G-RMI Keypoints Detection

COCO Visual Recognition Challenges Workshop @ ECCV 2016

Presenter: George Papandreou (gpapan@google.com)
Team: Tyler Zhu, Nori Kanazawa, George Papandreou, Alex Toshev, Hartwig Adam, Chris Bregler, Kevin Murphy, Jonathan Tompson

Google Research and Machine Intelligence
Summary: G-RMI Keypoint Detection System

- Two-stage system:
  - Box person detector
  - Human pose estimator
- Ranked #2. AP during competition:
  - CMU-Pose: 0.605 (challenge), 0.618 (testdev)
  - G-RMI: 0.598 (challenge), 0.605 (testdev)
  - DL-61: 0.533 (challenge), 0.544 (testdev)
- AP post-competition:
  - G-RMI: 0.668 (testdev)
- Key technical aspects:
  - State-of-art person box detector
  - Pose estimator featuring highly localized keypoint activation maps
  - Effective box proposal rescoring by the pose estimator
- Comparison of using COCO-only as well as COCO + in-house data for training
System Overview

(1) Person detection

(2) Pose estimation

photo credit: Moreseth
Person Detection

1. **G-RMI** Box Detection entry trained on COCO data (Inception-ResNet Faster-RCNN model ensemble)
   - 0.584 person keypoint AP.

2. Person-specific detector trained on COCO + ImageNet + in-house dataset (single model, multi-crop)
   - 0.592 person keypoint AP.

3. Our best result (before the deadline): Union of (1) + (2)
   - 0.605 person keypoint AP.

*all results obtained before deadline on COCO testdev

More info: G-RMI detection team presentation
From Box to Pose Proposals

- **Crop extraction**
  - height = 353
  - width = 257

- **Aspect ratio normalization**
  - height/width = 353/257

- **Box enlargement**
  - Train scale factor in [1.0, 1.5]
  - Eval scale factor 1.25

- **Box enlargement**
Pose Estimation Network

- Single ImageNet-pretrained Resnet-101 producing heatmaps and offsets fully-convolutionally\(^1\)
- Dense (stride=8) feature extraction via atrous convolution\(^2\), followed by bilinear interpolation
  - 353x257 crop → 45x33 feature maps
  → 353x257 feature maps
- Intermediate supervision, similar to MPII’s DeeperCut system\(^3\)

Powerhorse: Tensorflow and TF-Slim

- Scalable distributed infrastructure
- Multi-machine + Multi-GPU training
- Rich library of SoA vision models:
  - Inception
  - ResNet
  - Inception-Resnet
  - ...

Your laptop  Datacenters  Mobile  Raspberry Pi  Tensor Processing Unit
Pose Estimation Net: Heatmap Output

- Heatmap field for each keypoint
  - 17 channels (1 within a disk around each keypoint, 0 outside)
  - Sigmoid cross entropy loss
- CNN layers 52 (intermediate) and 101 (final)

photo credit: Andrew Taylor
Pose Estimation Net: Offset Output

- Offset field towards the center of the disk
  - 34 channels for x- and y- offsets
  - Huber loss, only active within disks
- Only at CNN layer 101 (final)
Fusing Heatmap and Offset Outputs

Algo: Offset-guided mass transfer
For each point in the heatmap:
(1) Transfer its mass by the corresponding offset.
(2) Accumulate into fused activation maps.
Final Pose Prediction: Keypoint Position and Score

Pose rescoring is crucial:
0.05 AP boost compared to using Faster-RCNN box scores.
Pose Estimation-Only Results

COCO pose estimation with oracle boxes

- How well can we predict the pose, given ground truth person boxes?
- Metric: AR in mini-val given ground truth box
- 0.730 using COCO only annotations.
- 0.756 also using in-house person keypoint annotations.

MPII Single-Person task

- Sanity check for our pose estimation network:
  - Training on MPII only data
  - 89.0% PCKh@0.5 (close to state-of-art)
Full System Results: COCO Images

photo sources: top left, bottom left, right
Full System Results: COCO Images

photo sources: top left, top right, bottom
COCO Quantitative Results (Competition)

- Competition results on “Challenge” split.

<table>
<thead>
<tr>
<th>TEAM</th>
<th>AP</th>
<th>AP@.5</th>
<th>AP@.75</th>
<th>AP (M)</th>
<th>AP (L)</th>
<th>AR</th>
<th>AR@.5</th>
<th>AR@.75</th>
<th>AR (M)</th>
<th>AR (L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU-Pose</td>
<td>0.605</td>
<td>0.834</td>
<td>0.664</td>
<td>0.551</td>
<td>0.681</td>
<td>0.659</td>
<td>0.864</td>
<td>0.713</td>
<td>0.594</td>
<td>0.748</td>
</tr>
<tr>
<td>G-RMI</td>
<td>0.598</td>
<td>0.81</td>
<td>0.651</td>
<td>0.567</td>
<td>0.667</td>
<td>0.664</td>
<td>0.865</td>
<td>0.712</td>
<td>0.618</td>
<td>0.726</td>
</tr>
<tr>
<td>DL-61</td>
<td>0.533</td>
<td>0.751</td>
<td>0.485</td>
<td>0.555</td>
<td>0.548</td>
<td>0.708</td>
<td>0.828</td>
<td>0.688</td>
<td>0.74</td>
<td>0.782</td>
</tr>
<tr>
<td>R4D</td>
<td>0.497</td>
<td>0.743</td>
<td>0.545</td>
<td>0.456</td>
<td>0.556</td>
<td>0.556</td>
<td>0.773</td>
<td>0.603</td>
<td>0.491</td>
<td>0.644</td>
</tr>
<tr>
<td>umich_vl7</td>
<td>0.434</td>
<td>0.722</td>
<td>0.449</td>
<td>0.364</td>
<td>0.534</td>
<td>0.499</td>
<td>0.758</td>
<td>0.52</td>
<td>0.387</td>
<td>0.652</td>
</tr>
</tbody>
</table>
COCO Quantitative Results (Post Competition)

- Latest results on “testdev” split.

<table>
<thead>
<tr>
<th>TEAM</th>
<th>AP</th>
<th>AP@.5</th>
<th>AP@.75</th>
<th>AP (M)</th>
<th>AP (L)</th>
<th>AR</th>
<th>AR@.5</th>
<th>AR@.75</th>
<th>AR (M)</th>
<th>AR (L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-RMI (competition entry)</td>
<td>0.605</td>
<td>0.822</td>
<td>0.662</td>
<td>0.576</td>
<td>0.666</td>
<td>0.662</td>
<td>0.866</td>
<td>0.714</td>
<td>0.619</td>
<td>0.722</td>
</tr>
<tr>
<td>G-RMI (post competition)</td>
<td>0.668</td>
<td>0.863</td>
<td>0.734</td>
<td>0.630</td>
<td>0.733</td>
<td>0.716</td>
<td>0.896</td>
<td>0.776</td>
<td>0.669</td>
<td>0.782</td>
</tr>
</tbody>
</table>

- Bug discovered: Aspect ratio mismatch between train and eval code.
- Added OKS-based non-maximum suppression on the pose result.
COCO Quantitative Results (Post Competition)

- Latest results on “testdev” split.
- AP using the G-RMI Object Detection team’s boxes trained only on COCO box annotations.
- Effect of training the pose estimator on COCO-only pose annotations vs. COCO+in-house pose annotations.
- Effect of OKS-based Non-Maximum Suppression.

<table>
<thead>
<tr>
<th></th>
<th>No OKS-based NMS</th>
<th>With OKS-based NMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO-only</td>
<td>0.601</td>
<td>0.636</td>
</tr>
<tr>
<td>COCO+in-house</td>
<td>0.628</td>
<td>0.668</td>
</tr>
</tbody>
</table>
Lessons from COCO Keypoints Challenge

- New dataset, new metric, first year the challenge runs
- Challenging problem:
  - Person detection and pose estimation
- Important differences compared to previous pose datasets:
  - Many example persons with severe occlusion
  - Large scale variations (and scale is not considered known)
- Lots of room for improvement!
- Interesting research problems:
  - Example: Two-stage or single-stage system?
- Pose estimation is becoming mature for in-the-wild deployment!
David Hallberg Dance Sequence Results
David Hallberg 3D Trajectory Reconstruction
Thanks!

- **Person detection and pose estimation team**
  - Tyler Zhu, Nori Kanazawa, George Papandreou, Alex Toshev, Hartwig Adam, Chris Bregler, Kevin Murphy, Jonathan Tompson

- **G-RMI object detection team**
  - Alireza Fathi, Ian Fischer, Sergio Guadarrama, Jonathan Huang, Anoop Korattikara, Kevin Murphy, Vivek Rathod, Yang Song, Chen Sun, Zbigniew Wojna, Menglong Zhu

- **TF-slim**
  - Sergio Guadarrama, Nathan Silberman

- **Person annotation team**
  - Akshay Gogia, Gursheesh Kour, Manish Arora

- **Special thanks:**
  - Georgia Gkioxari, Dumitru Erhan, Jitendra Malik, Chuck Rosenberg, Rahul Sukthankar, Jay Yagnik