STRUCTURED OUT PREDICTION (POSE)

HYUN SOO PARK
CHALLENGES OF VISUAL RECOGNITION

• Appearance
  • DOF: texture, illumination, material, shading, ...

• Shape
  • DOF: object category, geometric pose, viewpoint, ...
**Challenges of Visual Recognition**

- **Appearance**
  - DOF: texture, illumination, material, shading, ...
- **Shape**
  - DOF: object category, geometric pose, viewpoint, ...

$$f(I) = l_{\text{human}}$$

$$f(I) = \begin{bmatrix} x_1 \\ y_1 \\ \vdots \\ x_n \\ y_n \end{bmatrix}$$

Human

Human pose

Structured output
CHALLENGES OF VISUAL RECOGNITION

• Appearance
  • DOF: texture, illumination, material, shading, ...
• Shape
  • DOF: object category, geometric pose, viewpoint, ...

Landmark localization (structured prediction)

Learning appearance and spatial relation together
Topological Model

Star topology

Tree topology

Fully connected topology
PROBLEM DEFINITION

$w_{head} \cdot \phi(I, x_{head}) > 0$

Head detector

Feature extractor
**Problem Definition**

Head detector

\[ w_{head} \cdot \phi(I, x_{head}) > 0 \]

Arm detector

\[ w_{arm} \cdot \phi(I, x_{arm}) > 0 \]
**Problem Definition**

**Head detector**

\[ w_{\text{head}} \cdot \phi(I, x_{\text{head}}) > 0 \]

Feature extractor

\[ w_{\text{arm}} \cdot \phi(I, x_{\text{arm}}) > 0 \]

**Arm detector**

**Objective:**

\[ L = w_{h} \cdot \phi(I, x_{h}) + w_{a} \cdot \phi(I, x_{a}) \]
**Problem Definition**

\[ w_{\text{head}} \cdot \phi(I, x_{\text{head}}) > 0 \]

Feature extractor

\[ w_{\text{arm}} \cdot \phi(I, x_{\text{arm}}) > 0 \]

Head detector

Arm detector

Objective:

\[ L = w_{h} \cdot \phi(I, x_{h}) + w_{a} \cdot \phi(I, x_{a}) + \varphi(x_{h}, x_{a}) \]

Spatial relationship
**Problem Complexity**

Head detector

\[ w_{head} \cdot \phi(I, x_{head}) > 0 \]

Feature extractor

\# of configurations

\[ O(h^2) \]

Arm detector

\[ w_{arm} \cdot \phi(I, x_{arm}) > 0 \]

\[ h : \# \text{ of pixels} \]

Objective:

\[ L = w_h \cdot \phi(I, x_h) + w_a \cdot \phi(I, x_a) + \varphi(x_h, x_a) \]

Spatial relationship
**Problem Complexity**

\[ w_{head} \cdot \phi(I, x_{head}) > 0 \]

Feature extractor

\[ w_{arm} \cdot \phi(I, x_{arm}) > 0 \]

\[ \sum_{i} w_{i} \cdot \phi(I, x_{i}) + \sum_{i,j} \phi(x_{i}, x_{j}) \]

\[ O(h^{n}) \]

\( h \) : # of pixels

\( n \) : # of landmarks

Head detector

Arm detector

Objective:
**Model Topology**

- Star topology: $O(nh^2)$
- Tree topology: $O(nh^2)$
- Fully connected topology: $O(h^n)$

- Can be solved by dynamic programming
- NP-hard
Object Detection with Discriminatively Trained Part Based Models

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan

Slide inspired by Felzenszwalb and Girshick
**Deformable Part Model (DPM)**

- Root filter
- Part desc.
- Deformation
\[ L = \sum_i w_i \cdot \phi(I, x_i) + \sum_{i,j} \phi(x_i, x_j) \]
\[ L = \sum_i w_i \cdot \phi(I, x_i) + \sum_{i,j} \phi(x_i, x_j) \]

\[
\text{score}(x_0, x_1, \ldots, x_n) = \sum_{i=0} w_i \cdot \phi(I, x_i) - \sum_{i=1} d_i \cdot (dx_i^2)
\]

Appearance (filter response)
DPM

\[ L = \sum_i w_i \cdot \phi(I, x_i) + \sum_{i,j} \phi(x_i, x_j) \]

\[ \text{score}(x_0, x_1, \ldots, x_n) = \sum_i w_i \cdot \phi(I, x_i) - \sum_{i=1}^n d_i \cdot (dx_i^2) \]

\text{Spatial rel. (spring force)}

where \[ dx_i = x_0 + c_i - x_i \]

Offset
Score for root location based on part location

\[ score(x_0) = \max_{x_0, x_1, \ldots, x_n} score(x_0, x_1, \ldots, x_n) \]

\[ = w_0 \cdot \phi(I, x_0) + \sum_{i=1}^{n} \max_{x_i} score(x_0, x_i) \]

\[ = w_0 \cdot \phi(I, x_0) + \sum_{i=1}^{n} \max_{x_i} \left( w_i \cdot \phi(I, x_i) - d_i \cdot dx^2 \right) \]

Part evidence to predict the root

How does the head location tell us about the root location?
DPM

Input image

Filter response

Transformed response

Head filter

\[ R_h(x_h) = w_h \cdot \phi(I, x_h) \]

\[ D_h(x_0) = \max_{dx,dy} \left( R_h(x_0 + dx) - d_h \cdot dx^2 \right) \]
$$score(x_0) = R_0(x_0) + \sum_{i=1}^{\infty} \max_i \left( R_i(x_0 + dx) - d_i \cdot dx^2 \right)$$

$$R_h(x_h) = w_h \cdot \phi(I, x_h)$$

$$D_h(x_0) = \max_{dx, dy} \left( R_h(x_0 + dx) - d_h \cdot dx^2 \right)$$
Articulated pose estimation with flexible mixtures-of-parts

Yi Yang    Deva Ramanan
Dept. of Computer Science, University of California, Irvine
{yyang8,dramanan}@ics.uci.edu

CVPR 2011
2018 Longuet-Higgins Prize for fundamental contributions in computer vision

Slide inspired by Yang
**Limitation of DPM**

<table>
<thead>
<tr>
<th>In-plane rotation</th>
<th>Foreshortening</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="In-plane rotation images" /></td>
<td><img src="image2" alt="Foreshortening images" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scaling</th>
<th>Out-of-plane rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="Scaling images" /></td>
<td><img src="image4" alt="Out-of-plane rotation images" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intra-category variation</th>
<th>Aspect ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5" alt="Intra-category variation images" /></td>
<td><img src="image6" alt="Aspect ratio images" /></td>
</tr>
</tbody>
</table>
Part based model

Marr, Nishihara (1978)

3D deformation ~
Mixture of mini-parts
(a) Original  (b) Foreshortening  (c) Out-of-plane
**Mixture Part Model**

![Diagram of Mixture Part Model]

\[
L(I,x) = \sum_{i \in V} \alpha_i \cdot \phi(I,x_i) + \sum_{i,j \in E} \beta_{ij} \cdot \varphi(x_i,x_j)
\]

Mixture part model

\[
L(I,x,m) = \sum_{i \in V} \alpha^{m_i} \cdot \phi(I,x_i) + \sum_{i,j \in E} \beta_{ij}^{m_i m_j} \cdot \varphi(x_i,x_j)
\]

Appearance
**Mixture Part Model**

\[ L(I, x) = \sum_{i \in V} \alpha_i \cdot \phi(I_i, x_i) + \sum_{i, j \in E} \beta_{ij} \cdot \varphi(x_i, x_j) \]

Mixture part model

\[ L(I, x, m) = \sum_{i \in V} \alpha_i^m \cdot \phi(I_i, x_i) + \sum_{i, j \in E} \beta_{ij}^m \cdot \varphi(x_i, x_j) \]

- Appearance
- Spatial rel.
**Mixture Part Model**

\[ L(I, x) = \sum_{i \in V} \alpha_i \cdot \phi(I, x_i) + \sum_{i, j \in E} \beta_{ij} \cdot \varphi(x_i, x_j) \]

Mixture part model

\[ L(I, x, m) = \sum_{i \in V} \alpha_i^m \cdot \phi(I, x_i) + \sum_{i, j \in E} \beta_{ij}^{m_i m_j} \cdot \varphi(x_i, x_j) + S(m) \]

- Appearance
- Spatial rel.
- Mixture prior
Mixture Part Model

Mixture prior

\[ S(m) = \sum_{i \in V} b^{m_i}_i + \sum_{i, j \in E} b^{m_i, m_j}_{ij} \]

- \( b^{m_i}_i \) Prior on mixture \( m_i \)
- \( b^{m_i, m_j}_{ij} \) Co-occurrence prior between \( m_i \) and \( m_j \)
Mixture Part Model

\[
\text{score}_i(x_i, m_i) = \alpha_i^m \cdot \phi(I, x_i) + b_i^m + \sum_{k \in \text{kids}(i)} u_k(x_i, m_i)
\]

\[
u_k(x_i, m_i) = \max_{x_j, m_j} \left( \text{score}_k(x_j, m_j) + \beta_{ij}^{m_i,m_j} \cdot \phi(x_i, x_j) + b_{ij}^{m_i,m_j} \right)
\]

\[
\text{Cf)} \quad \text{score}(x_0) = R_0(x_0) + \sum_{i=1}^{n} \max_{x_i} \left( R_i(x_0 + dx) - d_i \cdot dx^2 \right)
\]
**Mixture Part Model**

Fig. 6: A visualization of our model for $K = 14$ parts and $T = 4$ local mixtures, trained on the Parse dataset. We show the local templates above, and the tree structure below, placing parts at their best-scoring location relative to their parent. Though we visualize 4 trees, there exists $T^K \approx 2^{17}$ global combinations, obtained by composing different part types together with different springs. The score associated with each combination decomposes into a tree, and so is efficient to search over using dynamic programming (1).
Pose Machines: Articulated Pose Estimation via Inference Machines

Varun Ramakrishna, Daniel Munoz, Martial Hebert, J. Andrew Bagnell, and Yaser Sheikh

The Robotics Institute, Carnegie Mellon University
**Pose Inference Machine**

Elbow estimation

\[ \sum_{ij} \beta_{ij}^{m_i m_j} \cdot \varphi(x_i, x_j) \]

All possible pairs of joints
**Pose Inference Machine**

Input Image

Image Location $z$

$\mathbf{x}_z$ → $g_0^p(\mathbf{x}_z)$

Regressor

$b_0(Y_p = z) = g_0^p(\mathbf{x}_z)$

$Y_p$ : location of $p$ joint
**Pose Inference Machine**

\[ b_0(Y_p = z) = g_0^p(x_z) \]

\[ b_1(Y_p = z) = g_1^p(x_z; \oplus \varphi(z; b_0^p)) \]

Context from previous prediction

\[
L(I, x) = \sum_{i \in V} \alpha_i \cdot \phi(I, x_i) + \sum_{i, j \in E} \beta_{ij} \cdot \varphi(x_i, x_j)
\]
**Context Features**

\[ b_1(Y_p = z) = g_1^p(x_z; \oplus \varphi(z; b_0^p)) \]

Context from previous prediction
Stage II Confidence

- Head
- Neck
- L-Shoulder
- L-Elbow
- L-Wrist
Stage III Confidence

Head  Neck  L-Shoulder  L-Elbow  L-Wrist
Hierarchical Representation

Level 1

$^1 g_1$ → Context Features → $^1 g_2$ → Context Features → $^1 g_T$

Level 2

$^2 g_1$ → Context Features → $^2 g_2$ → Context Features → $^2 g_T$

Level L

$L g_1$ → Context Features → $L g_2$ → Context Features → $L g_T$

Stage $t = 1$  Stage $t = 2$  Stage $t = (T = 3)$
Level 3 Confidence Maps
**DATASETS**

FLIC (Frame Labeled In Cinema) dataset (CVPR 2013)
- Human detection on Hollywood movies
- Mechanical Turker for annotation
- Upper body
- 20928 labeled images
DASETS

MPII dataset (CVPR 2014)
- YouTube videos
- 25k images
- 40k humans
- Full body
**DATASETS**

MSCOCO dataset (ECCV 2014)
- Internet images
- 200k images
- 250k humans
- Full body
DeepPose: Human Pose Estimation via Deep Neural Networks

Alexander Toshev       Christian Szegedy
Google
1600 Amphitheatre Pkwy
Mountain View, CA 94043
toshev,szegedy@google.com
**COORDINATE REGRESSION**

Initial stage

\[
\begin{bmatrix}
    x_1 \\
    y_1 \\
    \vdots \\
    x_n \\
    y_n
\end{bmatrix}
= z = f(I; \theta)
\]

\[
\arg\min_{\theta} \sum_i \left\| z_i - f(I_i; \theta) \right\|^2
\]
**Coordinate Regression**

**Initial stage**

DNN-based regressor

\[
\begin{bmatrix}
  x_1 \\
  y_1 \\
  \vdots \\
  x_n \\
  y_n
\end{bmatrix} = z = f(I; \theta)
\]

\[
\text{argmin}_{\theta} \sum_{i} \| z_i - f(I_i; \theta) \|^2
\]

**Stage s**

DNN-based refiner

\[
\text{argmin}_{\theta} \sum_{i} \| z_i - f(b(I_i); \theta) \|^2
\]

Pose refinement in the bounding box
<table>
<thead>
<tr>
<th>Method</th>
<th>Arm</th>
<th>Leg</th>
<th>Ave.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Upper</td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>DeepPose-st1</td>
<td>0.5</td>
<td>0.27</td>
<td>0.74</td>
</tr>
<tr>
<td>DeepPose-st2</td>
<td><strong>0.56</strong></td>
<td>0.36</td>
<td><strong>0.78</strong></td>
</tr>
<tr>
<td>DeepPose-st3</td>
<td><strong>0.56</strong></td>
<td>0.38</td>
<td>0.77</td>
</tr>
<tr>
<td>Dantone et al. [2]</td>
<td>0.45</td>
<td>0.25</td>
<td>0.65</td>
</tr>
<tr>
<td>Tian et al. [25]*</td>
<td>0.52</td>
<td>0.33</td>
<td>0.70</td>
</tr>
<tr>
<td>Johnson et al. [13]</td>
<td>0.54</td>
<td><strong>0.38</strong></td>
<td>0.75</td>
</tr>
<tr>
<td>Wang et al. [26]*</td>
<td><strong>0.565</strong></td>
<td>0.37</td>
<td>0.76</td>
</tr>
<tr>
<td>Pishchulin [18]°</td>
<td>0.49</td>
<td>0.32</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Figure 6. Predicted poses in red and ground truth poses in green for the first three stages of a cascade for three examples.
Convolutional Pose Machines

Shih-En Wei
shiwen@cmu.edu

Varun Ramakrishna
vramakri@cs.cmu.edu

Takeo Kanade
Takeo.Kanade@cs.cmu.edu

Yaser Sheikh
yaser@cs.cmu.edu

The Robotics Institute
Carnegie Mellon University
CONVOLUTIONAL POSE MACHINE (CPM)

(a) Stage 1

Pose machine

(b) Stage ≥ 2

\( g_1 \rightarrow b_1 \)

\( \psi_2 \)

\( g_2 \rightarrow b_2 \)

\[ \ldots \]

\( \psi_T \)

\( g_T \rightarrow b_T \)
**CONVOLUTIONAL POSE MACHINE (CPM)**

Convolutional Pose Machines (T-stage)

- **P** Pooling
- **C** Convolution

(a) Stage 1

x → \( g_1 \) → \( b_1 \)

(b) Stage \( \geq 2 \)

\( x' \) → \( \psi_2 \) → \( g_2 \) → \( b_2 \) → \( \psi_T \) → \( g_T \) → \( b_T \)

(c) Stage 1

Input Image \( h \times w \times 3 \)

\( \text{Loss } f_1 \)

(d) Stage \( \geq 2 \)

\( x' \) → \( \psi_2 \) → \( g_2 \) → \( b_2 \) → \( \psi_T \) → \( g_T \) → \( b_T \)

\( \text{Loss } f_2 \)

(e) Effective Receptive Field

- 9 x 9
- 26 x 26
- 60 x 60
- 96 x 96
- 160 x 160
- 240 x 240
- 320 x 320
- 400 x 400
Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields *

Zhe Cao    Tomas Simon    Shih-En Wei    Yaser Sheikh
The Robotics Institute, Carnegie Mellon University
{zhecao, shihenw}@cmu.edu    {tsimon, yaser}@cs.cmu.edu
**Part Affinity Fields**

(a) Input Image  
(b) Part Confidence Maps  
(c) Part Affinity Fields  
(d) Bipartite Matching  
(e) Parsing Results

Marr, Nishihara (1978)
**PART AFFINITY FIELDS**

- **Stage 1**
  - **Branch 1** $\rho^1$
  - **Loss $f_i^1$**
  - **Convolutions**
    - $3\times3$
    - $3\times3$
    - $3\times3$
    - $1\times1$
    - $1\times1$
  - **Output $S^1$**

- **Stage $t$, ($t \geq 2$)**
  - **Branch 1** $\rho^t$
  - **Loss $f_i^t$**
  - **Convolutions**
    - $7\times7$
    - $7\times7$
    - $7\times7$
    - $7\times7$
    - $1\times1$
    - $1\times1$
  - **Output $S^t$**

- **Joint Inference**
  - **Branch 2** $\phi^1$
  - **Loss $f_i^2$**
  - **Convolutions**
    - $3\times3$
    - $3\times3$
    - $3\times3$
    - $1\times1$
    - $1\times1$
  - **Output $L^1$**

- **PAF Inference**
  - **Stage 1**
  - **Stage $t$, ($t \geq 2$)**
  - **Joint Inference**
  - **Output $S^1$, $S^t$, $L^1$, $L^t$**
Part Affinity Fields Based Connection
<table>
<thead>
<tr>
<th>Method</th>
<th>Hea</th>
<th>Sho</th>
<th>Elb</th>
<th>Wri</th>
<th>Hip</th>
<th>Kne</th>
<th>Ank</th>
<th>mAP</th>
<th>s/image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deepcut [22]</td>
<td>73.4</td>
<td>71.8</td>
<td>57.9</td>
<td>39.9</td>
<td>56.7</td>
<td>44.0</td>
<td>32.0</td>
<td>54.1</td>
<td>57995</td>
</tr>
<tr>
<td>Iqbal et al. [12]</td>
<td>70.0</td>
<td>65.2</td>
<td>56.4</td>
<td>46.1</td>
<td>52.7</td>
<td>47.9</td>
<td>44.5</td>
<td>54.7</td>
<td>10</td>
</tr>
<tr>
<td>DeeperCut [11]</td>
<td>87.9</td>
<td>84.0</td>
<td>71.9</td>
<td>63.9</td>
<td>68.8</td>
<td>63.8</td>
<td>58.1</td>
<td>71.2</td>
<td>230</td>
</tr>
<tr>
<td>Ours</td>
<td>93.7</td>
<td>91.4</td>
<td>81.4</td>
<td>72.5</td>
<td>77.7</td>
<td>73.0</td>
<td>68.1</td>
<td>79.7</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Subset of 288 images as in [22]

<table>
<thead>
<tr>
<th>Method</th>
<th>Hea</th>
<th>Sho</th>
<th>Elb</th>
<th>Wri</th>
<th>Hip</th>
<th>Kne</th>
<th>Ank</th>
<th>mAP</th>
<th>s/image</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeeperCut [11]</td>
<td>78.4</td>
<td>72.5</td>
<td>60.2</td>
<td>51.0</td>
<td>57.2</td>
<td>52.0</td>
<td>45.4</td>
<td>59.5</td>
<td>485</td>
</tr>
<tr>
<td>Iqbal et al. [12]</td>
<td>58.4</td>
<td>53.9</td>
<td>44.5</td>
<td>35.0</td>
<td>42.2</td>
<td>36.7</td>
<td>31.1</td>
<td>43.1</td>
<td>10</td>
</tr>
<tr>
<td>Ours (one scale)</td>
<td>89.0</td>
<td>84.9</td>
<td>74.9</td>
<td>64.2</td>
<td>71.0</td>
<td>65.6</td>
<td>58.1</td>
<td>72.5</td>
<td>0.005</td>
</tr>
<tr>
<td>Ours</td>
<td>91.2</td>
<td>87.6</td>
<td>77.7</td>
<td>66.8</td>
<td>75.4</td>
<td>68.9</td>
<td>61.7</td>
<td>75.6</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Full testing set
https://www.youtube.com/watch?v=pW6nZXeWIGM