - p_n \log \hat{p}_n],
\]  

(6)

- The Soft-argmax layer can be implemented as a fully-connected layer with fixed parameters (not trainable).
# Results on 2D pose estimation (MPII)

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Shou.</th>
<th>Elbow</th>
<th>Wrist</th>
<th>Hip</th>
<th>Knee</th>
<th>Ankle</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Detection based methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wei et al. CVPR’16</td>
<td>97.8</td>
<td>95.0</td>
<td>88.7</td>
<td>84.0</td>
<td>88.4</td>
<td>82.8</td>
<td>79.4</td>
<td>88.5</td>
</tr>
<tr>
<td>Bulat and T. ECCV’16</td>
<td>97.9</td>
<td>95.1</td>
<td>89.9</td>
<td>85.3</td>
<td>89.4</td>
<td>85.7</td>
<td>81.7</td>
<td>89.7</td>
</tr>
<tr>
<td>Newell et al. ECCV’16</td>
<td>98.2</td>
<td>96.3</td>
<td>91.2</td>
<td>87.1</td>
<td>90.1</td>
<td>87.4</td>
<td>83.6</td>
<td>90.9</td>
</tr>
<tr>
<td>Chu et al. CVPR’17</td>
<td><strong>98.5</strong></td>
<td>96.3</td>
<td>91.9</td>
<td>88.1</td>
<td>90.6</td>
<td>88.0</td>
<td>85.0</td>
<td>91.5</td>
</tr>
<tr>
<td>Chou et al. arXiv’17</td>
<td>98.2</td>
<td><strong>96.8</strong></td>
<td>92.2</td>
<td>88.0</td>
<td><strong>91.3</strong></td>
<td>89.1</td>
<td>84.9</td>
<td>91.8</td>
</tr>
<tr>
<td>Chen et al. ICCV’17</td>
<td>98.1</td>
<td>96.5</td>
<td><strong>92.5</strong></td>
<td><strong>88.5</strong></td>
<td>90.2</td>
<td><strong>89.6</strong></td>
<td><strong>86.0</strong></td>
<td><strong>91.9</strong></td>
</tr>
<tr>
<td><strong>Regression based methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rogez et al. CVPR’17</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>74.2</td>
</tr>
<tr>
<td>Carreira et al. CVPR’16</td>
<td>95.7</td>
<td>91.7</td>
<td>81.7</td>
<td>72.4</td>
<td>82.8</td>
<td>73.2</td>
<td>66.4</td>
<td>81.3</td>
</tr>
<tr>
<td>Sun et al. arXiv’17</td>
<td>97.5</td>
<td>94.3</td>
<td>87.0</td>
<td>81.2</td>
<td>86.5</td>
<td>78.5</td>
<td>75.4</td>
<td>86.4</td>
</tr>
<tr>
<td><strong>Our method</strong></td>
<td><strong>98.1</strong></td>
<td><strong>96.6</strong></td>
<td><strong>92.0</strong></td>
<td><strong>87.5</strong></td>
<td><strong>90.6</strong></td>
<td><strong>88.0</strong></td>
<td><strong>82.7</strong></td>
<td><strong>91.2</strong></td>
</tr>
</tbody>
</table>
Samples from the MPII dataset
Figure – Image (a), specialized maps (b) and prediction (c), context maps (d, e), and the final aggregated prediction (f).
Conclusions and future work

Contributions

- Regression based pose estimation with results comparable to detection methods.
- Soft-argmax layer $\rightarrow$ fully differentiable architecture.
- Indirect heat map generation $\rightarrow$ allows for contextual maps.
- Smaller feature maps ($32 \times 32$) $\rightarrow$ uses less memory for training.
Conclusions and future work

Contributions

- Regression based pose estimation with results comparable to detection methods.
- Soft-argmax layer → fully differentiable architecture.
- Indirect heat map generation → allows for contextual maps.
- Smaller feature maps (32 × 32) → uses less memory for training.

Ongoing work

Performing action recognition on top of our pose estimation method.