Modeling Mutual Context of Object and Human Pose in Human-object Interaction Activities

- Bangpeng Yao
- Li Fei-Fei

Presented by Sahil Shah
Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results
Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results
Introduction

- Note on author
  - Pioneer of ImageNet dataset
  - Must see TED talk in March 2015
Introduction

• Problem: Detecting objects in cluttered scenes and estimating articulated human body parts especially in human object interaction activities
Introduction
Introduction
Introduction

• Key insight: Mutual Context
  – Automatically discover relevant poses
  – Automatically discover spatial relationships
  – Optimize for mutual co-occurrence of object and pose
Introduction

• Contribution
  – Builds up on Prof. Gupta’s work
  – First to use mutual context
  – Jointly solve object detection & pose estimation
Agenda

• Introduction
• Problem Formulation
• Learning
• Inference
• Results
Problem Formulation

• Goal: Given an image of HOI activity we need to estimate human pose (H), detect the object (O) and classify HOI activity (A)

• Model
  – Hierarchical Random Field
  – A, O and H contribute to detection of each other
  – H is a hidden variable
  – Body parts \( \{P_n\} \) are found using feature based detectors and they compose to form H
Problem Formulation

Golf Swing

Tennis Forehand
Problem Formulation
Problem Formulation

• Why need to learn structure?
  – The model captures important connections between object and the body parts
  – Which parts of the body should be connected to overall pose (H) and object (O)?
Problem Formulation

• Model
  – Overall model: \( \Psi = \sum w_e \psi_e \)
  – A,O,H: \( \psi_e(A, O), \psi_e(A, H), \) and \( \psi_e(O, H) \)
    • Counting co-occurrence frequencies
  – Spatial Relationships: \( \psi_e(O, P_n) \) & \( \psi e(P_m, P_n) \)
    • \( \text{bin}(l_O - l_{P_n}) \cdot \text{bin}(\theta_O - \theta_{P_n}) \cdot \mathcal{N}(s_O / s_{P_n}) \)
  – Compatibility: \( \psi_e(H,P_n) \)
    • \( \text{bin}(l_{P_n} - l_{P_1}) \cdot \text{bin}(\theta_{P_n}) \cdot \mathcal{N}(s_{P_n}) \)
  – Object & Body parts: \( \psi_e(O,f_O) \) and \( \psi_e(P_n,f_{P_n}) \)
    • Shape context feature based detectors
Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results
Learning

• Input and Output

Images with labeled objects, body parts & HOI → Model Learning → Set of models - each for one human pose in a particular HOI activity
Learning

• Overall Algorithm

Hill-climbing structure learning for each activity class.

\textbf{foreach} Iteration \textbf{do}
  - Model parameter estimation by max-margin learning;
  - Choose the model with the largest number of mis-classified images;
  - Cluster the images in the selected model into two sub-classes;
  - Structure learning for the two new sub-classes;
\textbf{end}
Learning

• Hill climbing structure learning
  – Each pose in each HOI activity class
  – Add/remove an edge and check for optima
  – Keep tabu list to avoid revisiting solutions
  – Randomly initialize thrice to avoid local optimas
Learning

- Max-margin for parameter estimation
  - Maximize discrimination between different A
  - Each A has subclasses, hence multiple models and multiple weight vectors
  - Training sample: \((x_i, c_i, y(c_i))\)  \(y\): maps \(c_i\) to class label
  - \(F: y(F(x_i)) = y(c_i)\)  \(F(x_i) = \text{argmax}_r\{w_r \cdot x_i\}\)  \(w_r\): weights for \(r\)th subclass.

\[
\min_{w, \xi} \frac{1}{2} \sum_r \|w_r\|_2^2 + \beta \sum_i \xi_i
\]

subject to:  \(\forall i, \xi_i \geq 0\)

\(\forall i, r\) where \(y(r) \neq y(c_i)\),  \(w_{c_i} \cdot x_i - w_r \cdot x_i \geq 1 - \xi_i\)
Learning

• Overall Algorithm

Hill-climbing structure learning for each activity class.

**foreach** Iteration **do**
- Model parameter estimation by max-margin learning;
- Choose the model with the largest number of mis-classified images;
- Cluster the images in the selected model into two sub-classes;
- Structure learning for the two new sub-classes;

**end**
Agenda

• Introduction
• Problem Formulation
• Learning
• Inference
• Results
Inference

• Given a test image (I), estimate pose and detect object and classify activity
  – To detect object (O) we maximize likelihood of the models given that object. Denoted as \( \max_{O,H} \Psi(A_k, O, H, I) \)
  – To detect human pose (H), compute \( \max_{O,H} \Psi(A_k, O, H, I) \) for each \( A_k \) and select the one corresponding to the ML score
Inference

\[ O_1, H_1 = \arg \max_{O, H} \Psi(A_1, O, H, I) \]

\[ O_k, H_k = \arg \max_{O, H} \Psi(A_k, O, H, I) \]

\[ A^*, H^* = \arg \max_k \Psi(A_k, O_k, H_k, I) \]
Agenda

- Introduction
- Problem Formulation
- Learning
- Inference
- Results
Results

- Cricket defensive shot
- Cricket bowling
- Croquet shot
Results

- Object Detection
  - Compare with two experiments
    1. Sliding window as baseline
    2. Pedestrian detector for human’s location context
Results

(a) Cricket Bat

(b) Cricket Ball

(c) Croquet Mallet

(d) Tennis Racket
## Results

- **Pose Estimation**

<table>
<thead>
<tr>
<th>Method</th>
<th>Torso</th>
<th>Upper Leg</th>
<th>Lower Leg</th>
<th>Upper Arm</th>
<th>Fore Arm</th>
<th>Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterative parsing [26]</td>
<td>52±19</td>
<td>22±11</td>
<td>22±10</td>
<td>21±9</td>
<td>28±16</td>
<td>24±16</td>
</tr>
<tr>
<td>Pictorial structure [1]</td>
<td>50±14</td>
<td>31±12</td>
<td>30±9</td>
<td>31±15</td>
<td>27±18</td>
<td>18±6</td>
</tr>
<tr>
<td>Class-based pictorial structure</td>
<td>59±9</td>
<td>36±11</td>
<td>26±17</td>
<td>39±9</td>
<td>27±9</td>
<td>30±12</td>
</tr>
<tr>
<td>Our model, only one pose per class</td>
<td>63±5</td>
<td>40±8</td>
<td>36±15</td>
<td>41±10</td>
<td>31±9</td>
<td>38±13</td>
</tr>
<tr>
<td>Our full model</td>
<td>66±6</td>
<td>43±8</td>
<td>39±14</td>
<td>44±10</td>
<td>34±10</td>
<td>44±9</td>
</tr>
</tbody>
</table>
Results

- HOI classification
  - Compare with SVM with BoW
  - Compare with Gupta et al.
Results

- Upper-left → object detection by mutual context
- Lower-left → object detection by a scanning window
- Upper-right → pose estimation by mutual context
- Lower-right → pose estimation by the state-of-the-art pictorial structure method
Results

- Upper-left → object detection by mutual context
- Lower-left → object detection by a scanning window
- Upper-right → pose estimation by mutual context
- Lower-right → pose estimation by the state-of-the-art pictorial structure method
Thank you!