Boulder

Computer Vision; Pose Estimation

Maziar Raissi

Assistant Professor

Department of Applied Mathematics

University of Colorado Boulder

maziar.raissi@colorado.edu

Convolutional Pose Machines

Pose Machines

$Y_p \in \mathcal{Z} \in \mathbb{R}^2 \rightarrow$ pixel location of the p-th (part: anatomical landmark)

$\mathcal{Z} \rightarrow$ set of all locations $z = (u, v)$ in an image

$Y = (Y_1, Y_2, \ldots, Y_P) \rightarrow$ image locations for all P parts (to be predicted)

$g_t(\cdot), t = 1, \ldots, T \rightarrow$ sequence of multi-class predictors

$t \rightarrow$ stage

$g_t \rightarrow$ predicts beliefs for assigning a location to each part (i.e., $Y_p = z, \forall z$)

$x_z \in \mathbb{R}^d \rightarrow$ features extracted from the image at location z

$g_1 : x_z \mapsto \{b_1^p(Y_p = z)\}_{p=1,\ldots,P+1}$ ($P+1$: background)

$b_1^p(Y_p = z)$: score for assigning the p-th part at image location z in the first stage

$b_t^p \in \mathbb{R}^{w \times h}, b_t^p(u, v) = b_t^p(Y_p = z)$ ($(u,v) = z$)

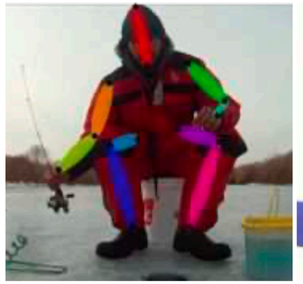
$b_t \in \mathbb{R}^{w \times h \times (P+1)}$

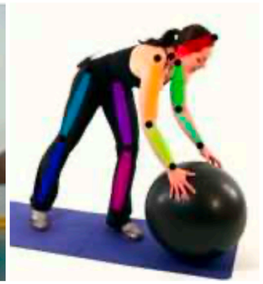
In subsequent stages ($t > 1$):

$g_t : (x'_z, \psi_t(z, b_{t-1})) \mapsto \{b_t^p(Y_p = z)\}_{p=1,\ldots,P+1}$

$\psi_t(z, b_{t-1})$: context features

boosted random forest and hand-crafted image & context features!

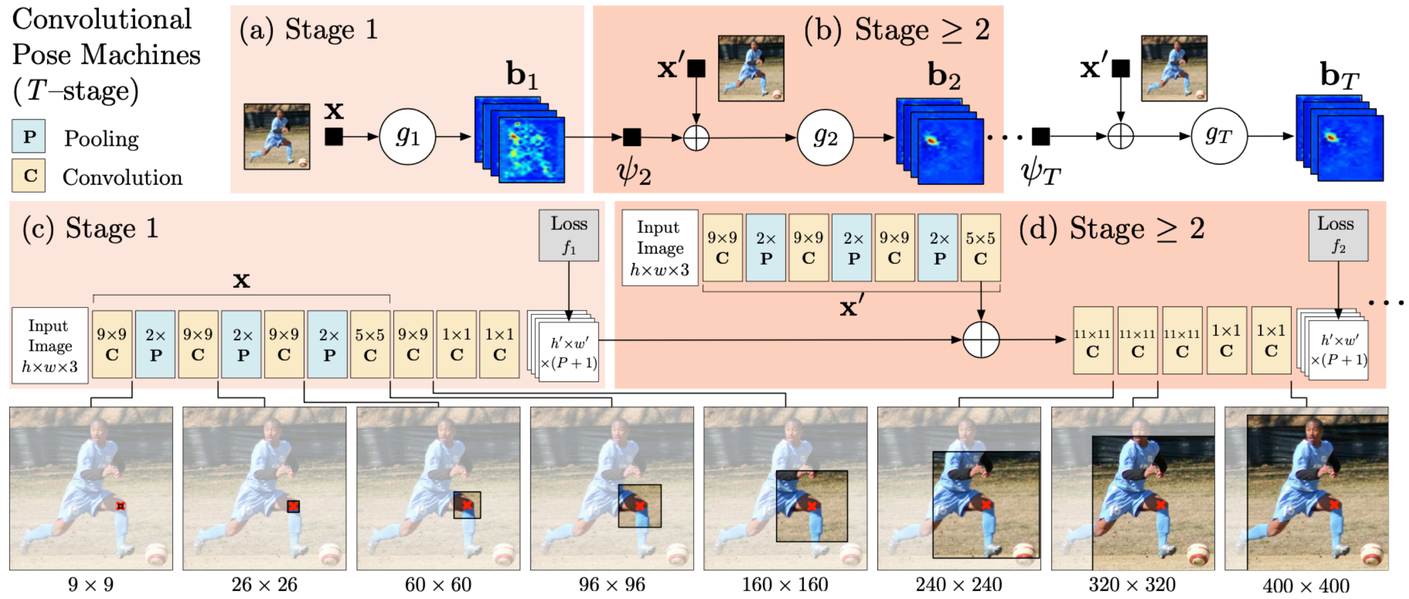


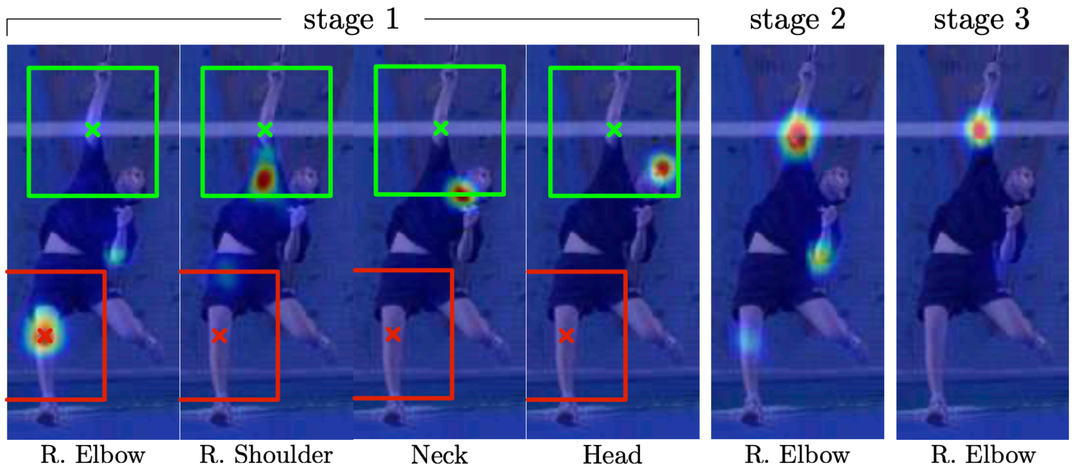


Convolutional Pose Machines

$$\mathcal{F} = \sum_{t=1}^{T} f_t, \; f_t = \sum_{p=1}^{P+1} \sum_{z \in \mathcal{Z}} |b_t^p(z) - b_*^p(z)|^2$$

ideal belief map for a part p, $b_*^p(Y_p = z)$: putting Gaussian peaks at ground truth locations of each body part p



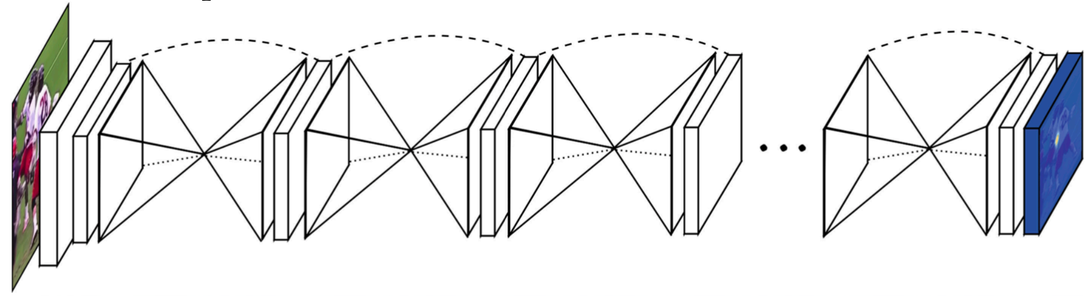


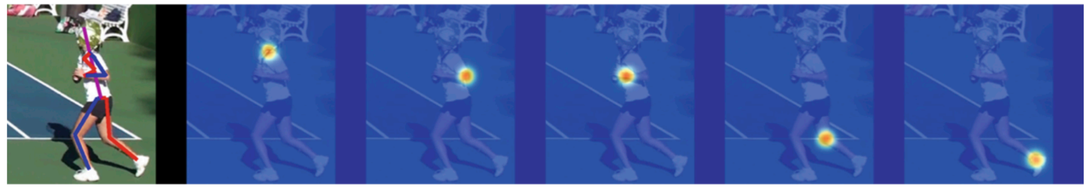
Wei, Shih-En, et al. "Convolutional pose machines." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016.

Stacked Hourglass Networks for Human Pose Estimation

Boulder

Human-computer interaction and animation

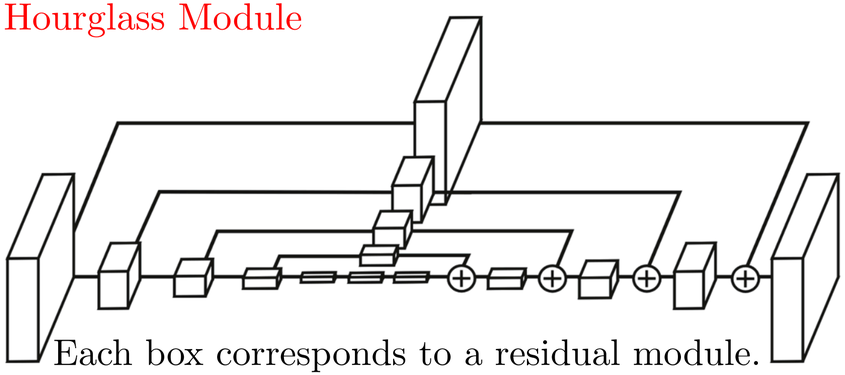




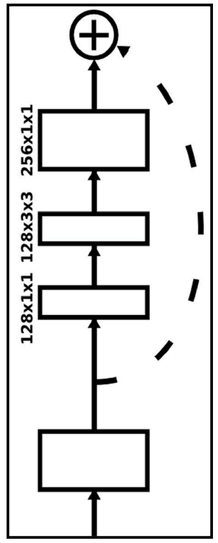
Motivation: Capturing information at every scale!

"Local evidence is essential for identifying features like faces and hands, while a final pose estimate requires a coherent understanding of the full body."

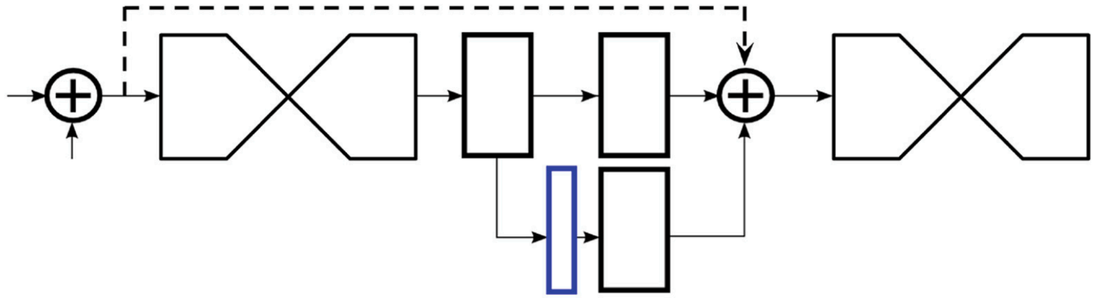
Hourglass Module



Each box corresponds to a residual module.



Intermediate Supervision

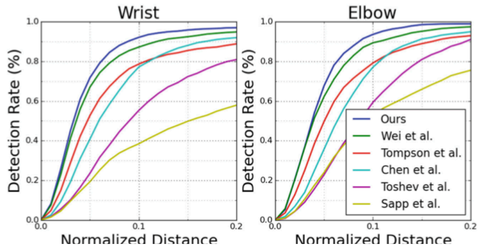


MSE Loss (predicted v.s. ground truth heatmaps)

Gaussian

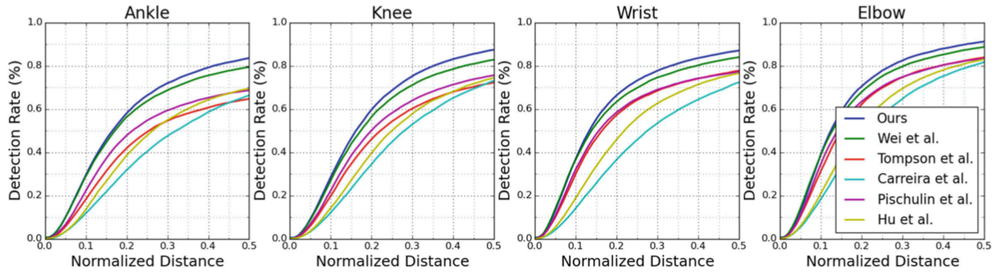
Percentage of Correct Keypoints (PCK): Percentage of detections that fall within a normalized distance of the ground truth.

FLIC Results



	Elbow	Wrist
Sapp et al. [1]	76.5	59.1
Toshev et al. [24]	92.3	82.0
Tompson et al. [16]	93.1	89.0
Chen et al. [25]	95.3	92.4
Wei et al. [18]	97.6	95.0
Our model	**99.0**	**97.0**

MPII Results

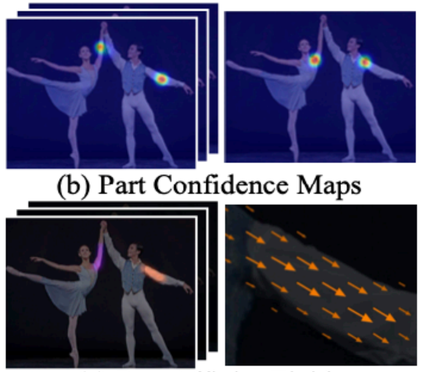


Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." *European conference on computer vision*. Springer, Cham, 2016.

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields

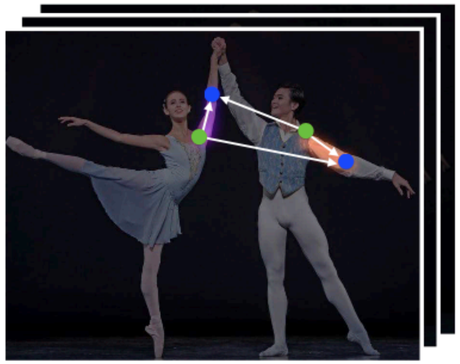
Human 2D pose estimation → localizing anatomical keypoints or "parts"

(a) Input Image



(b) Part Confidence Maps

(c) Part Affinity Fields



(d) Bipartite Matching

(e) Parsing Results

$S \rightarrow$ set of 2D confidence maps of body part locations

$L \rightarrow$ set of 2D vector fields of part affinities
(encode the degree of association between parts)

$S = (S_1, \ldots, S_J)$ has J confidence maps (one per part)

$S_j \in \mathbb{R}^{w \times h}$

$L = (L_1, \ldots, L_C)$ has C vector fields (one per limb)

$L_c \in \mathbb{R}^{w \times h \times 2} \rightarrow$ each image location encodes a 2D vector

$F \leftarrow$ image $\triangleright$ VGG-19 (first 10 layers)

$$S^1 = \rho^1(F) \qquad S^t = \rho^t(F, S^{t-1}, L^{t-1}) \text{ for } t \geq 2$$

$$L^1 = \phi^1(F) \qquad L^t = \phi^t(F, S^{t-1}, L^{t-1}) \text{ for } t \geq 2$$

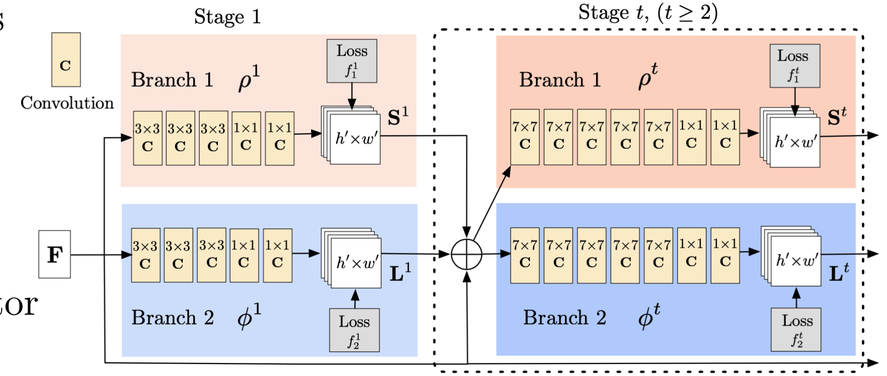
$W(p) = 0$ when the annotation is missing at an image location p

$$f_S^t = \sum_{j=1}^{J} \sum_{p} W(p) \| S_j^t(p) - S_j^*(p) \|_2^2$$

loss

$$f_L^t = \sum_{c=1}^{C} \sum_{p} W(p) \| L_c^t(p) - L_c^*(p) \|_2^2$$

$$f = \sum_{t=1}^{T} (f_S^t + f_L^t)$$

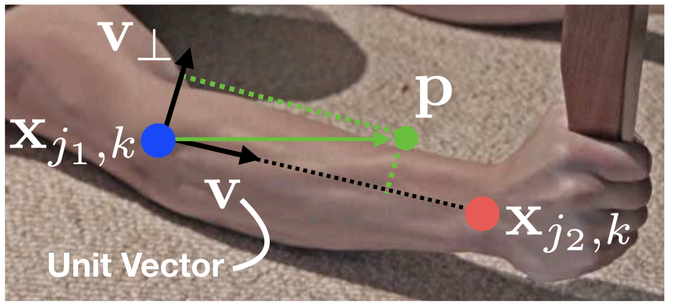


$S_j^*(p) = \max_k S_{j,k}^*(p) \rightarrow$ groundtruth confidence map

$S_{j,k}^*(p) \rightarrow$ confidence map for person k

$$S_{j,k}^*(p) = \exp\left(-\frac{\| p - x_{j,k} \|_2^2}{\sigma^2} \right)$$

$x_{j,k} \in \mathbb{R}^2 \rightarrow$ groundtruth position of body part j for person k in the image



$$L_c^*(p) = \frac{1}{n_c(p)} \sum_k L_{c,k}^*(p)$$

groundtruth part affinity field

$n_c(p) \rightarrow$ number of non-zero vectors at point p across all k people

$L_{c,k}^*(p) = v$ if p on limb c, k
$= 0$ otherwise

$$0 \leq \mathbf{v} \cdot (\mathbf{p} - \mathbf{x}_{j_1,k}) \leq l_{c,k} \text{ and } |\mathbf{v}_\perp \cdot (\mathbf{p} - \mathbf{x}_{j_1,k})| \leq \sigma_l$$

$$l_{c,k} = \| \mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k} \|_2$$

Testing

$d_{j_1}, d_{j_2} \rightarrow$ two candidate part locations

$p(u) = (1-u) d_{j_1} + u d_{j_2}$

$$E = \int_{u=0}^{u=1} L_c(p(u)) \cdot \frac{d_{j_2} - d_{j_1}}{\| d_{j_2} - d_{j_1} \|_2} du$$

association confidence

Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Boulder

Questions?