Introduction

Confidence Maps and Part Affinity Fields are two methods to show the possibility of a single pixel belongs to a certain body part. Thus we can use CNN to train and our networks learn to predict them.And with these predicts,we can use a bipartite graph matching algorithm to match body parts to do the pose estimation.

Implementing Details

Confidence Maps for Part Detection

$$\mathbf{S}^*_{j,k}(\mathbf{p}) = \exp\left(-\frac{||\mathbf{p} - \mathbf{x}_{j,k}||_2^2}{\sigma}\right) \quad \mathbf{S}^*_j(\mathbf{p}) = \max_k \mathbf{S}^*_{j,k}(\mathbf{p}),$$

x(j,k)be the ground-truth position of body part j for person k in the image,the value of location p in the confidence map S(j,k) for person k is defined as above.

Using the Confidence Maps,we can define E to be an association score between a part candidate of j1 at position d(j1) and a part candidate of j2 at position d(j2).

$$E = \int_{u=0}^{u=1} \mathbf{S}_c\left(\mathbf{p}(u)\right) du, \qquad \mathbf{p}(u) = (1-u)\mathbf{d}_{j_1} + u\mathbf{d}_{j_2}.$$

But above method encode only the location information and eschew orientation information of the limb.

Part Affinity Fields for Part Association

To solve this, we present a novel feature representation we call a part affinity field that preserves both location and orientation information required to perform body part association.

$$l_{c,k} = ||\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}||_2 \quad \mathbf{v} = l_{c,k}^{-1}(\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k})$$

L is the length of the limb,v is the unit vector in the direction of the limb.

We define the ideal part affinity vector field, L*(c,k), at an image point p as

$$\mathbf{L}^*_{c,k}(\mathbf{p}) = \begin{cases} \mathbf{v} & \text{if } \mathbf{p} \text{ on limb } c,k \\ \mathbf{0} & \text{otherwise,} \end{cases}$$

where the set of points on the limb is defined as those within a distance threshold of the line segment.

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

The ideal part affinity field to be predicted by the network combines the limbs of type c of all people into a single map.

$$\mathbf{L}_c^*(\mathbf{p}) = \frac{1}{n_p} \sum_k \mathbf{L}_{c,k}^*(\mathbf{p}),$$

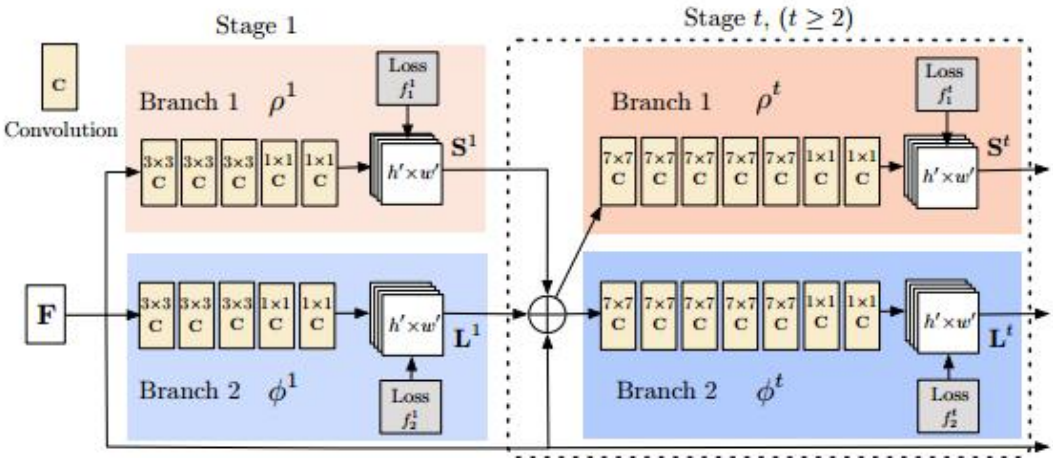
$$E = \int_{u=0}^{u=1} \mathbf{L}_c\left(\mathbf{p}(u)\right) \cdot \frac{\mathbf{d}_{j_2} - \mathbf{d}_{j_1}}{\|\mathbf{d}_{j_2} - \mathbf{d}_{j_1}\|_2} du,$$

Same as above,E is an association score .

Bipartite Graph Matching Algorithm

Now we know the weight of body part j1 connect to j2.We can use hungarian algorithm to match body parts consisting limbs in order to maximize the sum of weight.

Two-branch Multi-stage CNN



$$f_1^t = \sum_{j=1}^{J} \sum_{\mathbf{p} \in \mathbf{I}} \mathbf{W}(\mathbf{p}) \cdot \|\mathbf{S}_j^t(\mathbf{p}) - \mathbf{S}_j^*(\mathbf{p})\|_2^2,$$

$$f_2^t = \sum_{c=1}^{C} \sum_{\mathbf{p} \in \mathbf{I}} \mathbf{W}(\mathbf{p}) \cdot \|\mathbf{L}_c^t(\mathbf{p}) - \mathbf{L}_c^*(\mathbf{p})\|_2^2,$$

W is the binary mask, W(p) = 0 when p is in the area of persons without annotation.

$$f = \sum_{t=1}^{T} (f_1^t + \lambda f_2^t),$$

where λ is a weighting factor that is empirically found to work best when λ = 1.

Results

Results on the MPII Multi-Person Dataset

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	**mAP**	s/image
Subset of 288 images as in [27]									
Deepcut [27]	73.4	71.8	57.9	39.9	56.7	44.0	32.0	54.1	57995
Iqbal et al. [16]	70.0	65.2	56.4	46.1	52.7	47.9	44.5	54.7	10
DeeperCut [15]	87.9	84.0	71.9	63.9	68.8	63.8	58.1	71.2	230
Ours - 6 stages -ms	**93.7**	**91.4**	**81.4**	**72.5**	**77.7**	**73.0**	**68.1**	**79.7**	**1.1**
full testing set									
DeeperCut [15]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5	485
Iqbal et al. [16]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1	10
Ours-2 stages -ms	90.4	85.5	73.0	60.1	71.2	60.1	51.2	70.2	0.98
ours-6 stages	89.0	84.9	74.9	64.2	71.0	65.6	58.1	72.5	**0.6**
Ours-6 stages -ms	**91.2**	**87.6**	**77.7**	**66.8**	**75.4**	**68.9**	**61.7**	**75.6**	1.24

Using the model with 6 stages further increases the performance to

75.6% AP. The AP comparison with previous bottom-up approaches indicate the effectiveness of our novel feature representation, PAFs, to associate body parts.

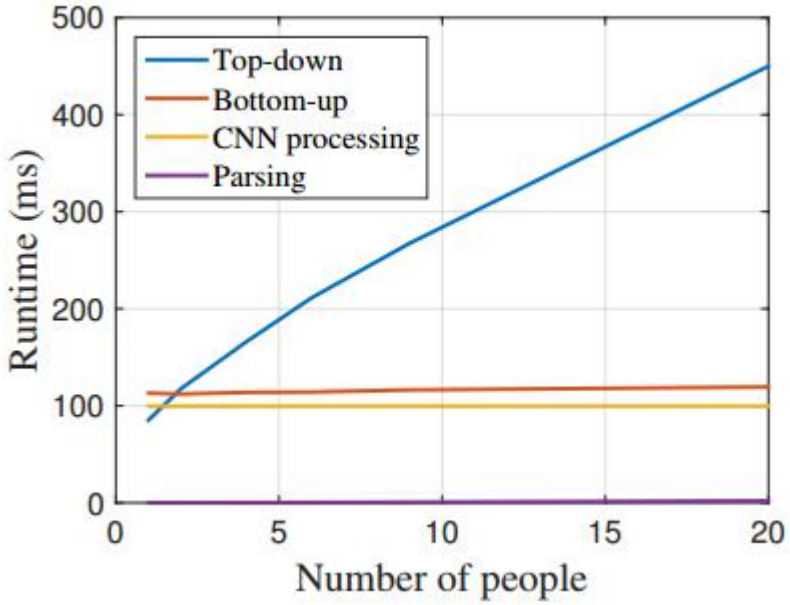
Results on the MSCOCO Keypoints Challenge

Team	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Test-challenge					
Ours	**60.5**	**83.4**	**66.4**	55.1	**68.1**
G-RMI	59.8	81.0	65.1	**56.7**	66.7
DL-61	53.3	75.1	48.5	55.5	54.8
R4D	49.7	74.3	54.5	45.6	55.6
Test-dev					
Ours	**61.8**	**84.9**	**67.5**	57.1	**68.2**
G-RMI	60.5	82.2	66.2	**57.6**	66.6
DL-61	54.4	75.3	50.9	58.3	54.3
R4D	51.4	75.0	55.9	47.4	56.7

We outperform other teams, which all use top-down methods. It is the noteworthy that our method has won, but has lower accuracy than the top-down methods on people of smaller scales (APM). The reason is that our method has to deal with a much larger scale range spanned by all people in the image in one shot. In contrast, top-down methods rescale the patch of each detected area to preferable size independently and thus suffer less from small people.

Method	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
GT + CPM [15]	62.7	86.0	69.3	58.5	70.6
SSD[22]+CPM[15]	52.7	71.1	57.2	47.0	64.2
Ours - 6 stages	58.4	81.5	62.6	54.4	65.1
Ours + refinement	61.0	84.9	67.5	56.3	69.3

Runtime Analysis



CNN's runtime complexity is O(1).Multi-person parsing's runtime complexity is O(n*n).However, the parsing time does not signifi- cantly influence the overall runtime because it is two orders of magnitude less than the CNN processing time.

Some Ideas

Maybe we can employ STN(spatial transform network) before our model.