FastPose: A Fast And Small Human Pose Estimator

Zexin Chen
Shanghai Jiao Tong University
czxsjtu@gmail.com

Abstract

FastPose is a bottom-up human pose estimation method. A neural networks is used to predict the 2D locations of anatomical keypoints (blue point in Fig2) as well as the middle point of adjoining anatomical keypoints (red point in Fig2) for all person in the image. Then the anatomical keypoints are grouped according to the middle points. Since middle points and anatomical keypoints share more similarity than Part Affinity Fields[1] and anatomical keypoints does, we can use a 46% smaller and faster neural network to do the prediction compared to OpenPose[1].

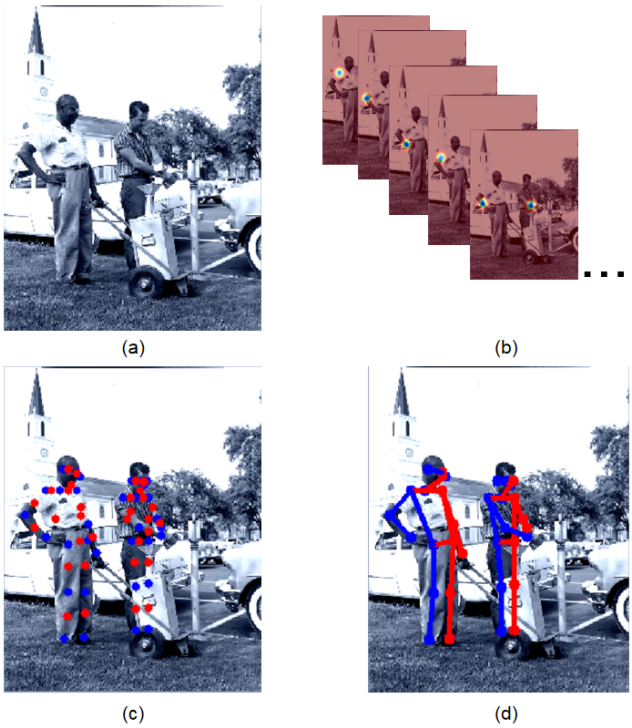

Figure 1. Overall Pipeline

1. Method

Fig. 1 illustrates the overall pipeline of our method.The system takes, as input, a color image of size w x h (Fig.1a) and produces, as output, the 2D locations of anatomical keypoints and the middle point of some anatomical keypoints for each person in the image (Fig. 1c). More specifically, a feedforward network predicts a set of 2D confidence maps S of body part locations (Fig. 1b). The set $S = (S_1; S_2; ...; S_J)$ has J confidence maps, one per part, where $Sj \in R^{wxh}, j \in \{1, ..., J\}$.The total number of the all body part is 34, including 17 anatomical key points, blue point in Fig. 2, and 17 middle point of limbs, red point in Fig. 2. Finally, the confidence maps are parsed and the middle points are used to group the anatomical key points to output the 2D keypoints for all people in the image (Fig. 1d). The grouping detail is in section 1.2.

Figure 2. All Predicted Body Part

1.1. Confidence Maps for Part Detection

During training, we first generate the groundtruth coordinate of the middle body part from the coordinates of the adjoining anatomical keypoints, which is simply the geometric middle point of two adjoining anatomical keypoints. Then we generate the groundtruth confidence maps S from

Figure 3. Grouping Process

the annotated 2D keypoints as well as the middle body part keypoints. Each confidence map is a 2D representation of the belief that a particular body part occurs at each pixel location. Ideally, if a single person occurs in the image, a single peak should exist in each confidence map if the corresponding part is visible; if multiple people occur, there should be a peak corresponding to each visible part j for each person k. The final groundtruth is J heatmaps each of which have up to k peaks.

To generate the final confidence maps, we first generate individual confidence maps $S^*_{j,k}$ for each person k. Let $x_{j,k} \in \mathbb{R}^2$ be the groundtruth position of body part j for person k in the image. The value at location $p \in \mathbb{R}^2$ in $S^*_{j,k}$ is defined as,

$$S^*_{j,k}(p) = exp(-\frac{||p - x_{j,k}||}{\sigma^2}), \tag{1}$$

where σ controls the spread of the peak. The groundtruth confidence map to be predicted by the network is an aggregation of the individual confidence maps via a max operator,

$$S^*_j(p) = max_k S^*_{j,k}(p), \tag{2}$$

We take the maximum of the confidence maps instead of the average so that the precision of close by peaks remains distinct. At test time, we predict confidence maps and obtain body part candidates by performing non-maximum suppression.

1.2. Middle Points for Anatomical Key Points Association

All of the adjoining anatomical key points of the same person are rigidly linked to each other, for example, the right wrist and the right albow are rigidly connected by the right forearm. In fact, from any perspective, the geometric middle point of the right wrist and the right albow in a picture is always the middle part of the forearm. The other adjoining anatomical key points share the same condition. This indicates that there are some certain pattern of the middle points of the adjoining anatomical key points that might be predictable.

we can use the middle points to group the anatomical key points into different persons. Since all of the adjoining anatomical key points of the same person are rigidly linked to each other, if two adjoining anatomical key points belongs to one single person, then there would definitely be a middle body part at the geometric middle point of the two adjoining anatomical key points. That is, only if there is a middle body part at the geometric middle point of the two adjoining anatomical key points, can two adjoining anatomical key points belongs to one single person. This feature make it possible to use middle point as a confidence measure of the association for each pair of body part detections, i.e., that they belong to the same person. As is shown in Fig.3 We start from one anatomical key point, saying right wrist. First, the adjoining anatomical key point (right albow) of the starting anatomical key point (right wrist) is found out by checking whether there is a middle body part (middle forearm) at the geometric middle point. And then the adjoining anatomical key point (right should) of the the adjoining anatomical key point (right albow) is found out by checking the middle body part (middle upper arm) and so on to get all of the anatomical key points of one person. The anatomical key points of the persons are determined in the same way. After linking all the anatomical key point predicted, we can get the anatomical key of all people in the picture.

To make the grouping more robust, we assume that the real adjoining anatomical key point pair is the one that has the shortest distance and the angle of some two interconnected keypoints pair, like the right-shoulder-right-shoulder pair and the right-shoulder-right-hip, are tend to be near 90 degree. These assumptions can exclude almost all the exception case where that middle part might no working due to duplicate middle points.

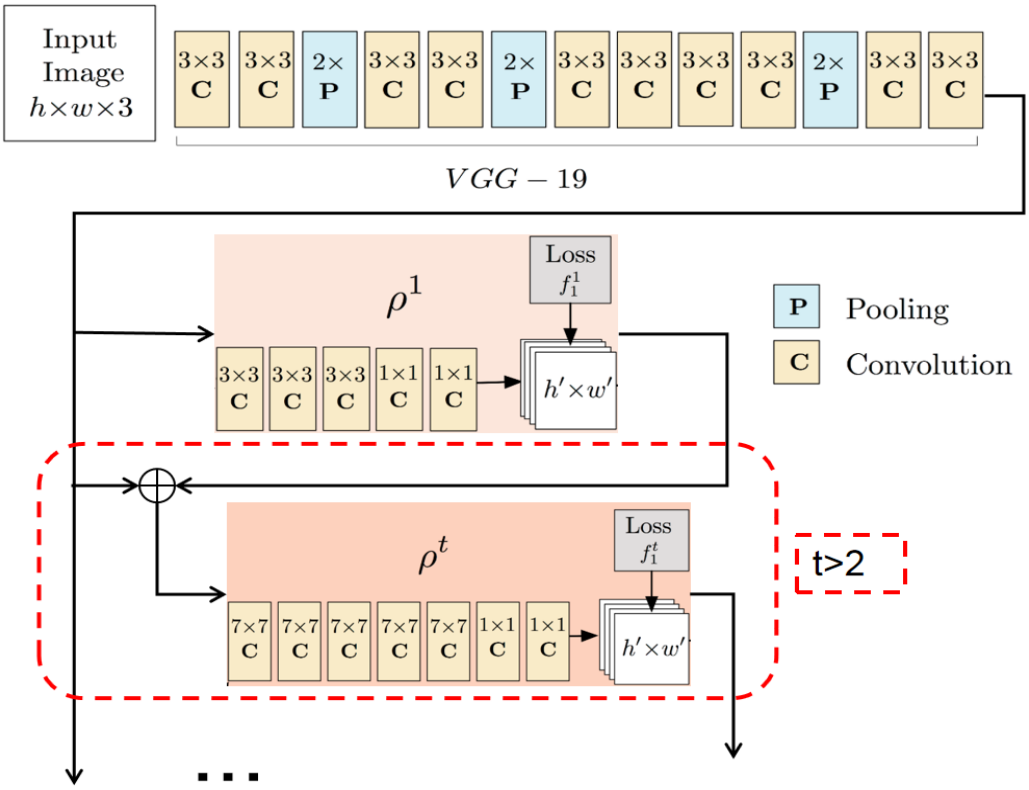

Figure 4. Feedforward Network Architecture

1.3. Feedforward Network

One advantage brought by using middle body part as a confidence measure of the association for each pair of body part detections is that there are some similarity between the middle points and the anatomical keypoints so that we can regard the middle body part as anatomical keypoints and predict them in the same way as predicting the anatomical keypoints. Without the need extra network branch, we can have a more light-weight and more efficient network.

The feedforward network shown in Fig4, take a three channels picture as input and predicts detection confidence maps of the anatomicla keypoints and the middle body part as output. We use VGG as our backbone and follow CMP[2] to use multi-stages to refine the heatmap prediction.

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[2] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.