Multi-person Pose Estimation

Zuoyue Li

1 Introduction

1.1 Problem Definition

The task of multi-person pose estimation in an image is to identify every person instance and to localize its facial and body keypoints. Specifically, given an RGB image $I = \{0,1,\dots,255\}^{H\times W\times 3}$, our goal is to find a set of person keypoints $K = \{K_{ij}\}_{i=1,\dots N, j=1,\dots,M}$, where N is the number of people in the image, M is the number of keypoints to be detected for a single person, K_{ij} refers to the pixel coordinates of i-th person's j-th keypoint. Usually, the value of M should be well set in advance, i.e. each index should correspond to a joint of a person.

1.2 Frameworks

There are two types of framework for tackling the problem of multi-person pose estimation: the top-down approach and the bottom-up approach, both of which are illustrated as follows:
(1) The top-down approach starts by identifying and localizing individual person instances roughly by means of a bounding box object detector, followed by single-person pose estimation.
(2) The bottom-up approach starts by localizing identity-free semantic entities, i.e. individual keypoint proposals, followed by grouping them into person instances.

Both of the two pipelines are exactly two-step frameworks. In the top-down approaches, the second step, single-person pose estimation, relies heavily on the first step, human detection. If the bounding box of a person found is determined as false positive, the second step would tend to make false prediction. As for top-down approaches, although they are box-free, it usually requires additional information to group keypoints into a person instance, because the only keypoints proposal itself is not enough. If the additional joints connection information gives false guidance, the pose decoding scheme would further break down. This often appears in people crowd.

As for the runtime comparison, the bottom-up approaches are usually more efficient than the top-down approaches, because the time spent by the second step, single-person pose estimation of top-down approaches is proportional to the number of people in the image. By contrast, bottom-up approaches do not have the problem of computing efficiency in general.

1.3 Motivation

It is mentioned above that in the top-down approach, the decision of person bounding box is very important. It determines whether the single-person pose estimator can accurately predict the human pose or not. Although the existing networks for object detection are already very mature, we still hope that it can be further improved.

We believe that the heatmaps generated by bottom-up approaches can help the human detection scheme. This is because if the heatmaps do not show any peak (or only have some smoothed peaks) within the area decided by a positive human bounding box, then this bounding box is likely to be false positive. Similarly, if there is no bounding box wrapping an area with some obvious peaks, then there may be one or more missing bounding boxes (false negative) which are not retrieved.

By contrast,

By contrast, We develop a fully convolutional system whose com- putational cost is essentially independent of the number of people present in the scene and only depends on the cost of the CNN feature extraction backbone.

In particular, our approach first predicts all keypoints for every person in the image in a fully convolutional way. We also learn to predict the relative displacement between each pair of keypoints, also proposing a novel recurrent scheme which greatly improves the accuracy of long-range predictions. Once we have localized the keypoints, we use a greedy decoding process to group them into instances. Our approach starts from the most confident detection, as opposed to always starting from a distinguished landmark such as the nose, so it works well even in clutter.

In addition to predicting the sparse keypoints, our system also predicts dense instance segmentation masks for each person. For this purpose, we train our network to predict instance-agnostic semantic person segmentation maps. For every person pixel we also predict offset vectors to each of the K keypoints of the corresponding person instance. The corresponding vector fields can be thought as a geometric embedding representation and induce basins of attraction around each person instance, leading to an efficient association algorithm: For each pixel xi, we predict the locations of all K keypoints for the corresponding person that xi belongs to; we then compare this to all candidate detected people j (in terms of average keypoint distance), weighted by the keypoint detection probability; if this distance is low enough, we assign pixel i to person j.

We train our model on the standard COCO keypoint dataset [1], which an- notates multiple people with 12 body and 5 facial keypoints. We significantly outperform the best previous bottom-up approach to keypoint localization [2], improving the keypoint AP from 0.655 to 0.687. In addition, we are the first bottom-up method to report competitive results on the person class for the COCO instance segmentation task. We get a mask AP of 0.417, which outper- forms the strong top-down FCIS method of [3], which gets 0.386. Furthermore our method is very simple and hence fast, since it does not require any second stage box-based refinement, or clustering algorithm. We believe it will therefore be quite useful for a variety of applications, especially since it lends itself to deployment in mobile phones.

Multi-person Pose Estimation