Human Pose Search using Deep Poselets

Nataraj Jammalamadaka *
Andrew Zisserman § C. V. Jawahar *

* CVIT, IIIT Hyderabad, India

§ Visual Geometry Group, Department of Engineering, University of Oxford
Human Pose: Gesture and action

Human pose is a very important precursor to gesture and action.
Pose Search: Motivation

Retrieve cover drive shots

Retrieve Bharatanatyam poses
Pose Search: System

1. Take a query
2. Build a feature
3. Search through video DB
4. Return the retrieved results
Overview

Deep Poselets

Poselet Discovery
- Cluster pose space

Training
- Train poselets using convolutional neural networks

Detection
- Detect poselets

Pose retrieval

- Given a query image
- Build Bag of Deep poselets
- Return the retrieved results
Datasets

Buffy Stickmen (Season 1, 5 episodes)

ETH Pascal dataset (Flickr Images)

H3D
(Flickr Images)
Datasets

FLIC dataset (30 Hollywood movies)

Movie dataset (Ours) (22 Hollywood movies)
No overlap with FLIC
## Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3D</td>
<td>238</td>
<td>0</td>
<td>0</td>
<td>238</td>
</tr>
<tr>
<td>ETHZ Pascal</td>
<td>0</td>
<td>0</td>
<td>548</td>
<td>548</td>
</tr>
<tr>
<td>Buffy</td>
<td>747</td>
<td>0</td>
<td>0</td>
<td>747</td>
</tr>
<tr>
<td>Buffy-2</td>
<td>396</td>
<td>0</td>
<td>0</td>
<td>396</td>
</tr>
<tr>
<td>Movie</td>
<td>1098</td>
<td>491</td>
<td>2172</td>
<td>3756</td>
</tr>
<tr>
<td>Flic</td>
<td>2724</td>
<td>2279</td>
<td>0</td>
<td>5003</td>
</tr>
<tr>
<td><strong>Total stickmen annotations</strong></td>
<td><strong>5198</strong></td>
<td><strong>2764</strong></td>
<td><strong>2720</strong></td>
<td><strong>10682</strong></td>
</tr>
<tr>
<td><strong>+ Flipped version</strong></td>
<td><strong>10396</strong></td>
<td><strong>5528</strong></td>
<td><strong>5440</strong></td>
<td><strong>21364</strong></td>
</tr>
</tbody>
</table>
Overview

Deep Poselets

Poselet Discovery
• Cluster pose space

Training
• Train poselets using convolutional neural networks

Detection
• Detect poselets

Pose retrieval

• Given a query image
• Build Bag of Deep poselets
• Return the retrieved results
Poselets

Poselets model body parts in a particular spatial configuration.
Poselets

Poselets model body parts in a particular spatial configuration.
Poselets

Poselets model body parts in a particular spatial configuration.
Poselets

Poselets model body parts in a particular spatial configuration.
Poselets: Discovery

Training data with ground truth stickmen annotations

Reorganize

For each body part, note the angle

Cluster on the angles

For each set, get pose descriptors

- For each body part, note the angle
- Cluster on the angles
Deep Poselets: CNNs

Convolutional layers

Convolution followed by pooling

Fully connected layers

Softmax layer

ReLU Non-linearity:

\[ f(x) = \max(0, x) \]

Softmax layer:

\[ f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \]
Deep Poselets: Training

Convolutional layers

Input image: $x$  Model parameters: $w$  Ground truth: $g$  Output: $y = f(x, w)$

Loss function: $L = \sum_j g_j \log(y_j)$

Training: Stochastic Gradient Descent

$$w = w - \frac{\eta \partial L}{\partial w}$$

Architecture from Krizhevsky et al., NIPS 2012
Deep Poselets: Fine tuning

Convolutional layers

Challenge:

-- Network has 40 million parameters.
-- Required training data ~1-2 million.
-- Available training data ~50K.

Solution:

-- Train the network on a task with enough data present.
-- Fine-tune the network to the current task.

Fine tuning procedure:

-- Train image classification task using imagenet data of size 1.2 million.
-- Replace the softmax layer with random initialization.
-- Run the gradient descent.
Deep Poselets: Detection

Given a test image, run all the deep poselets.

- Each poselet occurs in a localized regions within a upper body detection.
- Run the classifiers on the “Expected center points of poselets”.
- This improves both the speed and accuracy.
Deep Poselets: Spatial reasoning

Problem: The three detections fired in the same area.
Deep Poselets: Spatial reasoning

Problem: The three detections fired in the same area.

Objective: Rescore detection 2 to 1 and the detections 1,3 to 0.

Solution:
For each poselet, learn regression function whose
-- Input: Scores of other poselet detections
-- Output: New score
Deep Poselets: Results

- **Evaluation measure:** Mean average precision.
- **Comparison:** Poselets are trained using HOG feature.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td>32.6</td>
</tr>
<tr>
<td>CNN before fine-tuning</td>
<td>48.6</td>
</tr>
<tr>
<td>CNN after fine-tuning</td>
<td>56.0</td>
</tr>
</tbody>
</table>
Deep Poselets: Results

AP 78.1
#positives in train set 1863

AP 40.4
#positives in train set 698

Rank 1
Rank 6
Rank 11
Rank 16

Rank 21
Rank 26
Rank 31
Rank 36

Rank 21
Rank 26
Rank 31
Rank 36
Overview

Deep Poselets

Poselet Discovery

- Cluster pose space

Training

- Train poselets using convolutional neural networks

Detection

- Detect poselets

Pose retrieval

- Given a query image

- Build Bag of Deep poselets

- Return the retrieved results
Pose Search: Indexing

For each frame in the video DB collection:

- Detect the upper body.
- Run all the poselets.
- Perform spatial reasoning.

Descriptor: Max pool the Deep Poselet detections

Index in a database

122D vector
Pose Search: Retrieval

Given a query image

Build Bag of Deep poselets

Using cosine distance, search through the database

Return the retrieved results
Pose Search: Results

Experimental setup

- Database: Test data of size 5440 is used as the database.
- Queries: All the samples in the test data are used as query.
- Evaluation metric: Mean average precision (MAP).

Methods compared against

- **Bag of visual words (BOVW)**
  - Detect sift $\rightarrow$ K means ($K = 1000$) $\rightarrow$ VQ.

- **Berkeley Poselets (BPL)**
  - Run poselets $\rightarrow$ Bag of parts.

- **Human pose estimation [1] (HPE)**
  - Run human pose estimation algorithms
  - Concatenate ($\sin(x), \cos(x)$) of all the body part angles.

Results

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOVW</td>
<td>14.2</td>
</tr>
<tr>
<td>BPL</td>
<td>15.3</td>
</tr>
<tr>
<td>HPE [1]</td>
<td>17.5</td>
</tr>
<tr>
<td>Ours</td>
<td>34.6</td>
</tr>
</tbody>
</table>

Pose Search: Results

Comparison with the state-of-the-art

- HPE [1]: 17.5
  - 75% queries < 20% AP
  - 5% queries > 50% AP
- Ours: 34.6
  - 45% queries < 20% AP
  - 25% queries > 50% AP

Graph showing average precision distribution.
Pose Search: Analysis

- Pose detection algorithms often commit to wrong pose.
- Pose search systems based on them perform poorly.

- Bag of poselets descriptor encodes multiple proposals weighted by their likelihood
- Hence it can recover when some of the detections are wrong.
Pose Search: Results

Query

Rank 1

Rank 5

Rank 10

Rank 15

Rank 20

Rank 25

Precision

Recall

AP: 59.4
Pose Search: Results

Query

Precision

Recall

Rank 1

Rank 5

Rank 10

Rank 15

Rank 20

Rank 25
Pose Search: Results

Query

Rank 1

Rank 5

Rank 10

Rank 15

Rank 20

Rank 25

AP: 40.3
Summary

• We propose a novel Deep Poselets based method for human pose search system.

• Our Deep Poselet method outperforms HOG based poselets by 25% MAP.

• Our pose retrieval method improves the performance of the current state-of-art system by 17% MAP.
Thank you.
Questions?