Supplementary Material for Joint Multi-Person Pose Estimation and Semantic Part Segmentation

Fangting Xia[1] sukixia@gmail.com
Peng Wang[1] pengwangpku2012@gmail.com
Xianjie Chen[1] cxj@ucla.edu
Alan Yuille[2] alan.yuille@jhu.edu

[1]University of California, Los Angeles
Los Angeles, CA 90095

[2]Johns Hopkins University
Baltimore, MD 21218

Abstract

In this supplementary material, we provide a detailed introduction of our novel segment-joint smoothness term used in the fully-connected conditional random field (FCRF) for human pose estimation.

1. Segment-Joint Smoothness Term

In our pose estimation model, we train a FCRF to select and assemble joint location proposals into valid pose configurations. The graph of the FCRF is formulated as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the node set $\mathcal{V} = \{c_1, c_2, \ldots, c_n\}$ represents all the candidate locations of joints and the edge set $\mathcal{E} = \{(c_i, c_j)|i = 1, 2, \ldots, n, j = 1, 2, \ldots, n, i < j\}$ is the edges connecting all of the locations. The labels we want to predict are: $\{l_{c_i}|c_i \in \mathcal{V}\}$, the joint type for each node, and $\{l_{c_i,c_j}|(c_i, c_j) \in \mathcal{E}\}$ which indicates whether two nodes belong to the same person.

As explained in the main paper, the unary term for the FCRF is defined by deep-learned joint scores (*i.e.* output of Pose FCN) while the pairwise term is determined by both the joint neighbor score map $\mathbf{P}_n$ and the part segmentation score map $\mathbf{P}_s$. For any joint location pair c_i and c_j with label l_{c_i} and l_{c_j}, we use $\mathbf{P}_n$ and $\mathbf{P}_s$ to compute a feature vector $\mathbf{f}(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_n, \mathbf{P}_s)$ (see Equ. 1), based on which we perform logistic regression to predict whether c_i and c_j belong to the same person, and use the prediction score to compute the pairwise term for FCRF.

$$\mathbf{f}(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_n, \mathbf{P}_s) = [\mathbf{f}(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_n) \quad \mathbf{f}(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_s)] \tag{1}$$

Here, we focus on $\mathbf{f}(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_s)$, which describes "segment-joint smoothness", *i.e.* the compatibility between joint locations and part segmentation map. In our design, each joint type is associated with one or two semantic parts (see Tab. 1) while each neighbouring joint type pair is also associated with one semantic part (see Tab. 2). As shown in Equ. 2, $\mathbf{f}(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_s)$ includes the following three types of features: (1) $\mathbf{f}_u(c_i, l_{c_i}|\mathbf{P}_s)$, the compatibility feature of joint type l_{c_i} (with location c_i) and its associated semantic part/parts; (2) $\mathbf{f}_u(c_j, l_{c_j}|\mathbf{P}_s)$, the compatibility feature of joint type l_{c_j} (with location c_j) and its associated semantic part/parts; (3) $\mathbf{f}_p(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_s)$, the compatibility feature between joint type pair (l_{c_i}, l_{c_j}) and its associated semantic part.

Joint Type	Associated Semantic Part/Parts
forehead	head
neck	head torso
left/right shoulder	torso upper-arm
left/right elbow	upper-arm lower-arm
left/right wrist	lower-arm
left/right waist	torso upper-leg
left/right knee	upper-leg lower-leg
left/right ankle	lower-leg

Table 1: The full list of joint type and its associated semantic part/parts.

Joint Type Pair	Associated Semantic Part
forehead & neck	head
neck & left/right shoulder	torso
left/right shoulder & left/right elbow	upper-arm
left/right elbow & left/right wrist	lower-arm
neck & left/right waist	torso
left/right waist & left/right knee	upper-leg
left/right knee & left/right ankle	lower-leg

Table 2: The full list of joint type pair and its associated semantic part.

$$\begin{aligned}\mathbf{f}(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_s) = [&\mathbf{f}_u(c_i, l_{c_i}|\mathbf{P}_s) \\ &\mathbf{f}_u(c_j, l_{c_j}|\mathbf{P}_s) \quad \mathbf{f}_p(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_s)]\end{aligned} \tag{2}$$

For joint type l_{c_i} with location c_i, suppose its associated semantic parts are p_1 and p_2. We set $p_2 = p_1$ if l_{c_i} is associated with only one semantic part. $\mathbf{f}_u(c_i, l_{c_i}|\mathbf{P}_s)$ is a 4-d binary feature vector: the first dimension indicates whether c_i is inside semantic part p_1; the second dimension shows whether c_i is around the boundary of p_1 (within 10 pixels from the boundary); the third dimension indicates whether c_i is inside semantic part p_2; the fourth dimension shows whether c_i is around the boundary of p_2.

For joint type l_{c_i} with location c_i, suppose its associated semantic parts are p_1 and p_2. We set $p_2 = p_1$ if l_{c_i} is associated with only one semantic part. $\mathbf{f}_u(c_i, l_{c_i}|\mathbf{P}_s)$ is a 4-d binary feature vector: the first dimension indicates whether c_i is inside semantic part p_1; the second dimension shows whether c_i is around the boundary of p_1 (within 10 pixels from the boundary); the third dimension indicates whether c_i is inside semantic part p_2; the fourth dimension shows whether c_i is around the boundary of p_2.

For joint type pair l_{c_i} and l_{c_j} with location c_i and c_j respectively, suppose this joint type pair is associated with semantic part p_3. $\mathbf{f}_p(c_i, c_j, l_{c_i}, l_{c_j}|\mathbf{P}_s)$ is a 2-d feature, with the first dimension being the proportion of pixels on the line segment connecting c_i and c_j that fall inside semantic part p_3, and the second dimension being the intersection-over-union (IOU) between an oriented rectangle computed from c_i and c_j (with aspect ratio = 2.5 : 1) and part region p_3. If the joint type pair is not neighbouring to each other, this 2-d feature is set to 0.