

Deep Convolutional Neural Networks for Efficient Pose Estimation in Gesture Videos

Goal

Reliable **2D upper body pose estimation** in long videos in real-time for gesture recognition.



BBC TV Sign Language

Motivation

Address problems with previous approaches:

Unconstrained poses

Need foreground segmentation

Speed



2 fps

Temporal Pose ConvNet



SEG. NOT REQUIRED

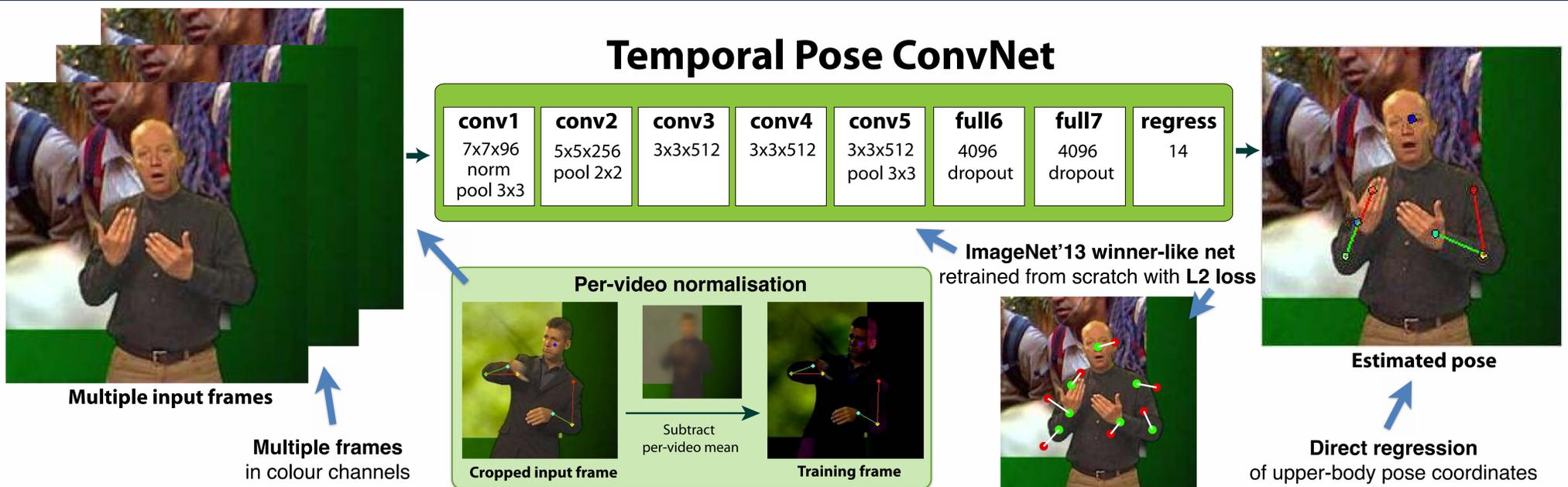
100 fps

Contributions

1. A pose estimation ConvNet for **learning from temporal information**
2. Investigation of search space reduction by **pre-segmenting the video foreground**
3. Benefits of estimating the pose from **multiple input frames**

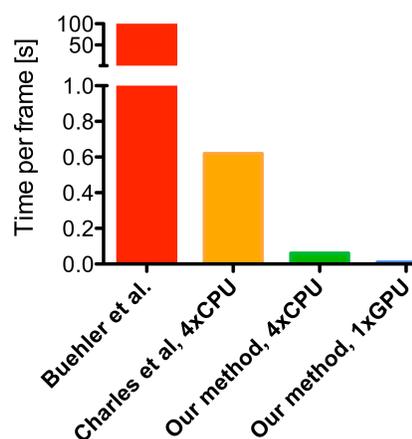
Method

Temporal Pose ConvNet



Experiments

Computation time



Dataset

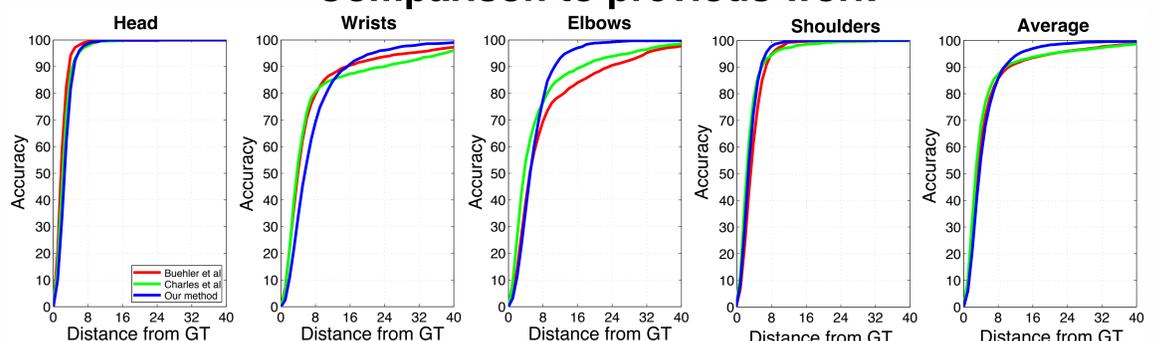
20 hours of video (15h train, 5h test) with automatically generated ground truth

Component evaluation

Evaluation measure: % of pixels within 6px of GT (scale on left)

Training	Augmentation	Multi-frame	Segmentation	Head	Wrists	Elbows	Shoulder	Average
Last layer only (ImNet)	✓			15.4	5.8	8.4	18.3	12.0
Finetune all layers (ImNet)	✓			95.6	44.0	53.6	80.8	68.5
Train from scratch				94.3	52.1	51.9	87.9	71.5
Train from scratch	✓			95.9	47.1	56.0	89.1	72.0
Train from scratch	✓	✓		95.6	50.1	58.1	89.5	73.3
Train from scratch	✓		✓	96.1	58.0	66.8	91.2	78.0
Train from scratch	✓	✓	✓	96.1	59.3	66.5	91.2	78.3

Comparison to previous work



[1] Buehler, P., Everingham, M., Huttenlocher, D.P., Zisserman, A.: Upper body detection and tracking in extended signing sequences. IJCV (2011)
 [2] Charles, J., Pfister, T., Everingham, M., Zisserman, A.: Automatic and efficient human pose estimation for sign language videos. IJCV (2013)