Deep neural nets for human pose estimation in videos

Tomas Pfister, James Charles, Andrew Zisserman
Department of Engineering Science
University of Oxford
http://www.robots.ox.ac.uk/~vgg
Aim:

Estimate 2D upper body joint positions (wrist, elbow, shoulder, head) with high accuracy in real-time.
Outline

• Two types of loss functions for pose estimation
  • Coordinate net
  • Heatmap net

• Optical flow for pose estimation in videos

• Results (cf state of the art)
Method overview: single frame learning

1. Coordinate Net

   e.g. DeepPose CVPR14, Pfister et al ACCV14

2. Heatmap Net

   e.g. Jain et al ICLR14, Tompson et al CVPR15
Coordinate Net: regress joint positions

Training loss: L2 on joint positions

OverFeat like architecture
Heatmap Net: regress heatmap for each joint

Represent joint position by Gaussian

Training loss: L2 on pixels
Comparison

Coordinate Net

Heatmap Net

Regression target

Coordinates

Heatmap
BBC sign language videos data set

Training:
15 videos each 0.5-1hr long, all frames annotated

Testing:
5 videos, 200 annotated frames per video

Extended Training:
72 videos with noisy automated annotations
Results on architecture comparison

- Heatmap net superior to coordinate net
- Performance of coordinate net saturates with more training data

Evaluated on BBC Pose
Why is the heatmap network superior?

1. Can represent multimodal estimates, so can model uncertainty/confidence

2. In training there is an error signal from every pixel, so better smoothing for back propagation

Also, it is easier to visualize (and understand) what is being learnt
Timelapse of training
Multiple modes example

early in training

late in training
What do the layers learn?

Three randomly selected activations from each layer
Learning from videos

• Temporal information

  – How do we learn from temporal information with a ConvNet?

Hand moving in x direction
Late fusion using flow

**Warp the heatmaps** from previous/next frames & **combine**

Charles et al., Upper Body Pose Estimation with Temporal Sequential Forests, BMVC 2014
Optical flow
Example: Heatmap Net & Optical flow

Tracks for optical flow for wrist positions

Flow: Brox et al GPU flow from OpenCV, or FastDeepFlow
Optical flow
Example: Heatmap Net & Optical flow

Warping heatmaps to frame t
Flowing ConvNets

- Learn the **pooling** of the warped heatmaps
Results: with/without optical flow

Method: Flowing ConvNets

Input
Results
Comparison of pooling types

![Graph showing comparison of pooling types for Wrist dataset. The x-axis represents the neighbourhood size (n), while the y-axis represents accuracy at d=5px. The graph compares different pooling methods: Sumpool, Maxout, and Parametric pooling. The accuracy varies with the neighbourhood size, indicating the effectiveness of each method under different conditions.](image-url)
Results
Learnt optical flow pooling weights

elbow

wrist
Results
Comparison to the state of the art

Poses in the Wild

12% improvement at $d = 10$px
Results: Example pose estimation

50fps on 1 GPU without optical flow, 5fps with optical flow
Results

Failure cases

Main failure case: Picking the wrong mode

BBC Pose

Correctable with a spatial model

ChaLearn
Additional Pooling Fusion Layers

Heatmap Pose ConvNet

Conv A 8x8x64
Conv B 13x13x64
Conv C 15x15x64
Conv D 1x1x128
Conv E 1x1x7

Implicit spatial model
Results: Additional Pooling Fusion Layers

Poses in the Wild

Heat map

CNNs

with fusion
and flow
with fusion
original

Accuracy [%]

Distance from GT [px]

Cherian et al. (2014)
Yang & Ramanan (2013)
SpatialNet Fusion Flow
SpatialNet Fusion
SpatialNet
Results: Additional Pooling Fusion Layers

FLIC: single image predictions

Average PCK for wrist & elbow

Accuracy [%]

Normalised distance from GT
Summary

• Deep Heatmap ConvNet achieves state of the art with implicit spatial models

• Performance improved by optical flow pooling

• Futures:
  – Robust regression
  – Data dependent flow channel pooling
  – More training data