Associative Embedding: End-to-End Learning for Joint Detection and Grouping

Troyle Thomas

Outline

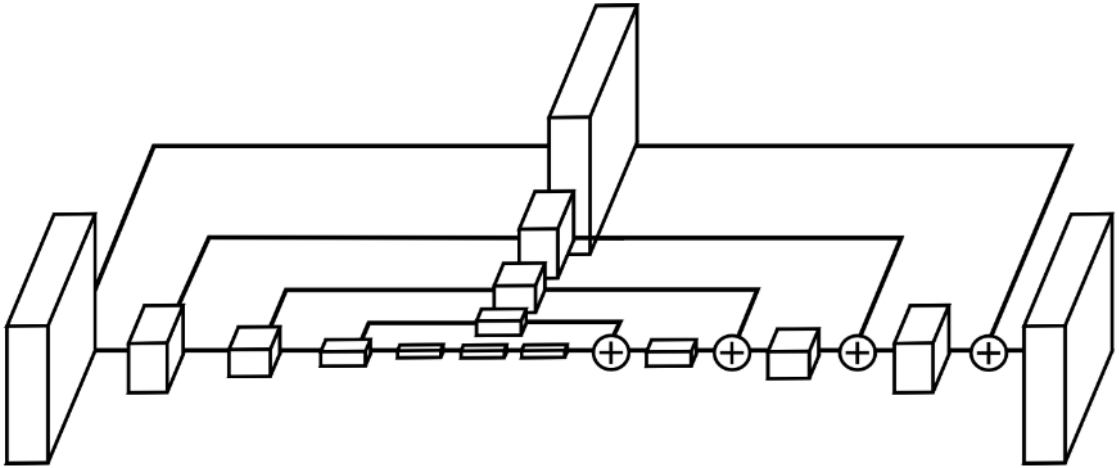
1. Problem & Motivation
2. Architecture
3. Multiperson Pose Estimation
4. Instance Segmentation

Problem



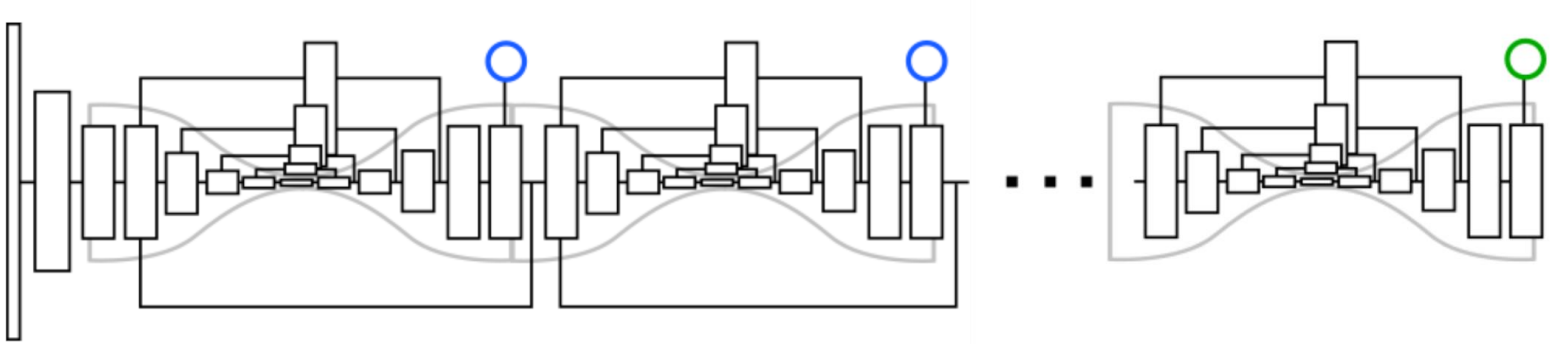
Hourglass Architecture



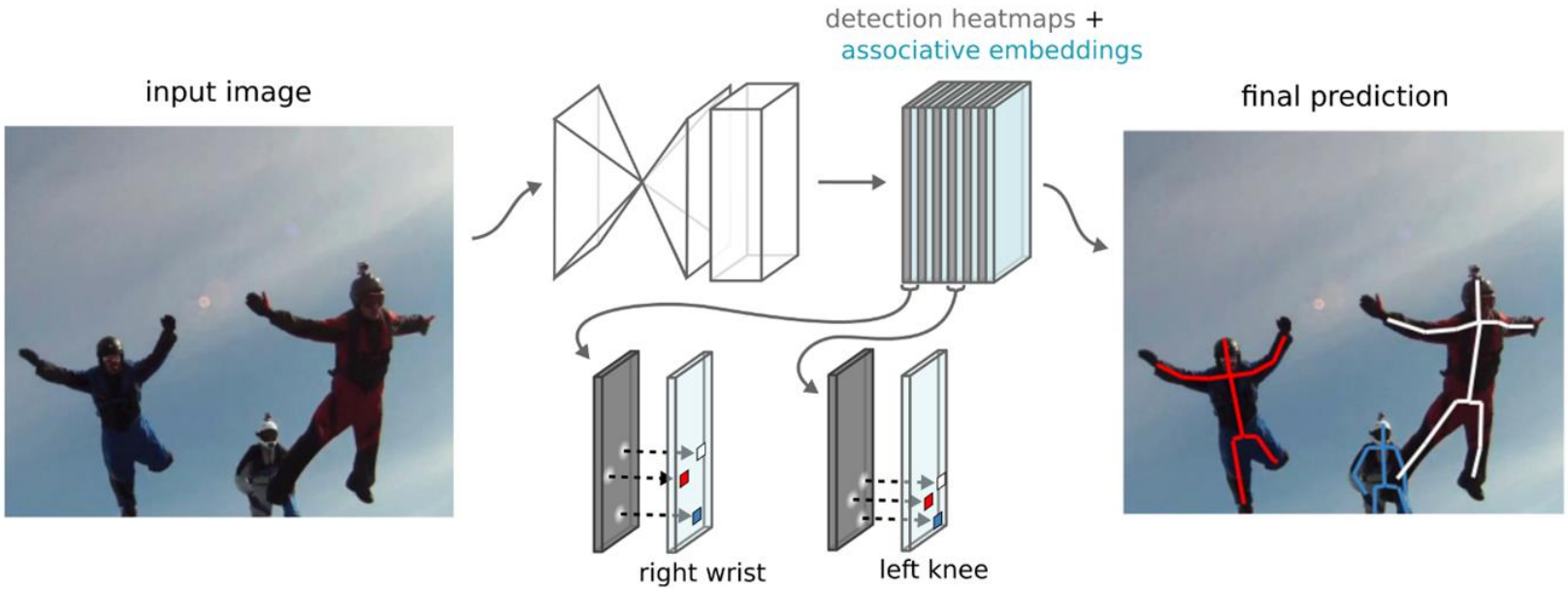


Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. ECCV, 2016.

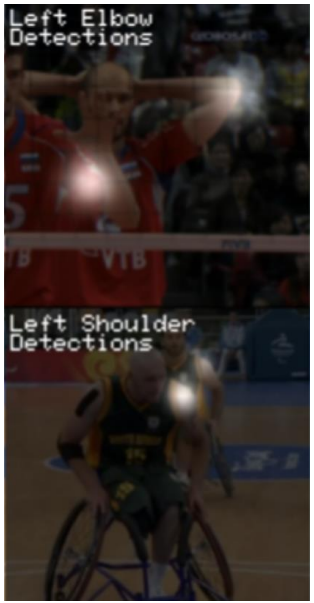
Stacked Hourglass Architecture

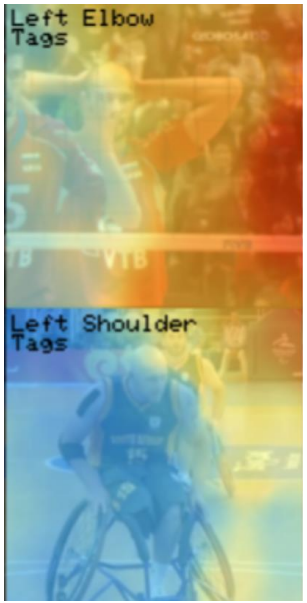


Multiperson Pose Estimation

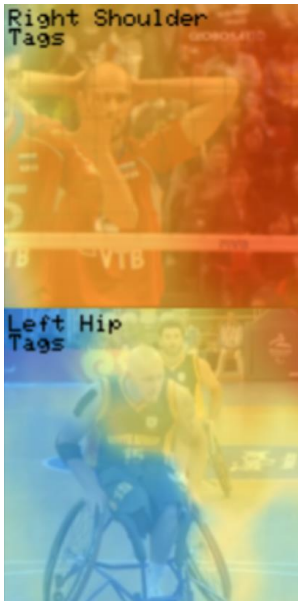





Left Elbow
Detections
Left Shoulder
Detections


Left Elbow
Tags
Left Shoulder
Tags


Right Shoulder
Detections
Left Hip
Detections


Right Shoulder
Tags
Left Hip
Tags

Loss Function

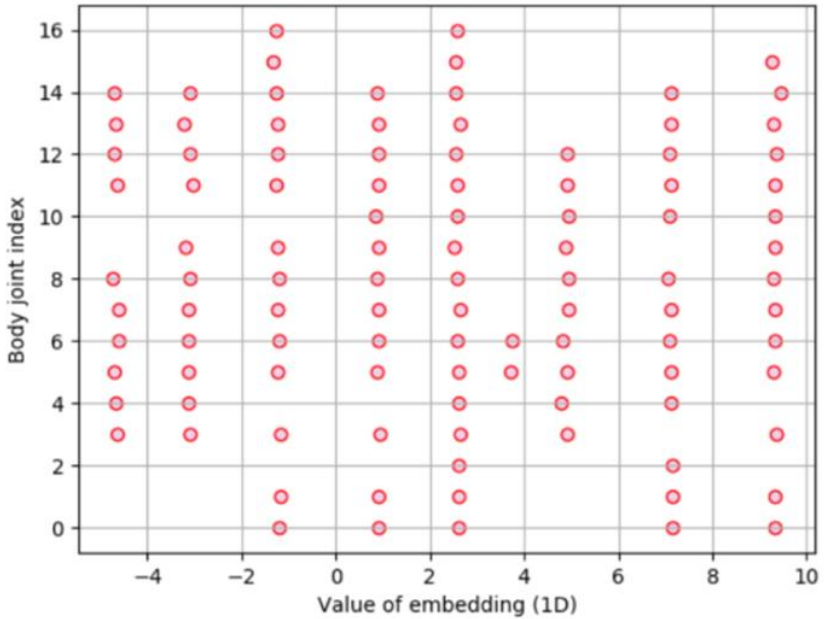
h_k - predicted tag heatmap for the kth joint
h(x) - tag value at pixel x
x_{nk}- ground truth pixel for nth person & kth joint

$$\bar{h}_n = \frac{1}{K} \sum_k h_k(x_{nk})$$

$$L_g(h, T) = \frac{1}{N} \sum_n \sum_k \left(\bar{h}_n - h_k(x_{nk},)\right)^2$$

$$+ \frac{1}{N^2} \sum_n \sum_{n'} \exp\{-\frac{1}{2\sigma^2} \left(\bar{h}_n - \bar{h}_{n'}\right)^2\},$$




Body joint index
Value of embedding (1D)
16
14
12
10
8
6
4
2
0
−4
−2
0
2
4
6
8
10

Multiscale Postprocessing

- Get test image at multiple scales
- Process through our network
- Average the results together

Implementation Details

- Four Stacked Hourglass Modules
- 512 x 512 Input
- 128 x 128 Output
- 32 Batch Size
- 2e-4 and 1e-5 after 100k iteration Learning Rate

Joints

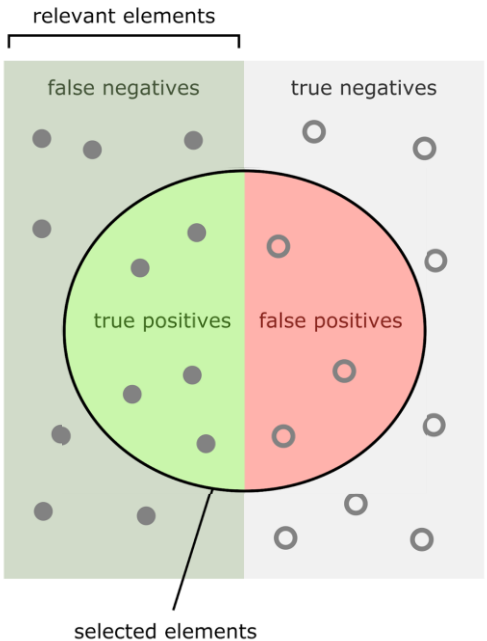
MS COCO

Nose, Eyes, Ears, Shoulders, Elbows, Wrists, Hips, Knees, Ankles

MPII

Head, Shoulders, Elbows, Wrists, Hip, Knees, Ankles

Precision and Recall



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

https://en.wikipedia.org/wiki/Precision_and_recall

MPII Multiperson Results

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Iqbal&Gall, ECCV16 [29]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1
Insafutdinov et al., ECCV16 [28]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5
Insafutdinov et al., arXiv16a [45]	89.4	84.5	70.4	59.3	68.9	62.7	54.6	70.0
Levinkov et al., CVPR17 [33]	89.8	85.2	71.8	59.6	71.1	63.0	53.5	70.6
Insafutdinov et al., CVPR17 [27]	88.8	87.0	75.9	64.9	74.2	68.8	60.5	74.3
Cao et al., CVPR17 [6]	91.2	87.6	77.7	66.8	75.4	68.9	61.7	75.6
Fang et al., arXiv17 [15]	88.4	86.5	78.6	**70.4**	74.4	**73.0**	**65.8**	76.7
Our method	**92.1**	**89.3**	**78.9**	69.8	**76.2**	71.6	64.7	**77.5**

Table 1. Results (AP) on MPII Multi-Person.

Object Keypoint Similarity

- MS-COCO
- Average Euclidean distance at different object scales from the ground truth
- Analogous to IOU

Object Keypoint Similarity



http://image-net.org/challenges/talks/2016/ECCV2016_workshop_presentation_keypoint.pdf

MS-COCO Results

	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
CMU-Pose [6]	0.611	0.844	0.667	0.558	0.684	0.665	0.872	0.718	0.602	0.749
G-RMI [42]	0.643	0.846	0.704	**0.614**	0.696	0.698	0.885	0.755	0.644	0.771
Our method	**0.663**	**0.865**	**0.727**	0.613	**0.732**	**0.715**	**0.897**	**0.772**	**0.662**	**0.787**

Table 2. Results on MS-COCO **test-std**, excluding systems trained with external data.

	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
CMU-Pose [6]	0.618	0.849	0.675	0.571	0.682	0.665	0.872	0.718	0.606	0.746
Mask-RCNN [25]	0.627	**0.870**	0.684	0.574	0.711	–	–	–	–	–
G-RMI [42]	0.649	0.855	0.713	**0.623**	0.700	0.697	0.887	0.755	0.644	0.771
Our method	**0.655**	0.868	**0.723**	0.606	**0.726**	**0.702**	**0.895**	**0.760**	**0.646**	**0.781**

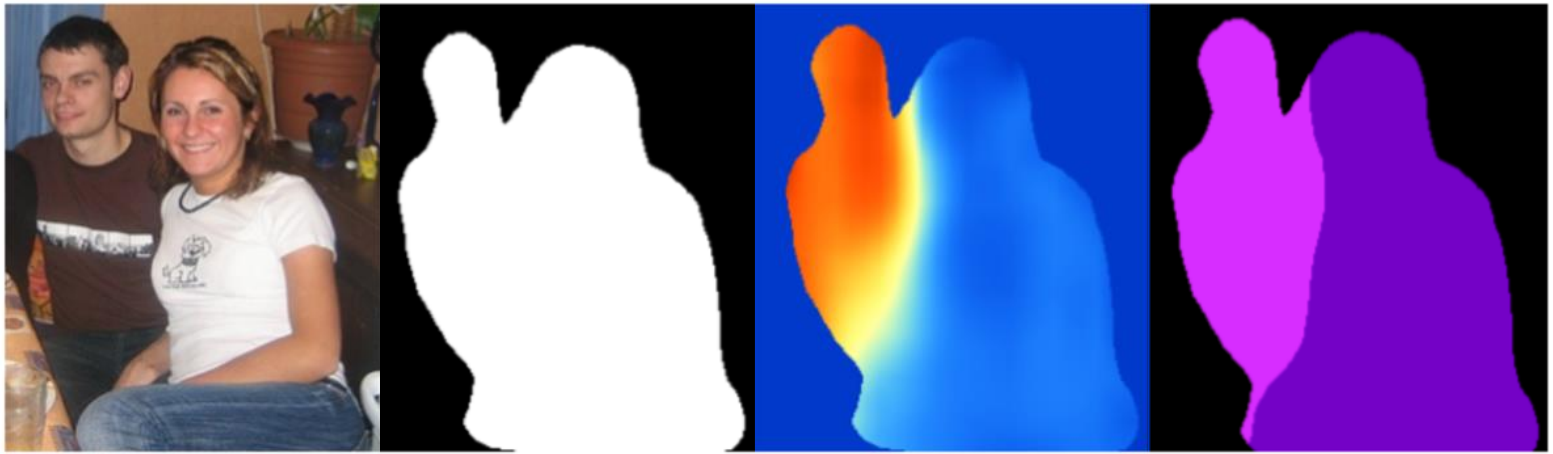
Table 3. Results on MS-COCO **test-dev**, excluding systems trained with external data.



Ablated Results

	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
single scale	0.566	0.818	0.618	0.498	0.670
single scale + refine	0.628	0.846	0.692	0.575	0.706
multi scale	0.630	0.857	0.689	0.580	0.704
multi scale + refine	0.655	0.868	0.723	0.606	0.726

Instance Segmentation



Loss Function

h - predicted tag heatmap
h(x) - tag value at pixel x
$S_n = x_{kn}$ - random sampled k locations with the nth object

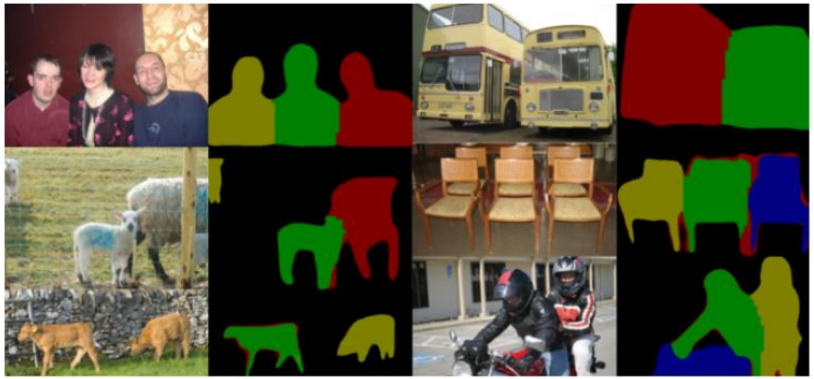
$$L_g(h,T) = \sum_n \sum_{x \in S_n} \sum_{x' \in S_n} (h(x) - h(x'))^2$$

$$+ \sum_n \sum_{n'} \sum_{x \in S_n} \sum_{x' \in S_{n'}} \exp\{-\frac{1}{2\sigma^2}(h(x) - h(x'))^2\}$$

Implementation Details

- Pascal VOC 2012 - 1449 evaluation images
- Four Stacked Hourglass Modules
- 256x 256 Input
- 64 x 64 Output
- Ignored objects that are too big or too small

Pascal VOC Results



	IOU=0.5	IOU=0.7
SDS [22]	49.7	25.3
Hypercolumn [23]	60.0	40.4
CFM [11]	60.7	39.6
MPA [36]	61.8	–
MNC [12]	63.5	41.5
Our method	35.1	26.0



Thank you!
Questions?