



STRUCTURED OUT PREDICTION (POSE)

HYUN SOO PARK

CHALLENGES OF VISUAL RECOGNITION

- Appearance
 - DOF: texture, illumination, material, shading, ...
- Shape
 - DOF: object category, geometric pose, viewpoint, ...



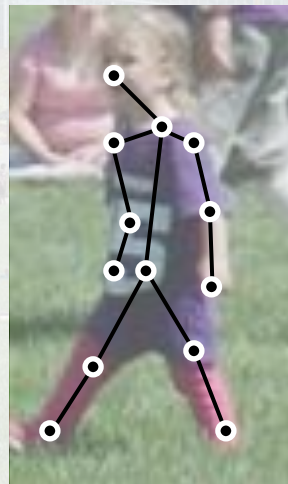
CHALLENGES OF VISUAL RECOGNITION

- Appearance
 - DOF: texture, illumination, material, shading, ...
- Shape
 - DOF: object category, geometric pose, viewpoint, ...



Human

$$f(I) = l_{human}$$



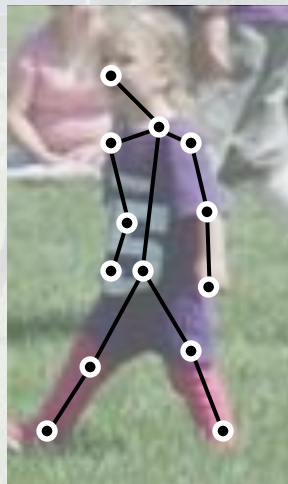
Human pose

$$f(I) = \begin{bmatrix} x_1 \\ y_1 \\ \vdots \\ x_n \\ y_n \end{bmatrix}$$

Structured output

CHALLENGES OF VISUAL RECOGNITION

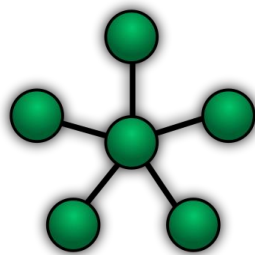
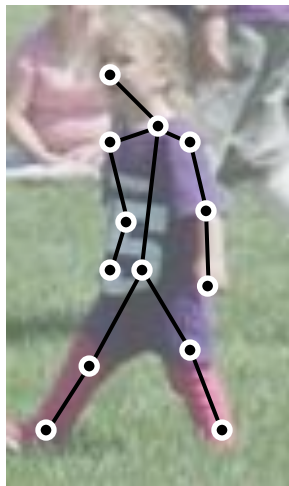
- Appearance
 - DOF: texture, illumination, material, shading, ...
- Shape
 - DOF: object category, geometric pose, viewpoint, ...



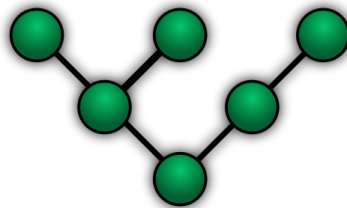
Landmark localization (**structured prediction**)

Learning appearance and spatial relation together

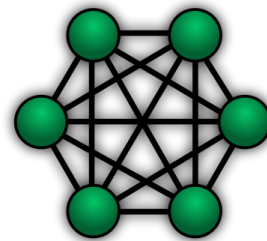
TOPOLOGICAL MODEL



Star topology

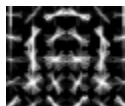


Tree topology



Fully connected topology

PROBLEM DEFINITION

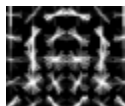
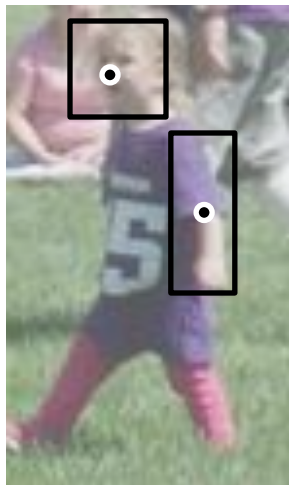


Head detector

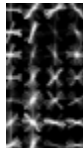
$$w_{head} \cdot \phi(I, x_{head}) > 0$$

Feature extractor

PROBLEM DEFINITION



Head detector

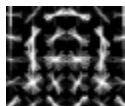
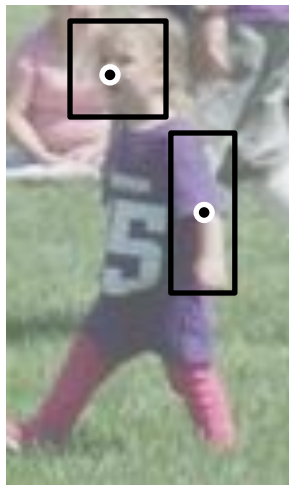


Arm detector

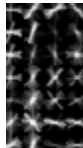
$$w_{head} \cdot \underbrace{\phi(I, x_{head})}_{\text{Feature extractor}} > 0$$

$$w_{arm} \cdot \phi(I, x_{arm}) > 0$$

PROBLEM DEFINITION



Head detector



Arm detector

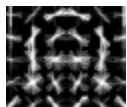
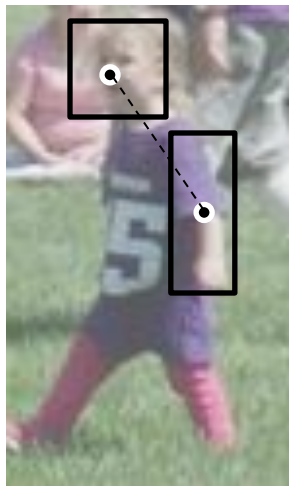
Objective:

$$\frac{w_{head} \cdot \phi(I, x_{head})}{\text{Feature extractor}} > 0$$

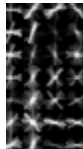
$$w_{arm} \cdot \phi(I, x_{arm}) > 0$$

$$L = w_h \cdot \phi(I, x_h) + w_a \cdot \phi(I, x_a)$$

PROBLEM DEFINITION



Head detector



Arm detector

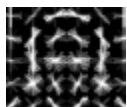
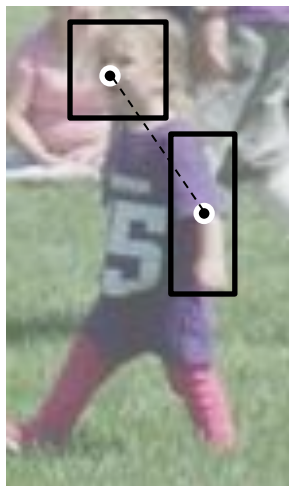
Objective:

$$\frac{w_{head} \cdot \phi(I, x_{head})}{\text{Feature extractor}} > 0$$

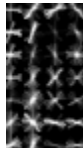
$$w_{arm} \cdot \phi(I, x_{arm}) > 0$$

$$L = w_h \cdot \phi(I, x_h) + w_a \cdot \phi(I, x_a) + \underbrace{\varphi(x_h, x_a)}_{\text{Spatial relationship}}$$

PROBLEM COMPLEXITY



Head detector



Arm detector

Objective:

$$w_{head} \cdot \phi(I, x_{head}) > 0$$

Feature extractor

of configurations

$$O(h^2)$$

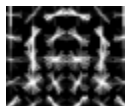
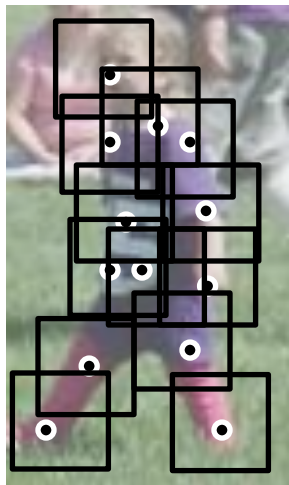
h : # of pixels

$$w_{arm} \cdot \phi(I, x_{arm}) > 0$$

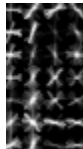
$$L = w_h \cdot \phi(I, x_h) + w_a \cdot \phi(I, x_a) + \underbrace{\phi(x_h, x_a)}$$

Spatial relationship

PROBLEM COMPLEXITY



Head detector



Arm detector

Objective:

$$w_{head} \cdot \phi(I, x_{head}) > 0$$

Feature extractor

of configurations

$$O(h^n)$$

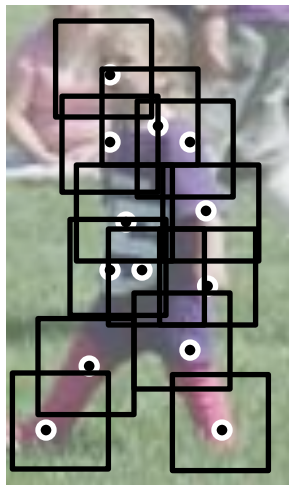
h : # of pixels

n : # of landmarks

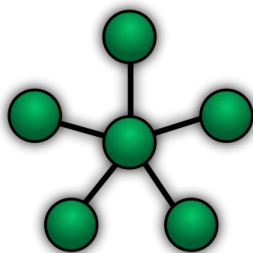
$$w_{arm} \cdot \phi(I, x_{arm}) > 0$$

$$L = \sum_i w_i \cdot \phi(I, x_i) + \sum_{i,j} \phi(x_i, x_j)$$

MODEL TOPOLOGY

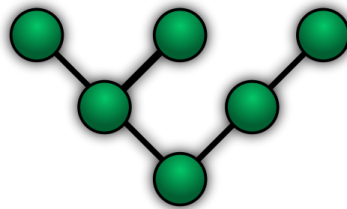


$$O(nh^2)$$



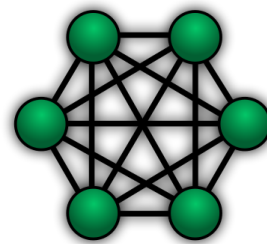
Star topology

$$O(nh^2)$$



Tree topology

$$O(h^n)$$



Fully connected topology

Can be solved by dynamic programming

NP-hard

CVPR 2008

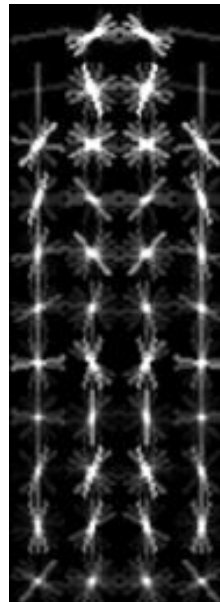
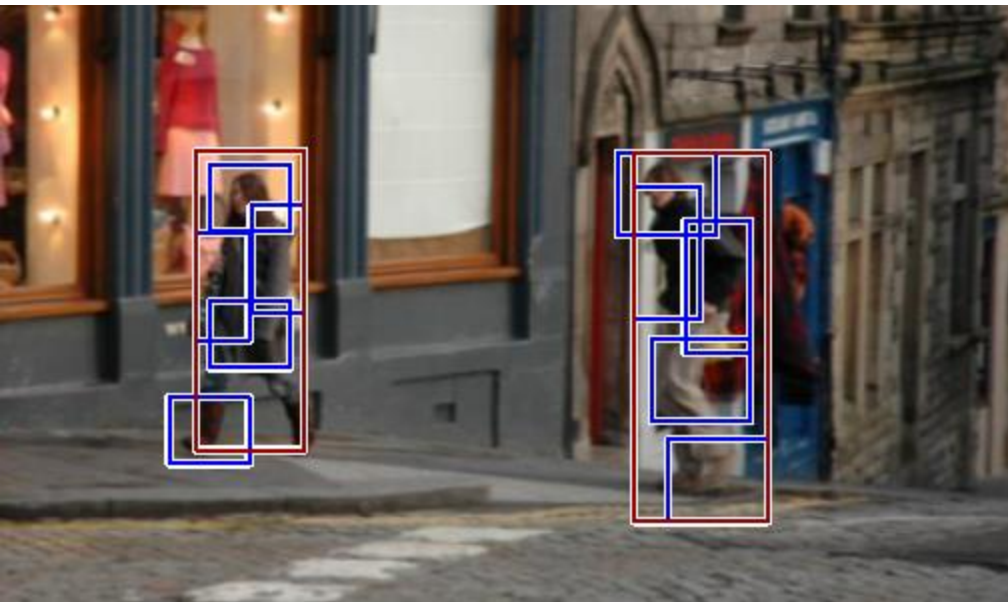
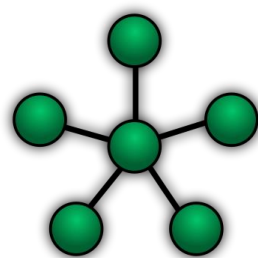
2018 Longuet-Higgins Prize for fundamental contributions in computer vision

Object Detection with Discriminatively Trained Part Based Models

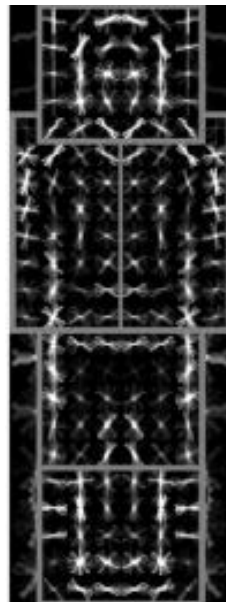
Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan

Slide inspired by Felzenszwalb and Girshick

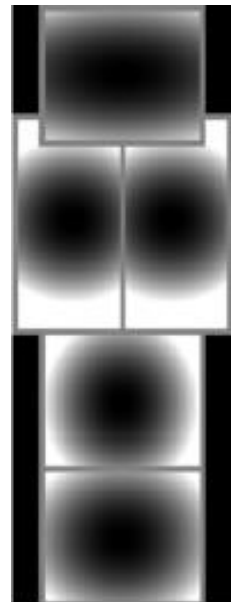
DEFORMABLE PART MODEL (DPM)



Root filter

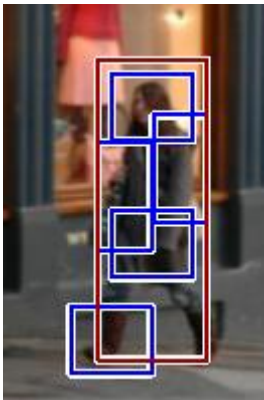


Part desc.

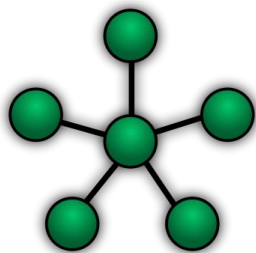


Deformation

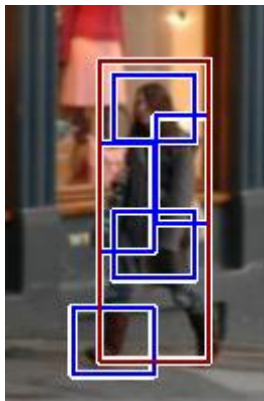
DPM



$$L = \sum_i w_i \cdot \phi(I, x_i) + \sum_{i,j} \varphi(x_i, x_j)$$

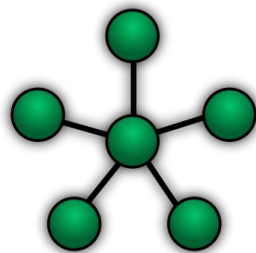


DPM

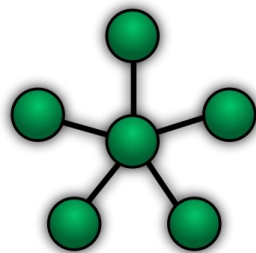
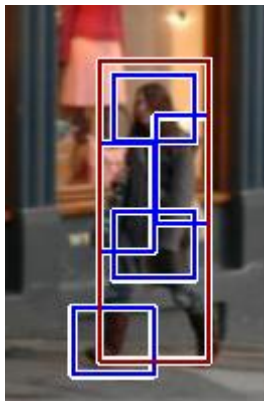


$$L = \sum_i w_i \cdot \phi(I, x_i) + \sum_{i,j} \phi(x_i, x_j)$$

$$score(x_0, x_1, \dots, x_n) = \underbrace{\sum_{i=0} w_i \cdot \phi(I, x_i) - \sum_{i=1} d_i \cdot (dx_i^2)}_{\text{Appearance (filter response)}}$$



DPM

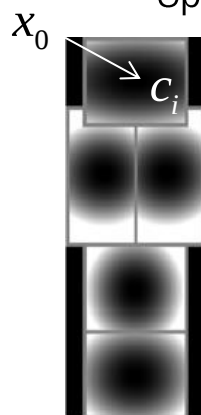


$$L = \sum_i w_i \cdot \phi(I, x_i) + \sum_{i,j} \varphi(x_i, x_j)$$

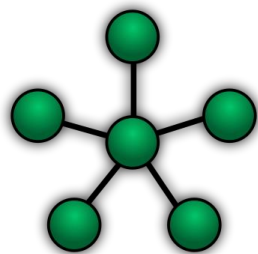
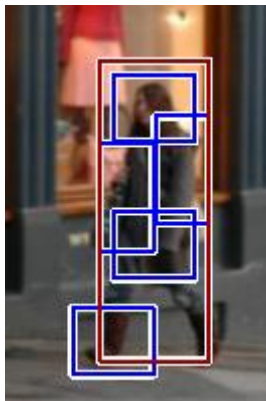
$$score(x_0, x_1, \dots, x_n) = \sum_{i=0} w_i \cdot \phi(I, x_i) - \sum_{i=1} d_i \cdot (dx_i^2)$$

Spatial rel. (spring force)

where $dx_i = x_0 + \underbrace{c_i}_{\text{Offset}} - x_i$

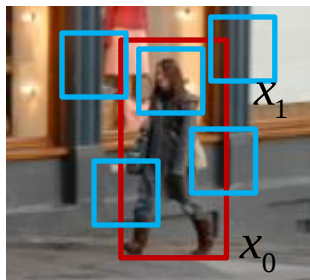


DPM



Score for root location based on part location

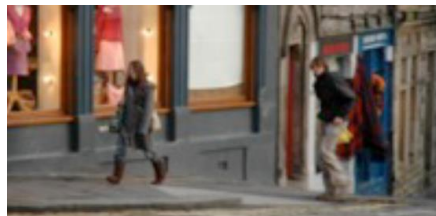
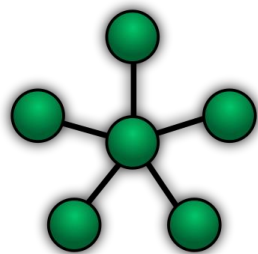
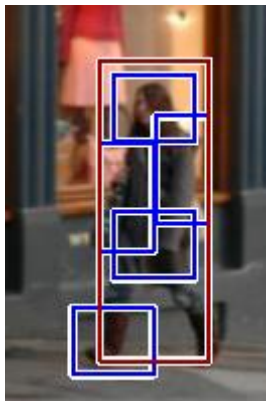
$$\begin{aligned} score(x_0) &= \max_{x_1, \dots, x_n} score(x_0, x_1, \dots, x_n) \\ &= w_0 \cdot \phi(I, x_0) + \sum_{i=1} \max_{x_i} score(x_0, x_i) \\ &= w_0 \cdot \phi(I, x_0) + \underbrace{\sum_{i=1} \max_{x_i} (w_i \cdot \phi(I, x_i) - d_i \cdot dx^2)} \end{aligned}$$



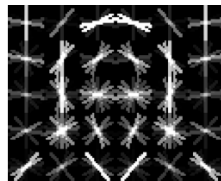
Part evidence to predict the root

How does the head location tell us about the root location?

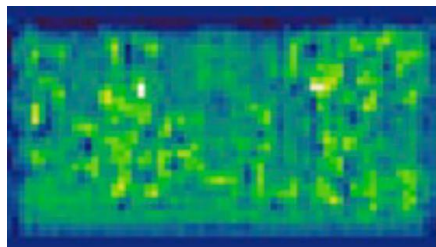
DPM



Input image



Head filter



Filter response

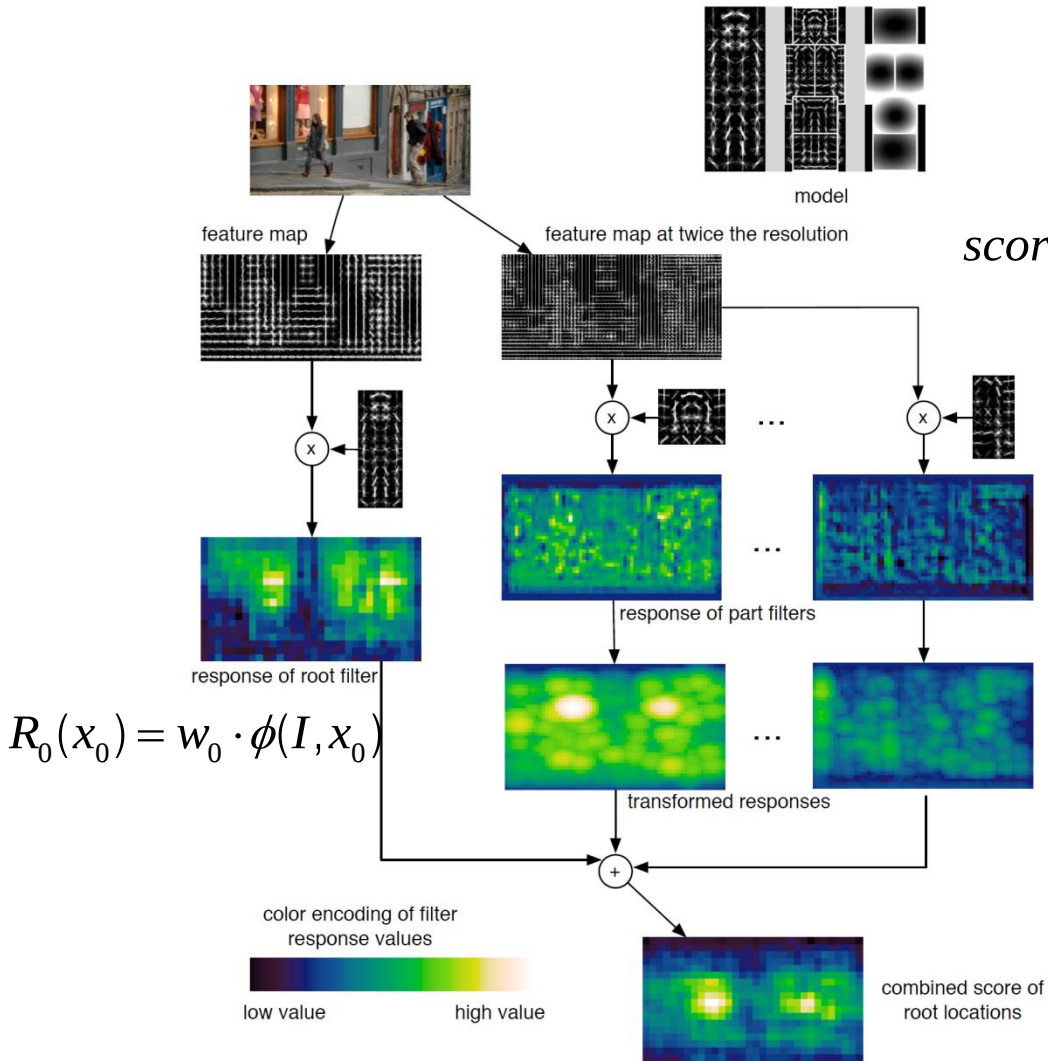


Transformed response

$$R_h(x_h) = w_h \cdot \phi(I, x_h)$$



$$D_h(x_0) = \max_{dx, dy} (R_h(x_0 + dx) - d_h \cdot dx^2)$$



CVPR 2011

2018 Longuet-Higgins Prize for fundamental contributions in computer vision

Articulated pose estimation with flexible mixtures-of-parts

Yi Yang Deva Ramanan

Dept. of Computer Science, University of California, Irvine

`{yyang8, dramanan}@ics.uci.edu`

Slide inspired by Yang

LIMITATION OF DPM

In-plane rotation



Foreshortening



Scaling



Out-of-plane rotation

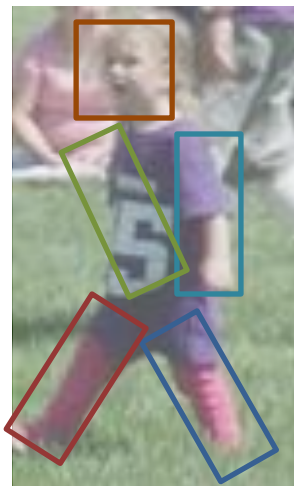


Intra-category variation

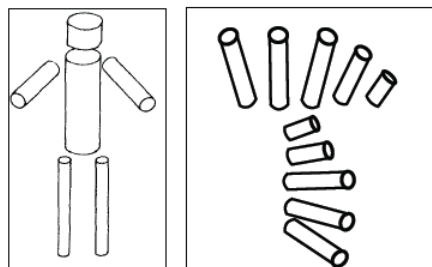


Aspect ratio

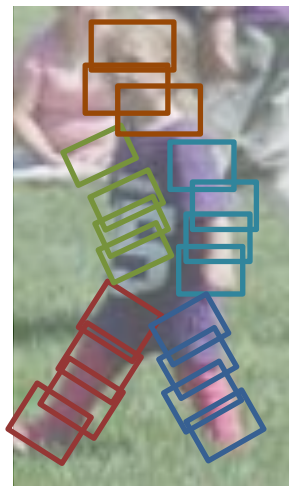




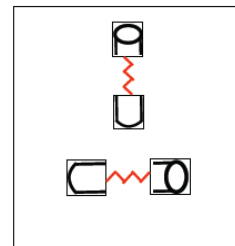
Part based model



Marr, Nishihara (1978)



3D deformation ~
Mixture of mini-parts



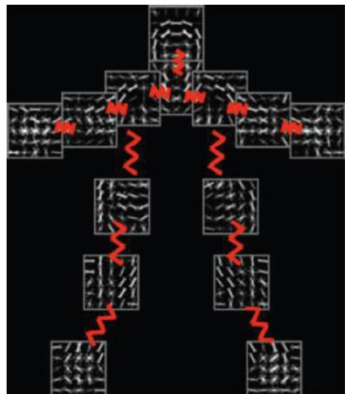


(a) Original

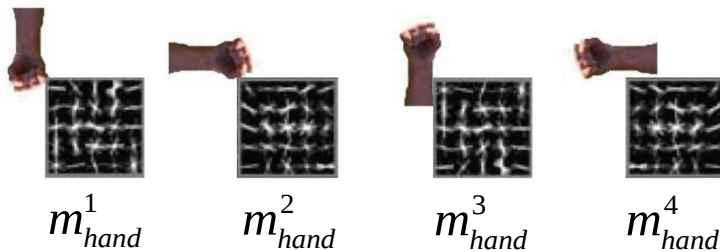
(b) Foreshortening

(c) Out-of-plane

MIXTURE PART MODEL



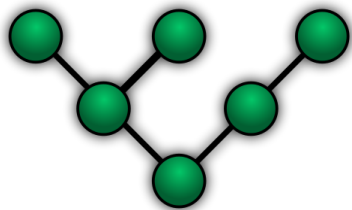
$$L(I, x) = \sum_{i \in V} \alpha_i \cdot \phi(I, x_i) + \sum_{i, j \in E} \beta_{ij} \cdot \phi(x_i, x_j)$$



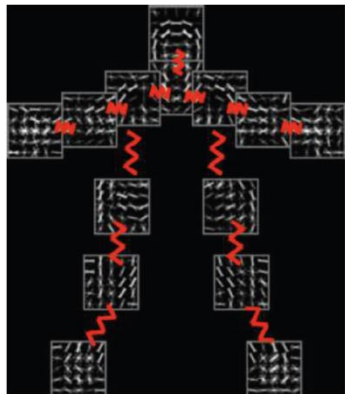
Mixture part model

$$L(I, x, m) = \sum_{i \in V} \alpha_i^{m_i} \cdot \phi(I, x_i) + \sum_{i, j \in E} \beta_{ij}^{m_i m_j} \cdot \phi(x_i, x_j)$$

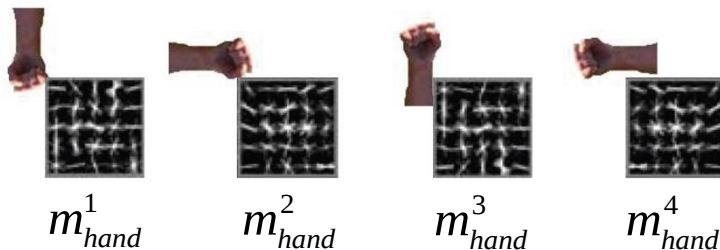
Appearance



MIXTURE PART MODEL

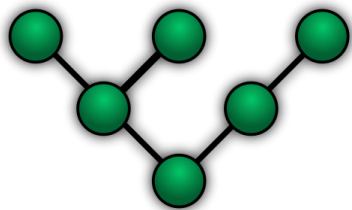


$$L(I, x) = \sum_{i \in V} \alpha_i \cdot \phi(I, x_i) + \sum_{i, j \in E} \beta_{ij} \cdot \phi(x_i, x_j)$$

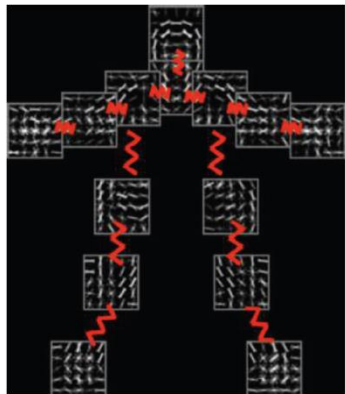


Mixture part model

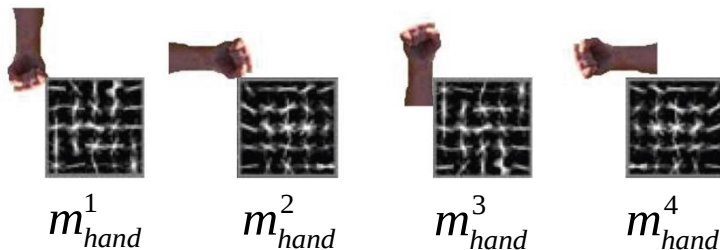
$$L(I, x, m) = \underbrace{\sum_{i \in V} \alpha_i^{m_i} \cdot \phi(I, x_i)}_{\text{Appearance}} + \underbrace{\sum_{i, j \in E} \beta_{ij}^{m_i m_j} \cdot \phi(x_i, x_j)}_{\text{Spatial rel.}}$$



MIXTURE PART MODEL

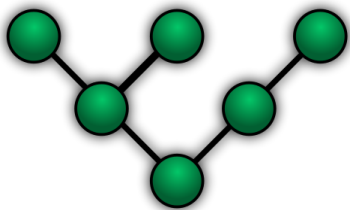


$$L(I, x) = \sum_{i \in V} \alpha_i \cdot \phi(I, x_i) + \sum_{i, j \in E} \beta_{ij} \cdot \phi(x_i, x_j)$$

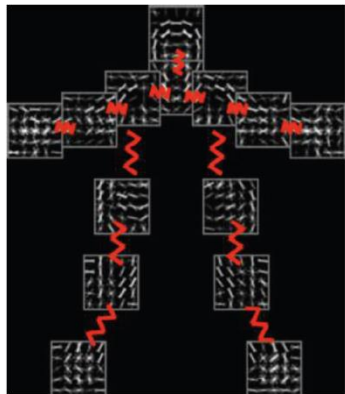


Mixture part model

$$L(I, x, m) = \underbrace{\sum_{i \in V} \alpha_i^{m_i} \cdot \phi(I, x_i)}_{\text{Appearance}} + \underbrace{\sum_{i, j \in E} \beta_{ij}^{m_i m_j} \cdot \phi(x_i, x_j)}_{\text{Spatial rel.}} + \underbrace{S(m)}_{\text{Mixture prior}}$$



MIXTURE PART MODEL

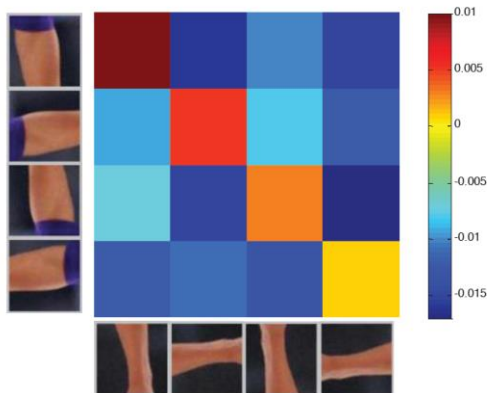
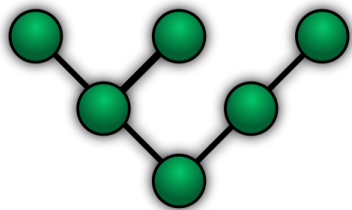


Mixture prior

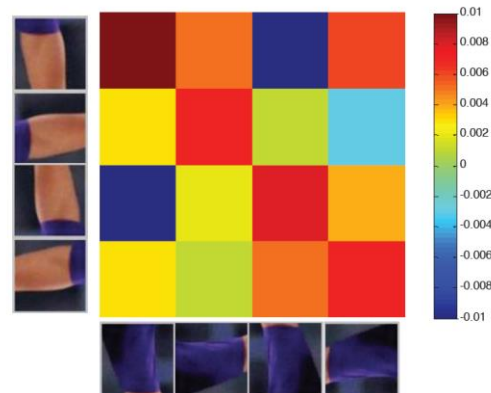
$$S(m) = \sum_{i \in V} b_i^{m_i} + \sum_{i, j \in E} b_{ij}^{m_i m_j}$$

$b_i^{m_i}$ Prior on mixture m_i

$b_{ij}^{m_i m_j}$ Co-occurrence prior between m_i and m_j

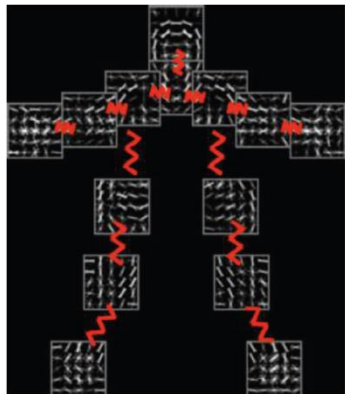


Same part



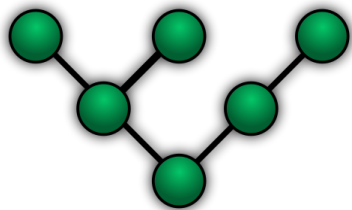
Different part

MIXTURE PART MODEL



$$score_i(x_i, m_i) = \alpha_i^{m_i} \cdot \phi(I, x_i) + b_i^{m_i} + \sum_{k \in kids(i)} u_k(x_i, m_i)$$

$$u_k(x_i, m_i) = \max_{x_j, m_j} \left(score_k(x_j, m_j) + \beta_{ij}^{m_i m_j} \cdot \phi(x_i, x_j) + b_{ij}^{m_i m_j} \right)$$



$$\text{Cf) } score(x_0) = R_0(x_0) + \sum_{i=1} \max_{x_i} \left(R_i(x_0 + dx) - d_i \cdot dx^2 \right)$$



MIXTURE PART MODEL

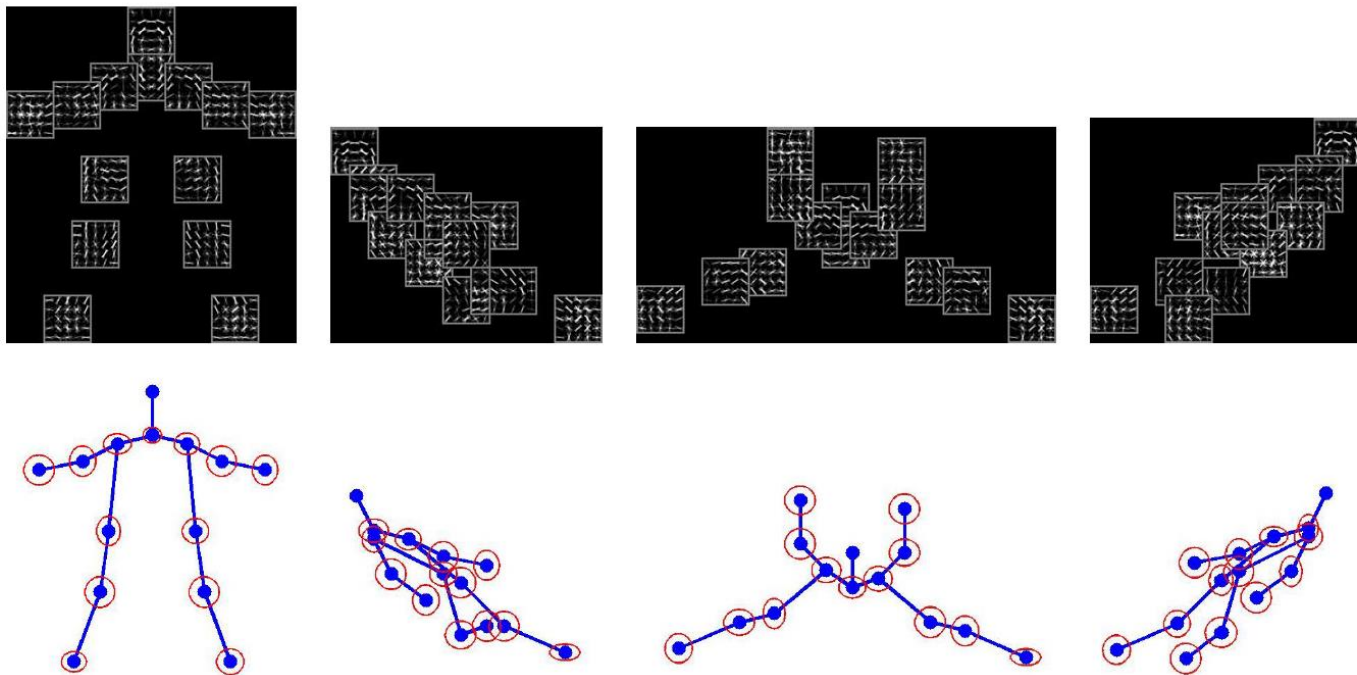
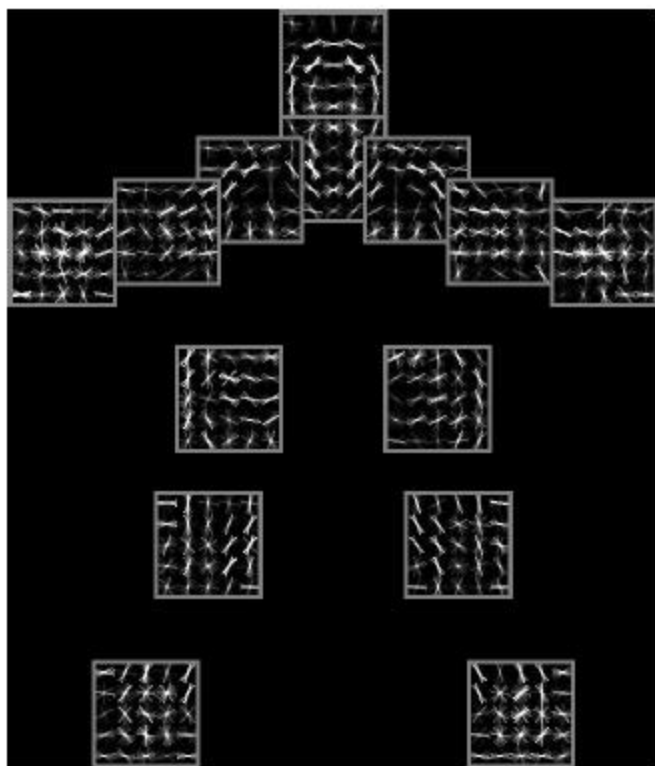
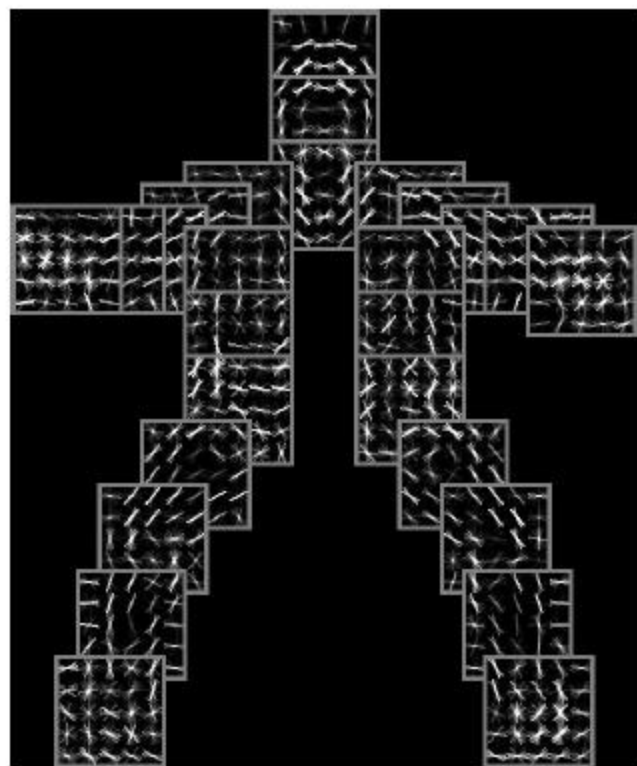


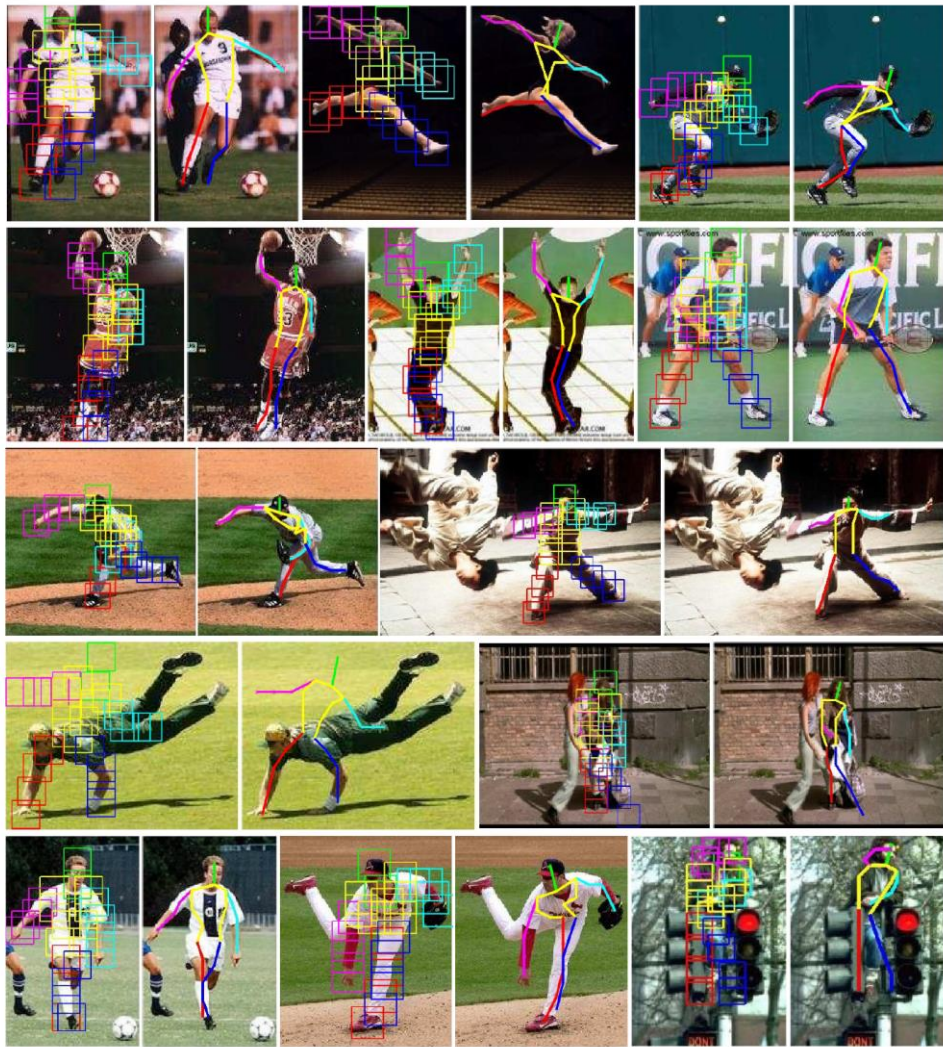
Fig. 6: A visualization of our model for $K = 14$ parts and $T = 4$ local mixtures, trained on the Parse dataset. We show the local templates **above**, and the tree structure **below**, placing parts at their best-scoring location relative to their parent. Though we visualize 4 trees, there exists $T^K \approx 2e7$ global combinations, obtained by composing different part types together with different springs. The score associated with each combination decomposes into a tree, and so is efficient to search over using dynamic programming (1).



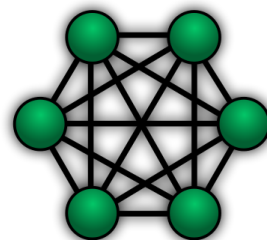
$K = 14$ parts



$K = 26$ parts



ECCV 2014



Pose Machines: Articulated Pose Estimation via Inference Machines

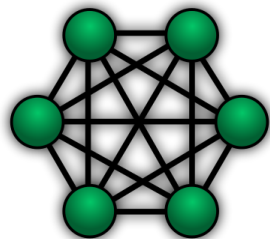
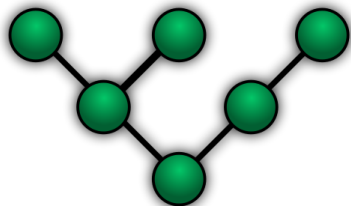
Varun Ramakrishna, Daniel Munoz, Martial Hebert,
J. Andrew Bagnell, and Yaser Sheikh

The Robotics Institute, Carnegie Mellon University

Slide inspired by Ramakrishna

POSE INFERENCE MACHINE

Elbow estimation

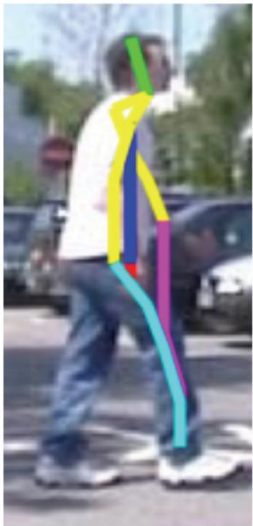


$$\sum_{ij} \beta_{ij}^{m_i m_j} \cdot \varphi(x_i, x_j)$$

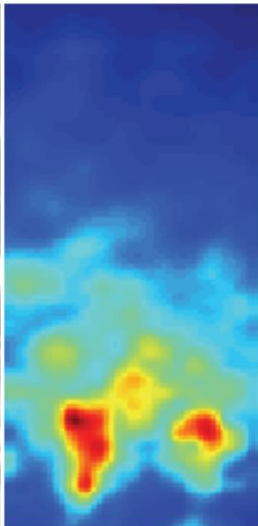
All possible pairs of joints



Input Image



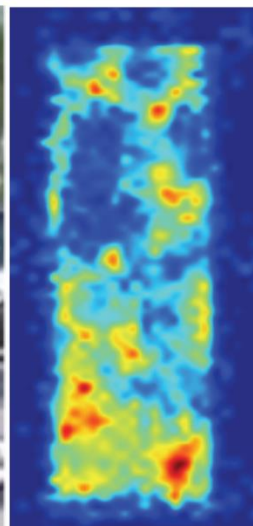
Estimated Pose



Max Marginal
(left ankle)



Estimated Pose



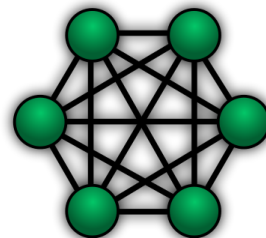
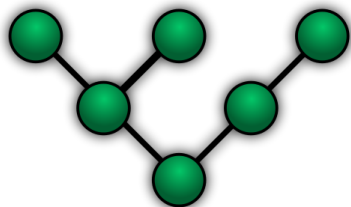
Stage I
Confidence



Stage II
Confidence



Stage III
Confidence



POSE INFERENCE MACHINE

Image Location z

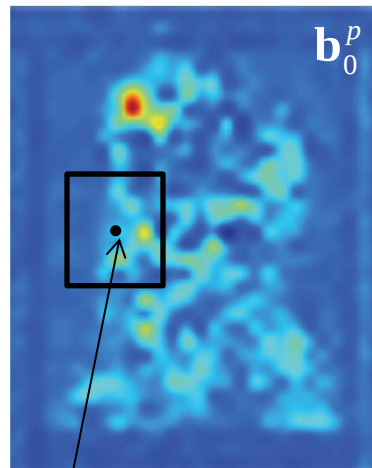


Input Image



$\mathbf{x}_z$

$g_0^p(\mathbf{x}_z)$
—
Regressor

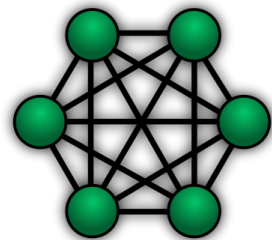
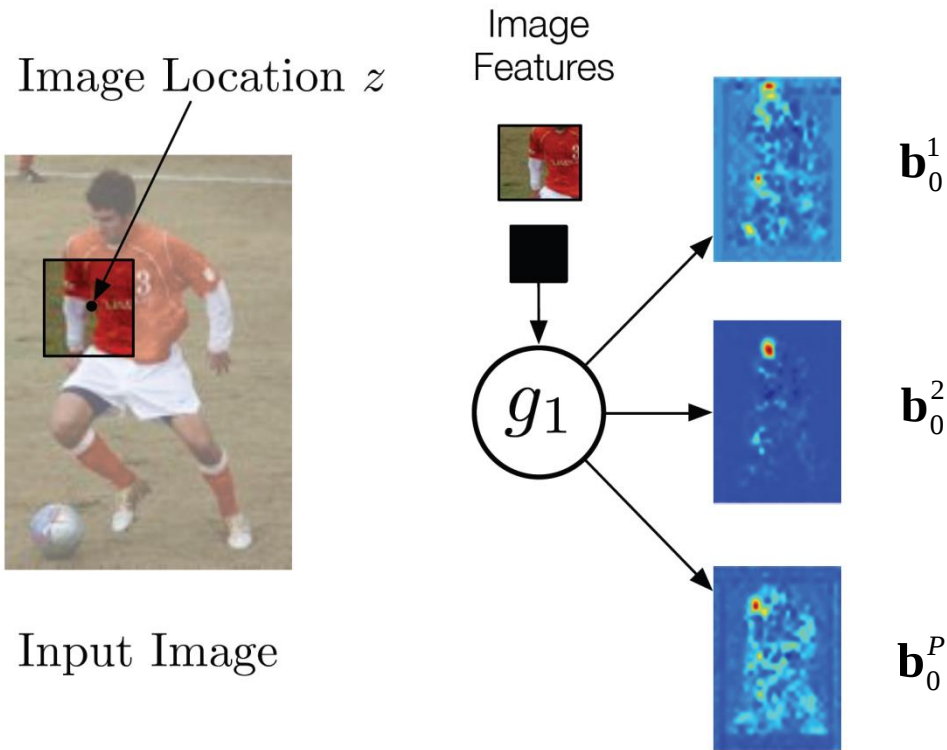


$\mathbf{b}_0^p$

$$b_0(Y_p = z) = g_0^p(\mathbf{x}_z)$$

Y_p : location of p joint

POSE INFERENCE MACHINE



$$b_0(Y_p = z) = g_0^p(\mathbf{x}_z)$$

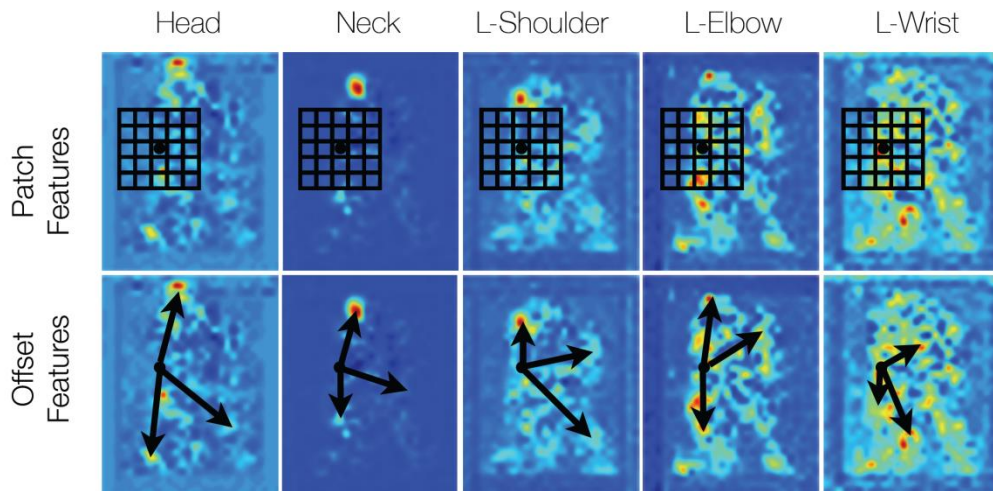
$$b_1(Y_p = z) = g_1^p(\mathbf{x}_z; \underbrace{\oplus \varphi(z; \mathbf{b}_0^p)}_{\text{Context from previous prediction}})$$

Context from previous prediction

Cf)

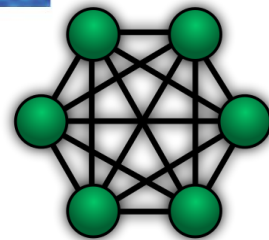
$$L(I, x) = \sum_{i \in V} \alpha_i \cdot \phi(I, x_i) + \sum_{i, j \in E} \beta_{ij} \cdot \phi(x_i, x_j)$$

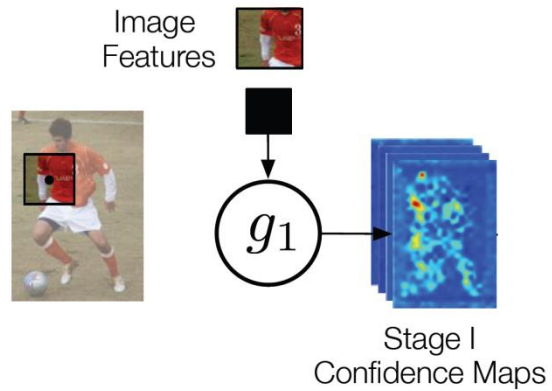
CONTEXT FEATURES



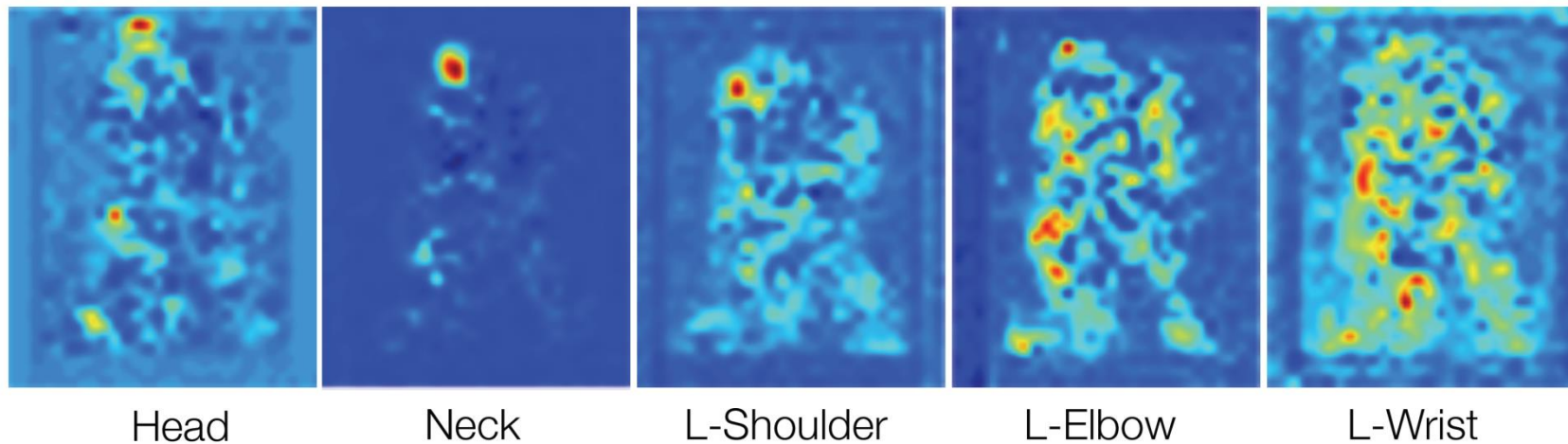
$$b_1(Y_p = z) = g_1^p(\mathbf{x}_z; \underbrace{\oplus \varphi(z; \mathbf{b}_0^p)}_{\text{Context from previous prediction}})$$

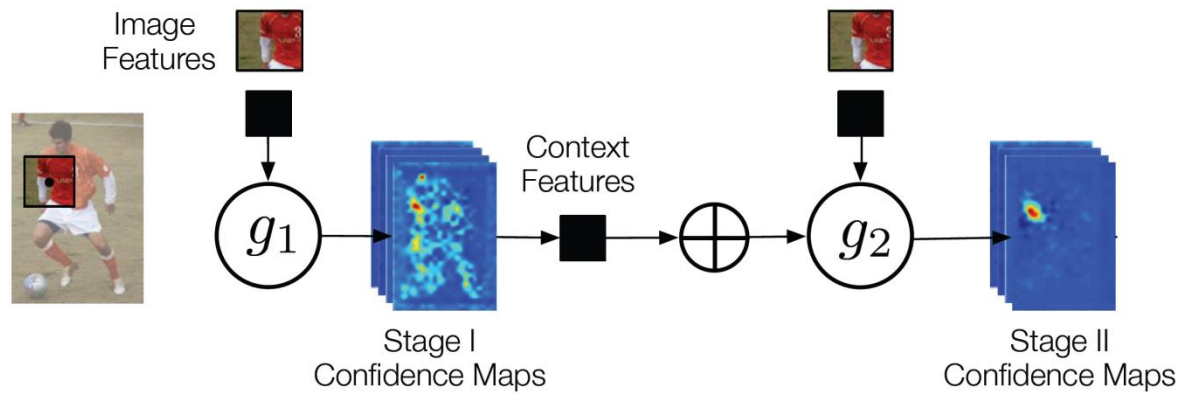
Context from previous prediction



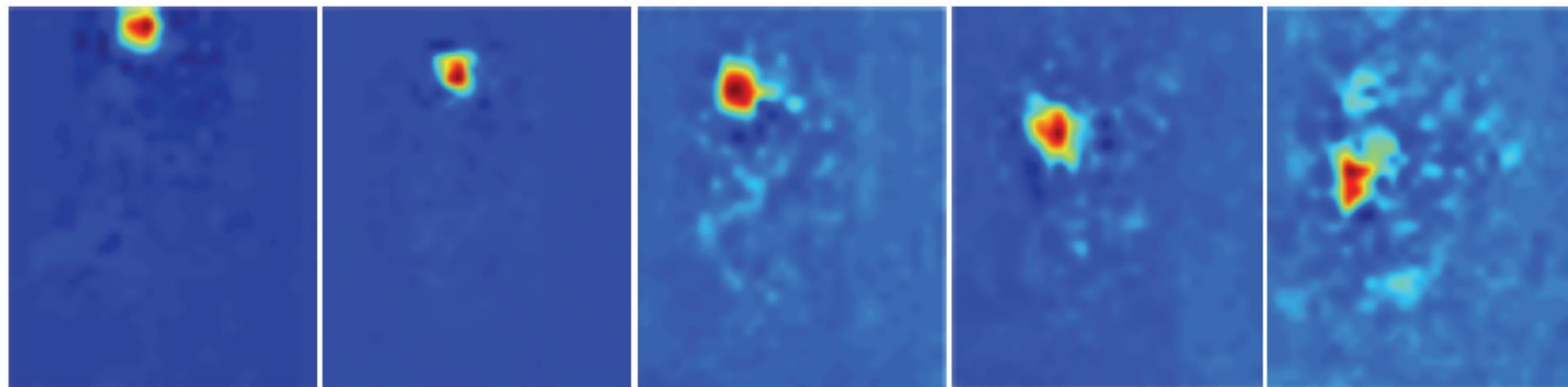


Stage I Confidence





Stage II Confidence



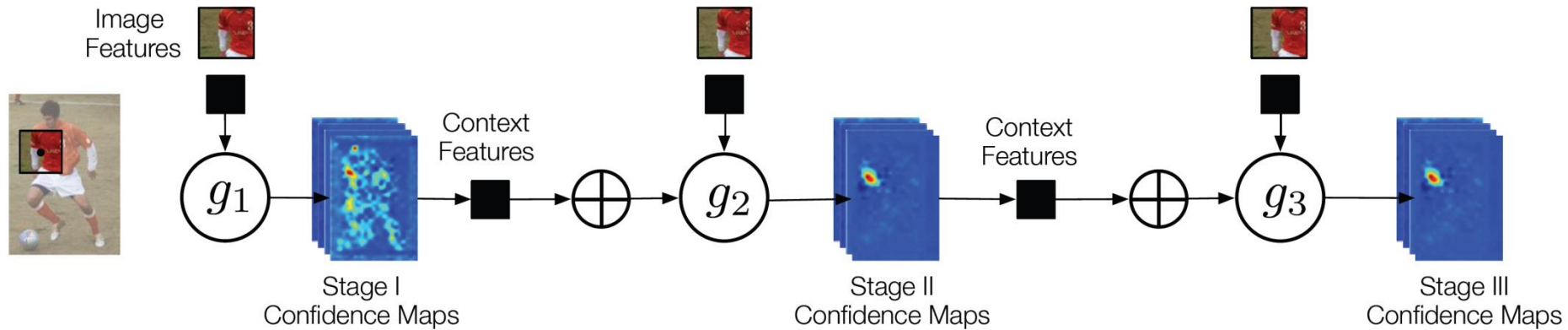
Head

Neck

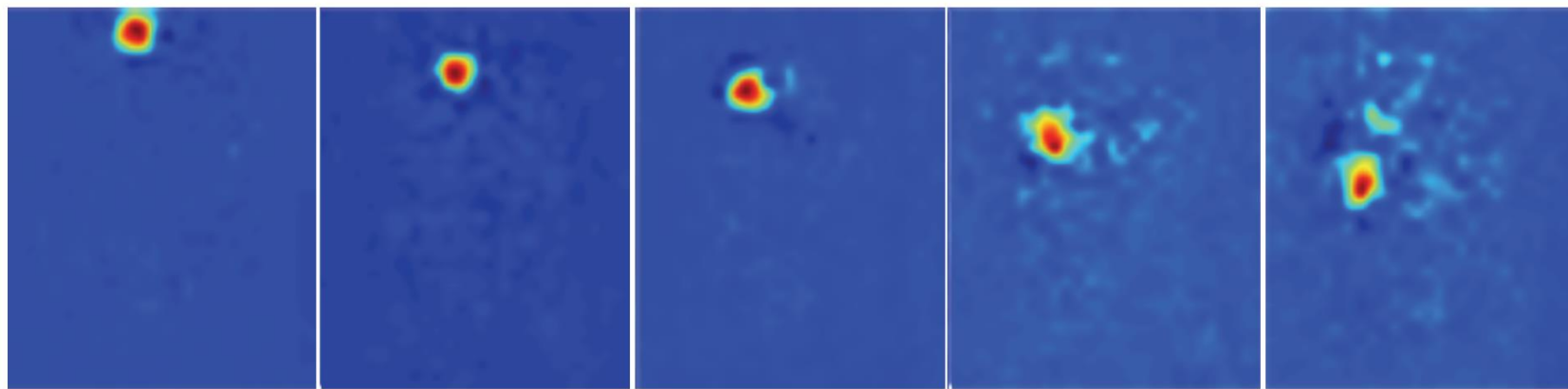
L-Shoulder

L-Elbow

L-Wrist



Stage III Confidence



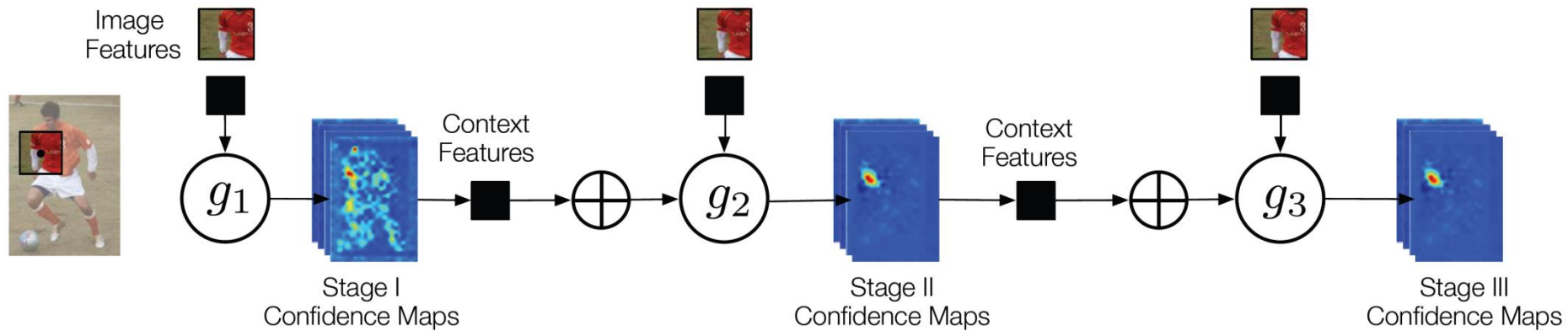
Head

Neck

L-Shoulder

L-Elbow

L-Wrist



Stage III Confidence



Head

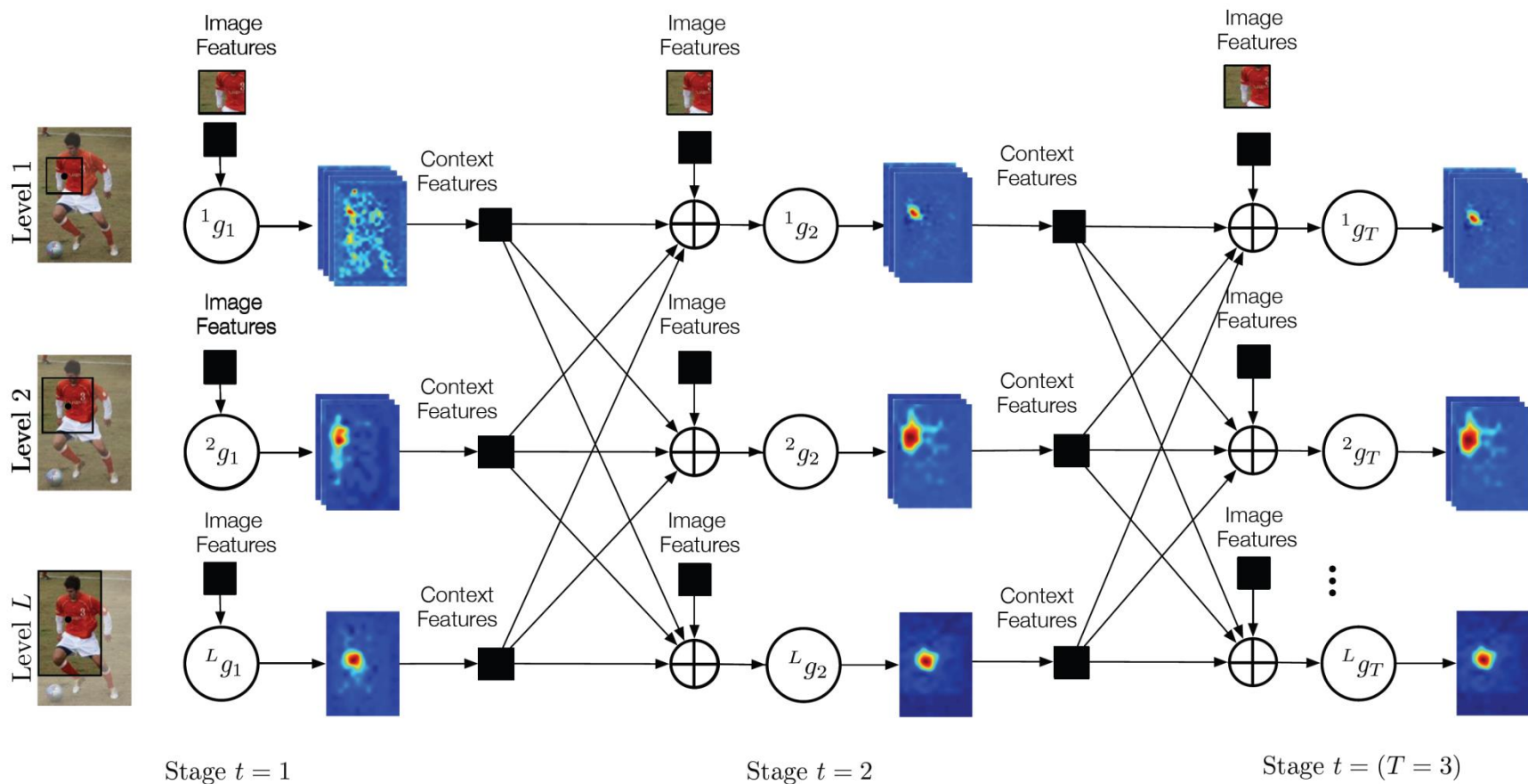
Neck

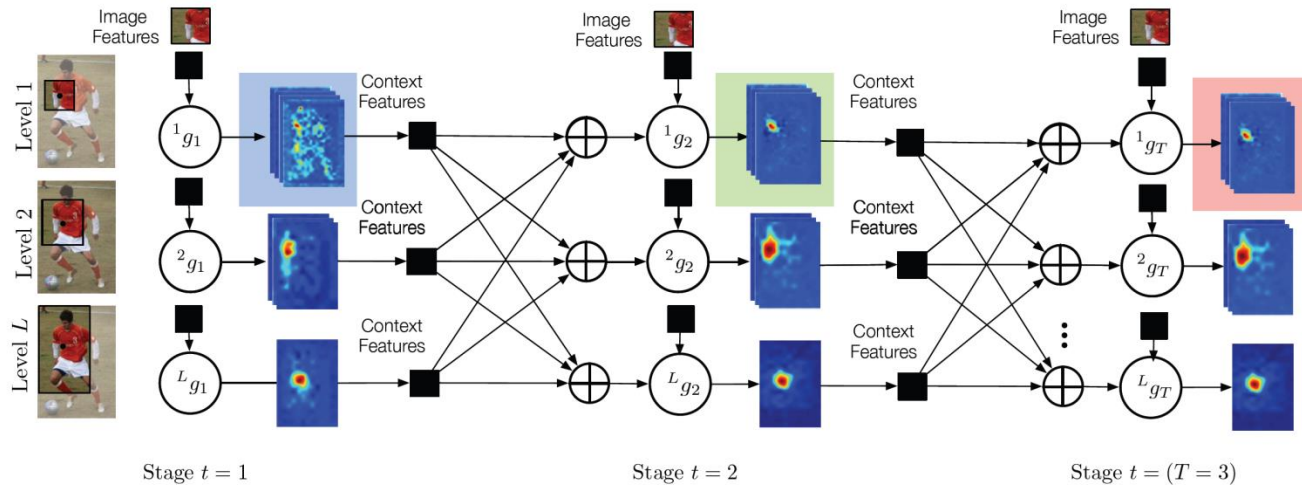
L-Shoulder

L-Elbow

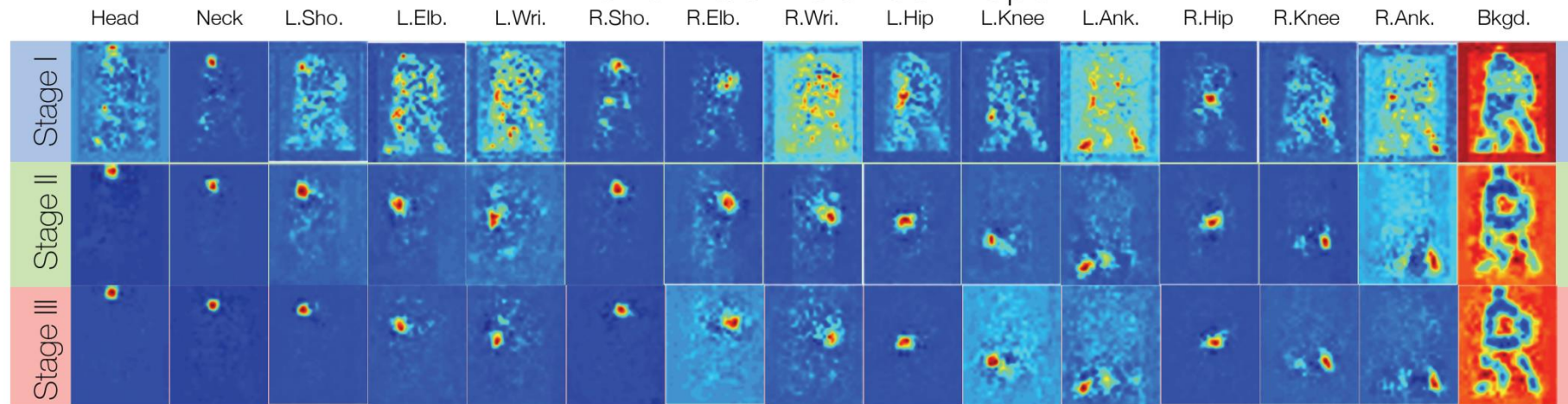
L-Wrist

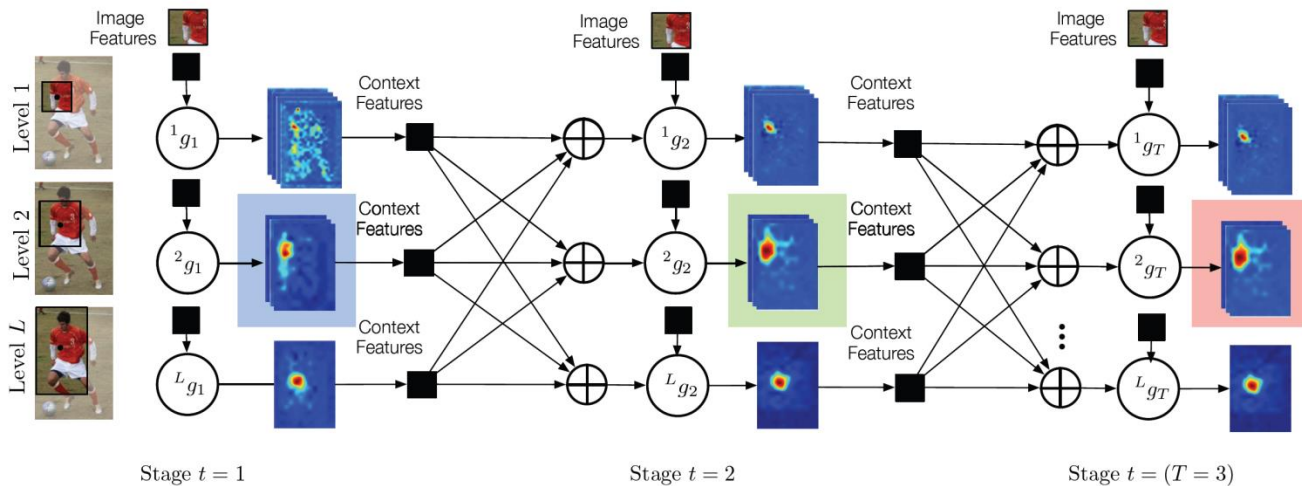
HIERACHICAL REPRESENTATION



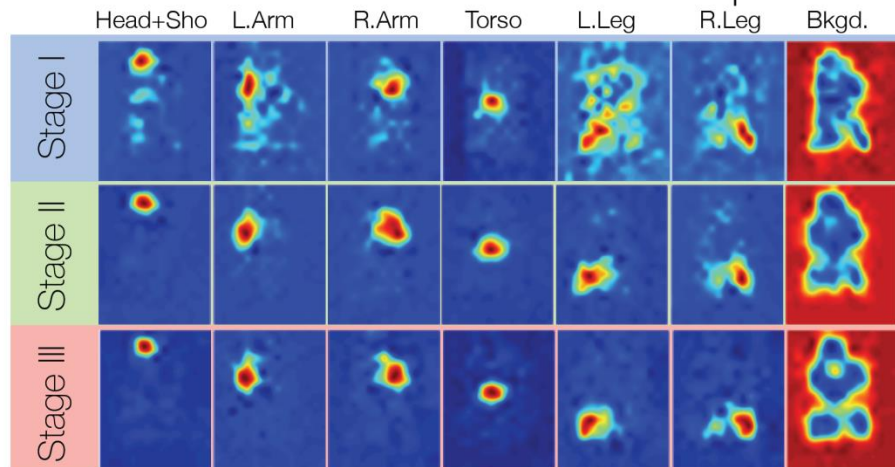


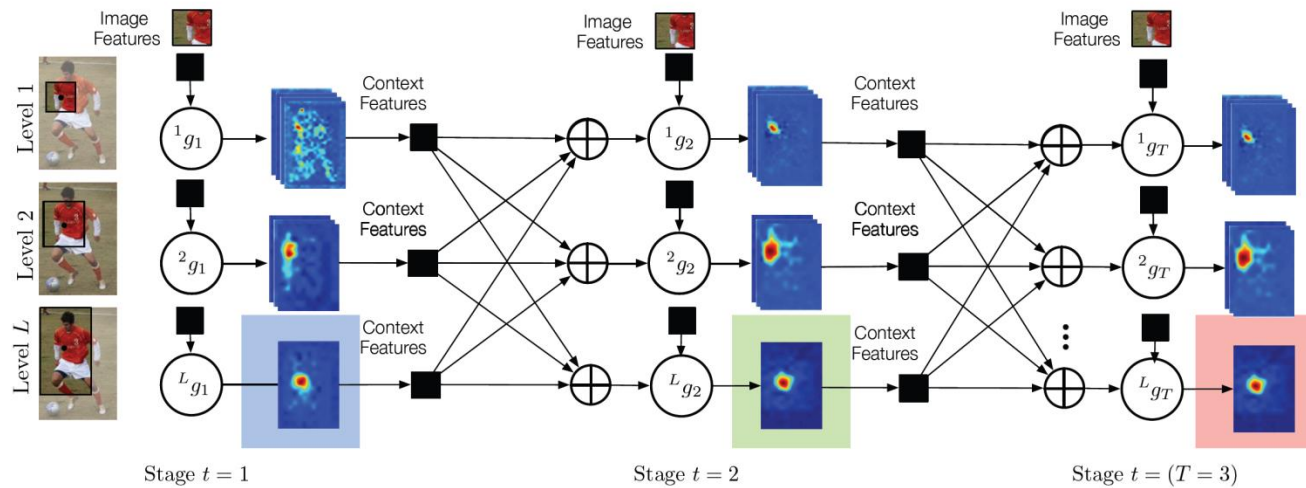
Level I Confidence Maps



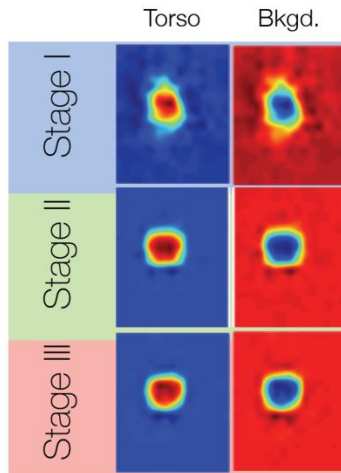


Level 2 Confidence Maps





Level 3 Confidence Maps





DATASETS



FLIC (Frame Labeled In Cinema) dataset (CVPR 2013)

- Human detection on Hollywood movies
- Mechanical Turker for annotation
- Upper body
- 20928 labeled images

DATASETS



MPII dataset (CVPR 2014)

- YouTube videos
- 25k images
- 40k humans
- Full body

DATASETS



MSCOCO dataset (ECCV 2014)

- Internet images
- 200k images
- 250k humans
- Full body

CVPR 2014

DeepPose: Human Pose Estimation via Deep Neural Networks

Alexander Toshev Christian Szegedy

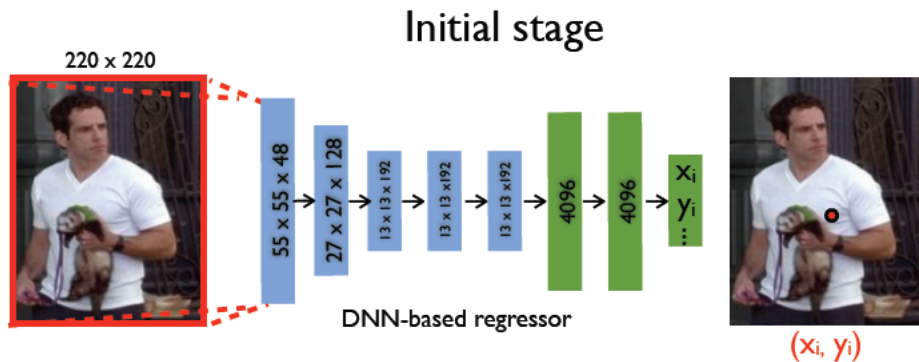
Google

1600 Amphitheatre Pkwy

Mountain View, CA 94043

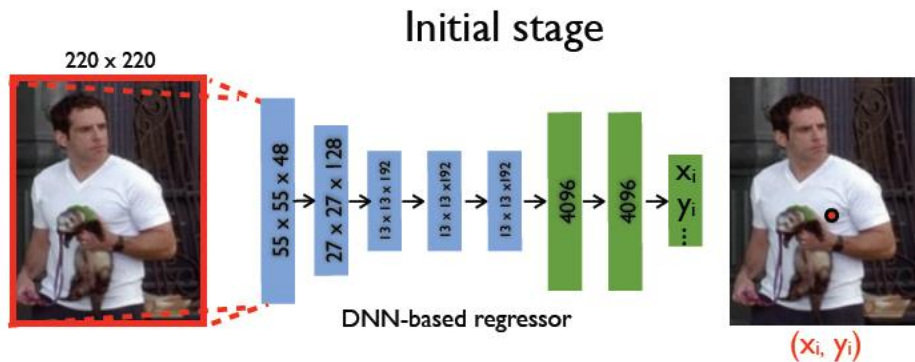
`toshev,szegedy@google.com`

COORDINATE REGRESSION



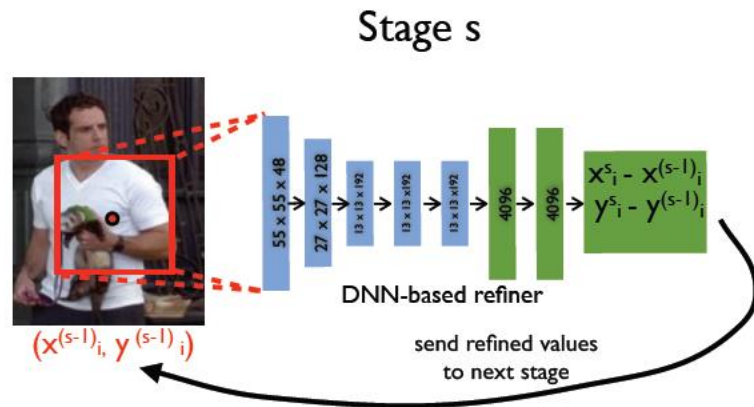
$$\begin{bmatrix} x_1 \\ y_1 \\ \vdots \\ x_n \\ y_n \end{bmatrix} = z = f(I; \theta) \quad \argmin_{\theta} \sum_i \|z_i - f(I_i; \theta)\|^2$$

COORDINATE REGRESSION



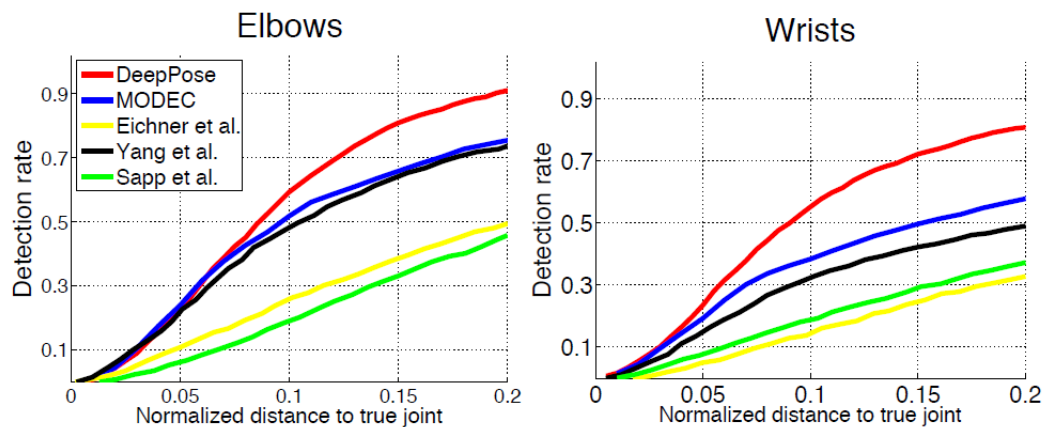
$$\begin{bmatrix} x_1 \\ y_1 \\ \vdots \\ x_n \\ y_n \end{bmatrix} = z = f(I; \theta)$$

$$\operatorname{argmin}_{\theta} \sum_i \|z_i - f(I_i; \theta)\|^2$$



$$\operatorname{argmin}_{\theta} \sum_i \|z_i - f(b(I_i); \theta)\|^2$$

Pose refinement in the bounding box



Method	Arm		Leg		Ave.
	Upper	Lower	Upper	Lower	
DeepPose-st1	0.5	0.27	0.74	0.65	0.54
DeepPose-st2	0.56	0.36	0.78	0.70	0.60
DeepPose-st3	0.56	0.38	0.77	0.71	0.61
Dantone et al. [2]	0.45	0.25	0.65	0.61	0.49
Tian et al. [25]*	0.52	0.33	0.70	0.60	0.56
Johnson et al. [13]	0.54	0.38	0.75	0.66	0.58
Wang et al. [26]*	0.565	0.37	0.76	0.68	0.59
Pishchulin [18] ^o	0.49	0.32	0.74	0.70	0.56



Figure 6. Predicted poses in red and ground truth poses in green for the first three stages of a cascade for three examples.

CVPR 2016

Convolutional Pose Machines

Shih-En Wei

shihenw@cmu.edu

Varun Ramakrishna

vramakri@cs.cmu.edu

Takeo Kanade

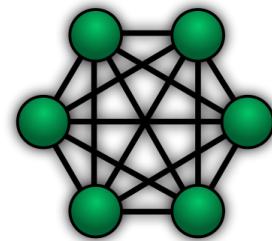
Takeo.Kanade@cs.cmu.edu

Yaser Sheikh

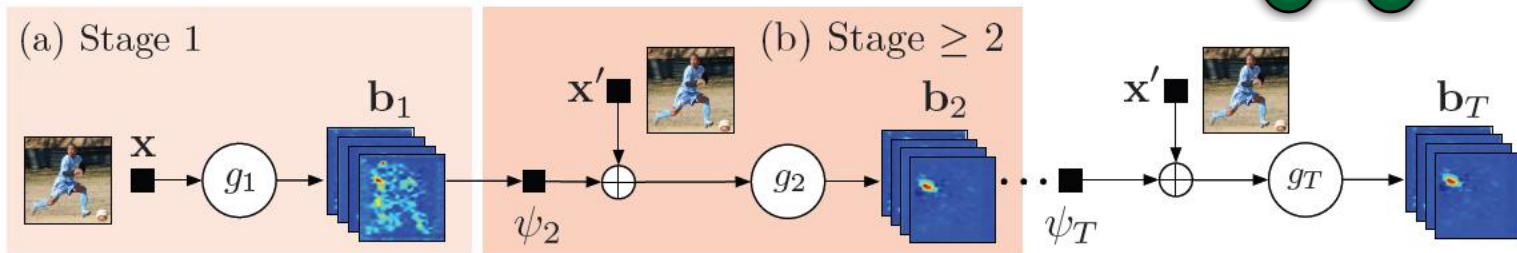
yaser@cs.cmu.edu

The Robotics Institute
Carnegie Mellon University

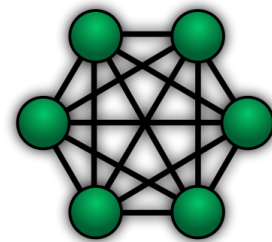
CONVOLUTIONAL POSE MACHINE (CPM)



Pose machine



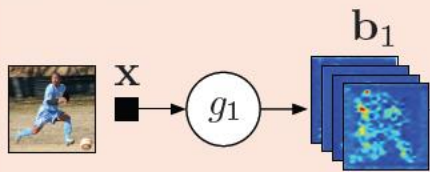
CONVOLUTIONAL POSE MACHINE (CPM)



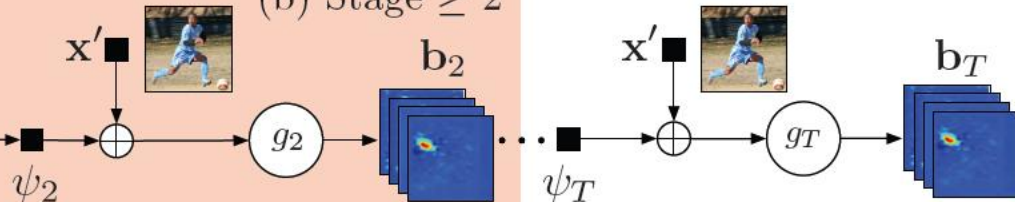
Convolutional
Pose Machines
(T -stage)



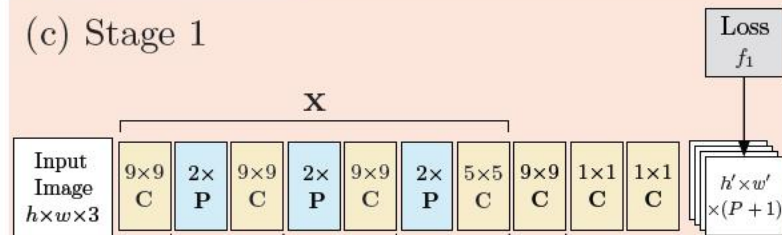
(a) Stage 1



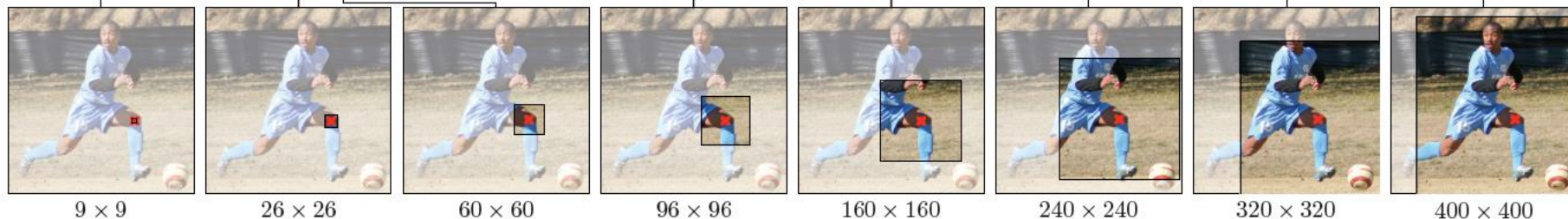
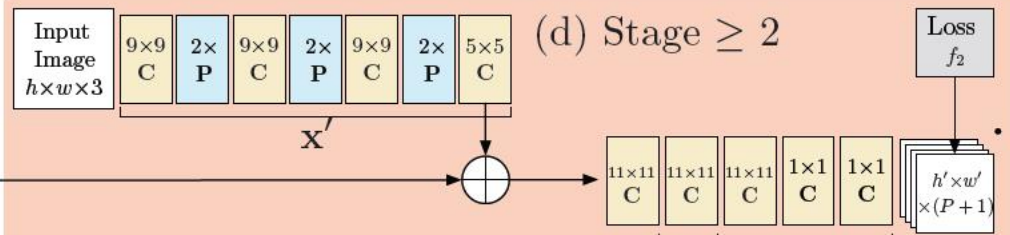
(b) Stage ≥ 2



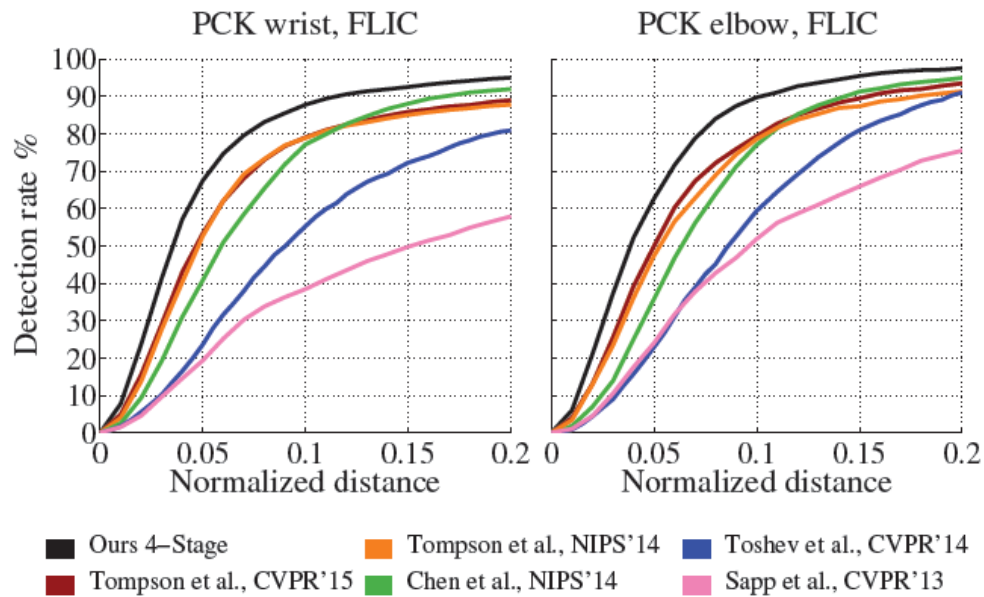
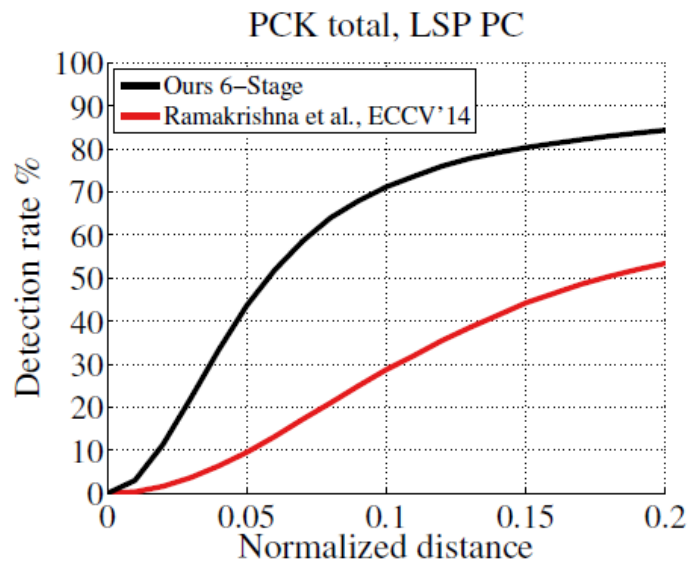
(c) Stage 1

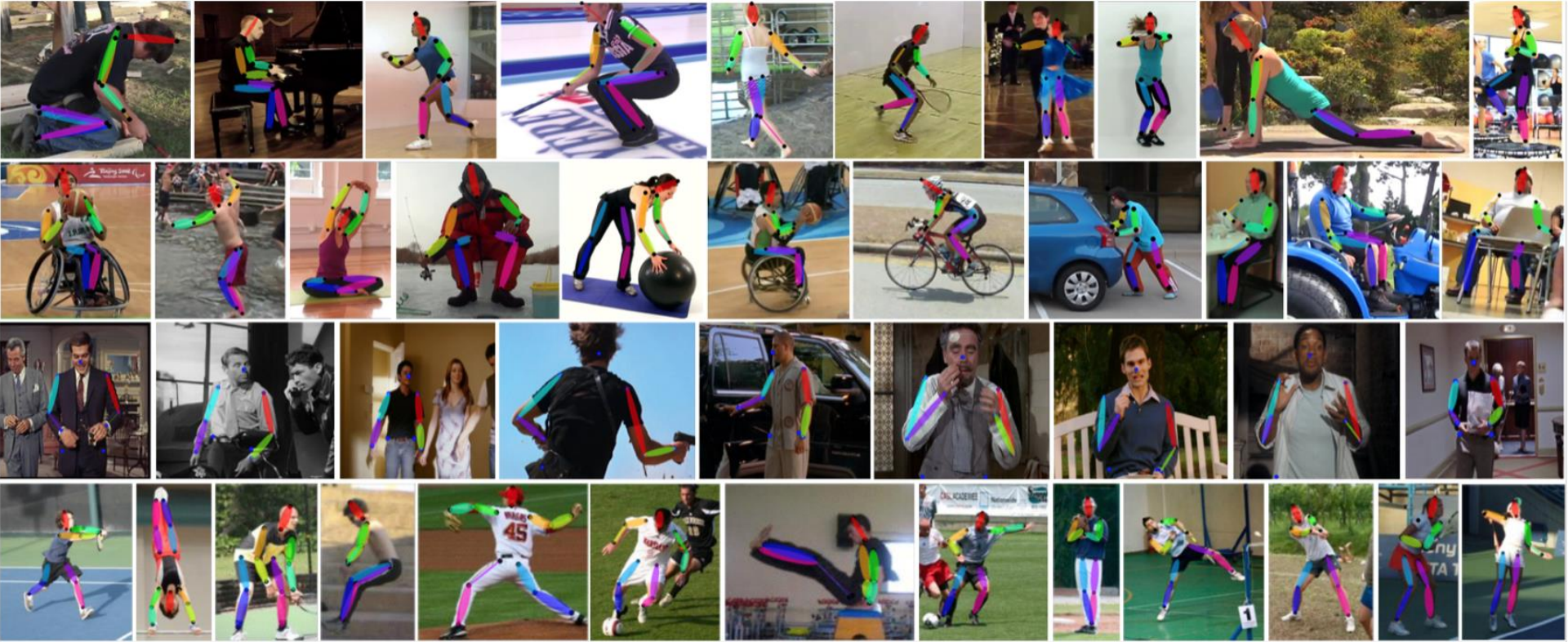


(d) Stage ≥ 2



(e) Effective Receptive Field





CVPR 2017

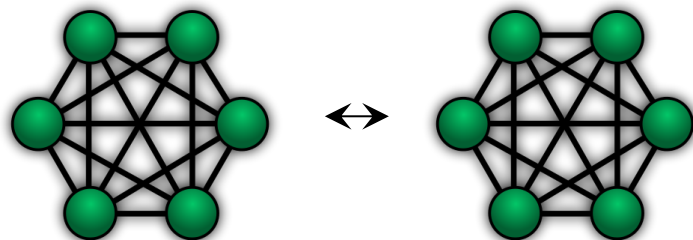
Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields *

Zhe Cao Tomas Simon Shih-En Wei Yaser Sheikh

The Robotics Institute, Carnegie Mellon University

{zhecao, shihenw}@cmu.edu

{tsimon, yaser}@cs.cmu.edu



PART AFFINITY FIELDS



(a) Input Image



(b) Part Confidence Maps



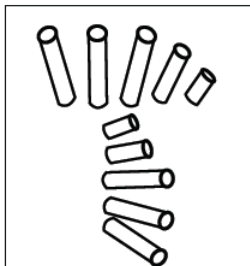
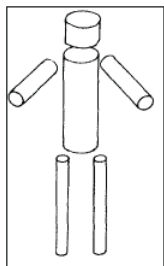
(c) Part Affinity Fields



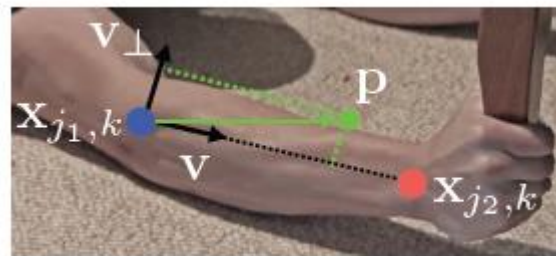
(d) Bipartite Matching



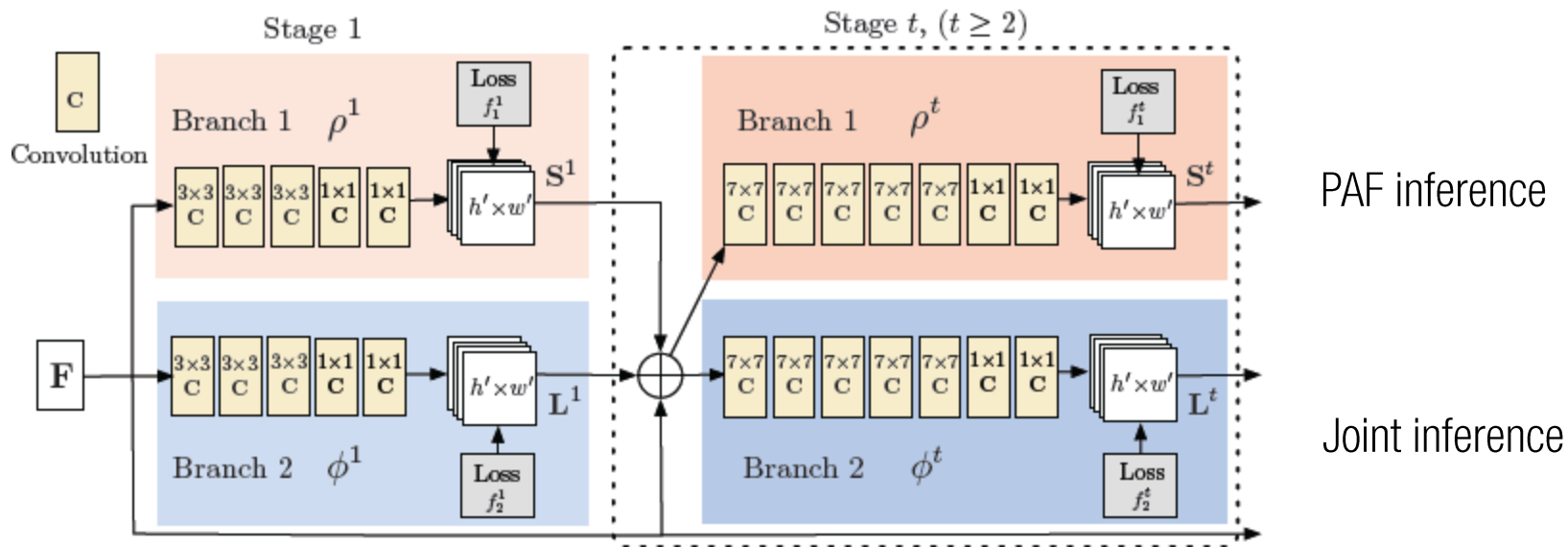
(e) Parsing Results



Marr, Nishihara (1978)



PART AFFINITY FIELDS



PART AFFINITY FIELDS BASED CONNECTION



Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	mAP	s/image
Subset of 288 images as in [22]									
Deepcut [22]	73.4	71.8	57.9	39.9	56.7	44.0	32.0	54.1	57995
Iqbal et al. [12]	70.0	65.2	56.4	46.1	52.7	47.9	44.5	54.7	10
DeeperCut [11]	87.9	84.0	71.9	63.9	68.8	63.8	58.1	71.2	230
Ours	93.7	91.4	81.4	72.5	77.7	73.0	68.1	79.7	0.005
Full testing set									
DeeperCut [11]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5	485
Iqbal et al. [12]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1	10
Ours (one scale)	89.0	84.9	74.9	64.2	71.0	65.6	58.1	72.5	0.005
Ours	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6	0.005

- <https://www.youtube.com/watch?v=pW6nZXeWIGM>