Biposelets for Human Pose Estimation

Sam Toyer

Supervisor: Anoop Cherian, RSCS
Roadmap

- Motivation
- Existing approaches
- My contribution
- Experiments
- Conclusion and future work
Problem setup (II)

Why?

3D pose estimation
Andriluka, Roth & Schiele. Monocular 3D pose estimation and tracking by detection. CVPR ’10

Describing objects
Delaitre, Fouhey, Laptev, Sivic, Gupta & Efros. Scene semantics from long-term observation of people. In ECCV ’12

Action recognition
**Existing work**


**Graphical models**

- Find joint locations with sensible appearance, but also make sure distances between joints are realistic
- For videos: add constraints on movement between frames
- In practice, evaluate frames independently then stitch results

**Joint heat maps**

- Convolutional neural network (CNN) produces heat map for each joint
- Network takes image patch as input, produces heatmap for patch directly
- Using videos: open problem

My contribution

• New approach: look at two frames at a time, use CNN to produce heatmap for each sub-pose instead of each joint

• Two-frames-at-a-time formulation enables use of temporal information at CNN level

• Preliminary results: competitive with temporally-aware baseline
• CNNs are good at classification
• Existing approaches try to classify individual joints
• Instead, try to classify entire subposes
• CNN gets to see context of surrounding joints
Sub-pose classification

- Do K-means on all instances of each sub-pose in dataset, find joint location centroids
  - Centroids referred to as poselets
- Sliding window classifier determines which poselet is in each patch
- Use expectation over poselets w.r.t. CNN output distribution to find joint locations
Biposelets

- Take two frames and the flow (per-pixel displacement) between them
- Output a poselet for both frames at once
- A poselet spanning two frames is called a *biposelet*
- Biposelets allow the CNN to learn relationships between frames
From biposelets to poses

- CNN gives us heatmap for each subpose (above: head, left and right arms) and each biposelet for that subpose
- Use SVM to find locations and biposelets for each subpose; gives pose for each frame
- Keep top K instead of just best pose
Putting it all together

CNN

Biposelet 0
0.87
Biposelet 1
0.08
Biposelet 2
0.05

Final pose for each frame
Results: MPII Cooking Activities (I)

<table>
<thead>
<tr>
<th></th>
<th>Upper arm</th>
<th>Lower arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cherian et al.</td>
<td>80.9%</td>
<td>75.9%</td>
</tr>
<tr>
<td>Biposelets</td>
<td>87.4%</td>
<td>85.4%</td>
</tr>
</tbody>
</table>

Above: Percentage Correct Keypoints (PCK) at image scale

Left: Percentage Correct Parts (PCP); strict variant
**Top:** A well-localised sequence. **Bottom:** Failure modes.
Conclusion

- New approach to video pose estimation:
  - Casts pose estimation as problem of classifying biposelets
  - Biposelets allow temporal relationships to be learnt at CNN-level
- Results competitive with temporally-aware past work

Future work

- Improve current approach: fine-tune training process, decrease subpose size, improve expressiveness of graphical model
- Address lack of video pose estimation datasets