JOINT TRAINING OF A CONVOLUTIONAL NETWORK AND A GRAPHICAL MODEL FOR HUMAN POSE ESTIMATION (NIPS’14)

JONATHAN TOMPSON
PRESENTED BY TOMAS

ADVISOR: CHRISTOPH BREGLER
**HUMAN BODY POSE INFERENCE**

Problem definition:

Track human body joints on a monocular RGB image
Arbitrary pose and background
WHY DIFFICULT...
TWO WAYS TO ESTIMATE POSE...

1. Coordinate Net

Input frame(s) ➞ Coordinate ConvNet ➞ Estimated pose

2. Heatmap Net

Input frame(s) ➞ Heatmap ConvNet ➞ Heatmap ➞ Estimated pose
Regression Target

Gaussian centered at each joint (w/ fixed variance)

One heatmap per joint

L2 loss (avg over x,y & joints)
Basic Idea of This Paper

Two Parts:

- ConvNet to track joints
- Graphical Model to stitch them together

Jointly train them!
**PART DETECTOR OVERVIEW**

Full Image 320x240px

Half-res Image 160x120px

Conv + ReLU + Pool (3 stages)

Conv + ReLU + Pool (3 stages)

98x68x128

90x60x512

1x1conv + ReLU

9x9 Conv + ReLU

Point-wise Upscale

90x60x128

45x30x128

90x60x512

Fully-connectionned equivalent model

90x60x4
WHAT IS THE ARCHITECTURE ACTUALLY

Their CVPR’15 basic model (slightly tweaked architecture)
PART DETECTOR RESULTS

What’s wrong so far

Independent joint terms in objective function
We’re hoping the network implicitly learns joint consistency
Failure cases are usually stupid:
\[
\bar{p}_A = \frac{1}{Z} \prod_{v \in V} (P_{A|v} * P_v + b_{v \rightarrow A})
\]
**Spatial Model**

Two additional details

1. Spatial model kernel size is 128x128! → Have to use FFT\(^1\)
2. For standard datasets → add (noisy) torso location

---

**Spatial Model**

Implement it as a network

\[
\bar{p}_A = \frac{1}{Z} \prod_{v \in V} (p_{A|v} * p_v + b_{v \to A})
\]

\[
\bar{e}_A = \exp \left( \sum_{v \in V} \left[ \log (\text{SoftPlus} (e_{A|v}) * \text{ReLU} (e_v) + \text{SoftPlus} (b_{v \to A})) \right] \right)
\]

where: \( \text{SoftPlus} (x) = \frac{1}{\beta} \log (1 + \exp (\beta x)) , \frac{1}{2} \leq \beta \leq 2 \)

\( \text{ReLU} (x) = \max (x, \epsilon) , 0 < \epsilon \leq 0.01 \)
JOINT TRAINING

Joint training

Pre-train both models separately
Joint train (BPROP) through both models
J**OINT TRAINING**

Joint training

- Pre-train both models separately
- Joint train (BPROP) through both models

Performance (Wrist):

\[
\text{DetectionRate} (R) = \frac{100}{N} \sum_{t=1}^{N} \left( \frac{\|x - x^t\|_2}{\text{torso height } t} / 100 \leq R \right)
\]
**JOINT TRAINING**

Joint training

Pre-train both models separately

Joint train (BPROP) through both models

**Performance (Wrist):**

\[
\text{DetectionRate} (R) = \frac{100}{N} \sum_{t=1}^{N} \left( \frac{\|x - x^t\|_2}{\text{torso height } t} \right) / 100 \leq R
\]
RESULTS

FLIC\(^{(1)}\)
Elbow

LSP\(^{(2)}\)
Arms

FLIC\(^{(1)}\)
Wrist

LSP\(^{(1)}\)
Legs

(1) B. Sapp and B. Taskar. MODEC: Multimodel decomposition models for human pose estimation. CVPR’13
(2) S. Johnson and M. Everingham. Learning Effective Human Pose Estimation for Inaccurate Annotation. CVPR’11
MULTICAMERA RESULTS

From their CVPR’15 paper
RESULTS

Joint work with MPII

Use our detector + analysis by synthesis technique
RESULTS

Joint work with MPII

Use our detector + analysis by synthesis technique