End-to-End Learning of Deformable Mixture of Parts and Deep Convolutioanal Neural Networks for Human Pose Estimation

Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang

Department of Electronic Engineering, The Chinese University of Hong Kong
Outline

- Introduction
- Methodology
- Experiments
What is Human Pose Estimation?

Results are generated by the proposed methods (without temporal constraints).

Video Credit: Peter Jasko solo - M-idzomer 2013
Human Pose Estimation

is to recover the joint positions of articulated limbs of human body given an image or a video.

Results are generated by the proposed methods (without temporal constraints).

Video Credit: Peter Jasko solo - M-idzomer 2013
Applications

Activity Recognition  Game and Animation  Clothing Parsing

[ Yamaguchi et al. CVPR’14 ]
Challenges

- Articulation
- Forshortening
- Clothing

Dantone et al. CVPR 2013
Structure Modeling

Figure Drawing
- Cylinders for each body parts
- Join up the cylinders

Pictorial Structures
- Unary templates
- Pairwise springs

❌ Unable to handle large variations

Image credit: artintegrity.wordpress.com
Fischler & Elschlager 1973
Felzenszwalb & Huttenlocher 2005
Structure Modeling

Figure Drawing
- Cylinders for each body part
- Join up the cylinders

Pictorial Structures
- Unary templates
- Pairwise springs
  - Unable to handle large variations

Mixtures of each part
- Unary template for each mixture type
- Clustering on $(\Delta x, \Delta y)$

Graph Structures
- Trees
- Latent trees
- Loopy graphs
- Fully connected graphs

Image credit: artintegrity.wordpress.com
Fischler & Elschlager 1973
Felzenszwalb & Huttenlocher 2005
Yang & Ramanan 2011
Deep Learning Based Methods

- Benefits from better neural network architectures
  - VGG
  - GoogLeNet,
  - ResNet
- Structures are only used for post processing

Heat map prediction

Fully Convolutional Networks
Gap Between Deep Models and Structure Modeling
Gap Between Deep Models and Structure Modeling

End-to-End Learning
Motivation: Geometric Constraints Among Body Parts Helps in Learning Better Representation

Local appearance is ambiguous + Global consistency helps training
Outline

- Introduction
- Methodology
- Experiments
Graph Models

$G = (V, E)$

**Vertices**
- Locations and mixture types of body parts
- Modeled by a front-end CNN

**Edges**
- Pairwise spatial relationships between body parts
- Modeled by message passing layers
Framework

CNN → Softmax → Logarithm → Message Passing Layers

- Max poolings at different levels:
  - Max → head
  - Max → neck
  - Max → l.shou
  - Max → r.shou
  - Max → r.knee
  - Max → r.ankle
Framework

\[
F(l, t|I; \theta, \omega) = \sum_{i \in V} \phi(l_i, t_i|I; \theta) + \sum_{(i,j) \in E} \psi(l_i, l_j, t_i, t_j|I; \omega_{i,j}^{t_i, t_j})
\]
**Framework**

\[ l = \{l_i\} = \{(x_i, y_i)\}: \text{location of part } i \]

\[
F(l, t | I; \theta, \omega) = \sum_{i \in V} \phi(l_i, t_i | I; \theta) + \sum_{(i, j) \in E} \psi\left(l_i, l_j, t_i, t_j | I; \omega^{t_i, t_j}_{i, j}\right)
\]

**Message Passing Layers**
Framework

\[ l = \{l_i\} = \{(x_i, y_i)\}: \text{location of part } i \]

\[ F(l, t|I; \theta, \omega) = \sum_{i \in V} \phi(l_i, t_i|I; \theta) \]
\[ + \sum_{(i,j) \in E} \psi(l_i, l_j, t_i, t_j|I; \omega_{i,j}) \]

\[ t = \{t_i\}: \text{mixture type of part } i \]
Front-End CNN
Part Appearance Terms

\[ F(l, t| I; \theta, \omega) = \sum_{i \in V} \phi(l_i, t_i | I; \theta) \]
\[ + \sum_{(i,j) \in E} \psi(l_i, l_j, t_i, t_j | I; \omega_{i,j}) \]

\[ \phi(l, t | I; \theta) = \log \frac{1}{\sum_{l'} \exp \left( \theta \cdot \mathcal{D}(l, l') \right) \exp \left( \theta \cdot \log \mathcal{D}(l, l') \right) \mathcal{D}(l, l') } \]

\[ \psi(l, l', t, t' | I; \omega) = \exp \left( \omega \cdot \log \mathcal{D}(l, l') \right) \mathcal{D}(l, l') \]

\[ \mathcal{D}(l, l') = \frac{\sum_{I} l \cdot l'}{\sum_{I} l^2} \]
Message Passing Layers
Spatial Relationship Terms

\[
F(l, t|I; \theta, \omega) = \sum_{i \in V} \phi(l_i, t_i|I; \theta) + \sum_{(i, j) \in E} \psi(l_i, l_j, t_i, t_j|I; \omega_{i, j}^{t_i, t_j})
\]
Local confidence of the appearance of part $i$ with mixture type $t_i$:

$$
\phi(l_i, t_i | I; \theta) = \log p(l_i, t_i | I; \theta) = \log \sigma(f(l_i, t_i | I; \theta))
$$
Local confidence of the appearance of part $i$ with mixture type $t_i$:

$$
\phi(l_i, t_i | I; \theta) = \log p(l_i, t_i | I; \theta) = \log \sigma(f(l_i, t_i | I; \theta))
$$

output of the front-end CNN
Local confidence of the appearance of part $i$ with mixture type $t_i$:

$$\phi(l_i, t_i|I; \theta) = \log p(l_i, t_i|I; \theta) = \log \sigma(f(l_i, t_i|I; \theta))$$
Front end CNN

Local confidence of the appearance of part $i$ with mixture type $t_i$:

$$\phi(l_i, t_i|I; \theta) = \log p(l_i, t_i|I; \theta) = \log \sigma(f(l_i, t_i|I; \theta))$$

Softmax function
Local confidence of the appearance of part $i$ with mixture type $t_i$:

$$\phi(l_i, t_i | I; \theta) = \log p(l_i, t_i | I; \theta) = \log \sigma(f(l_i, t_i | I; \theta))$$

Logarithm
Message Passing Layers
Spatial Relationship Terms

\[ F(l, t|I; \theta, \omega) = \sum_{i \in V} \phi(l_i, t_i|I; \theta) + \sum_{(i,j) \in E} \psi(l_i, l_j, t_i, t_j | I; \omega_{i,j}) \]
Spatial Relationship Terms

Quadratic deformation constraints

\[
\psi(l_i, l_j, t_i, t_j | I; \omega_{i,j}^{t_i,t_j}) = \langle \omega_{i,j}^{t_i,t_j}, d(l_i, l_j) \rangle
\]

- \( d(l_i, l_j) = [\Delta x, \Delta x^2, \Delta y, \Delta y^2] \)
- \( \Delta x = x_i - x_j, \Delta y = y_i - y_j \)

Message Passing: Max-Sum Algorithm

\[ m_{ij}(l_j, t_j) \quad \text{The message passed from part } i \text{ to } j \]
Message Passing: Max-Sum Algorithm

$m_{ij}(l_j, t_j)$ \text{ The message passed from part } i \text{ to } j

$u_i(l_i, t_i)$ \text{ The belief of part } i
Message Passing: Max-Sum Algorithm

\[ m_{ij}(l_j, t_j) \leftarrow \alpha_m \max_{l_i, t_i} \left( u_i(l_i, t_i) + \psi(l_i, l_j, t_i, t_j) \right) \]

\[ u_i(l_i, t_i) \leftarrow \alpha_u \left( \phi(l_i, t_i) + \sum_{k \in N(i)} m_{ki}(l_i, t_i) \right) \]

\[ (l_i^*, t_i^*) = \arg\max_{l_i, t_i} u_i^*(l_i, t_i) \]
Diagnosis on Message Passing Layers

1st
Diagnosis on Message Passing Layers
Diagnosis on Message Passing Layers

3rd
Outline

- Introduction
- Methodology
- Experiments
Datasets

LSP

Sports
1000 training
1000 testing

FLIC

Films
3987 training
1016 testing
Qualitative Results on the LSP Dataset
PCP on the LSP dataset

[Percentage of Correct Parts]
STRICT PCP ON THE LSP DATASET

- Yang&Ramanan, CVPR'11
- Eichner&Ferrari, ACCV'13
- Pose Machines, ECCV'14
- Pishchulin et al., ICCV'13
- Ouyang et al., CVPR'14
- Ours
- Kiefel&Gehler, ECCV'14
- DeepPose, CVPR'14
- Chen&Yuille, NIPS'14
Percentage of Detected Joints (PDJ)

**Our method**

![Graphs showing detection rates for wrists, elbows, ankles, and knees across different methods with normalized precision thresholds.](image-url)
Datasets

LSP
Sports
1000 training
1000 testing

FLIC
Films
3987 training
1016 testing

Image Parse
Activities
100 training
205 testing
Cross-dataset validation
Generalization

Image Parse Dataset

- Ours
- Ouyang et al., CVPR'14
- Yang&Ramanan, TPAMI'13
- Pishchulin et al., ICCV'13
- Pishchulin et al., CVPR'13
- Pishchulin et al., CVPR'12
- Johnson&Everingham, CVPR'11
- Yang&Ramanan, CVPR'11
Component Analysis
Unary Term vs. Full Model

**STRICT PCP ON THE LSP DATASET (VGG-LG)**
Tree-Structured Model vs. Loopy Model

<table>
<thead>
<tr>
<th></th>
<th>Tree Model</th>
<th>Loopy Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>81.1</td>
<td>80.7</td>
</tr>
<tr>
<td>Lower Legs</td>
<td>81.7</td>
<td>81.1</td>
</tr>
<tr>
<td>Upper Legs</td>
<td></td>
<td>88.7</td>
</tr>
<tr>
<td>Lower Arms</td>
<td>66.7</td>
<td>65.8</td>
</tr>
<tr>
<td>Upper Arms</td>
<td></td>
<td>78.8</td>
</tr>
<tr>
<td>Head</td>
<td>83.1</td>
<td>83.4</td>
</tr>
<tr>
<td>Torso</td>
<td>96.5</td>
<td>96.2</td>
</tr>
</tbody>
</table>
Human Pose Estimation in This CVPR

Chu et al. Structured Feature Learning for Pose Estimation.

Carreira et al. Human Pose Estimation With Iterative Error Feedback.

Wei et al. Convolutional Pose Machines.
Summary

End-to-End Learning

- Bridge the gap between structure modeling and deep learning
- A new message passing layer, which is flexible to tree-structured/loopy relational graphs.
Thank you.

Welcome to our poster session #5