Combining Local Appearance and Holistic View: Dual-Source Deep Neural Networks for Human Pose Estimation

Xiaochuan Fan, Kang Zheng, Yuewei Lin, Song Wang
Department of Computer Science & Engineering, University of South Carolina, Columbia, SC 29208, USA

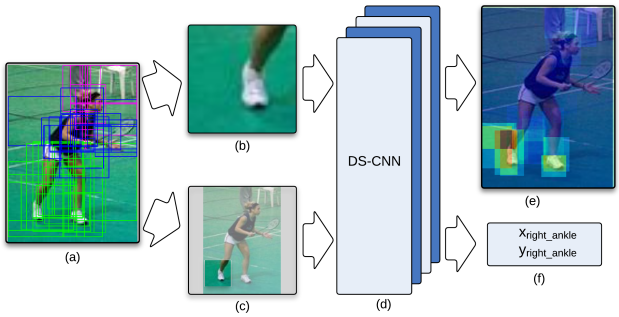

Figure 1: An illustration of the proposed method based on DS-CNN. (a) Input image and generated image patches. (b) DS-CNN input on an image patch (containing a local part – ankle). (c) DS-CNN input on full body and holistic view of the local part in the full body. (d) DS-CNN for learning. (e) DS-CNN output on joint detection. (f) DS-CNN output on joint localization.

By accurately locating the important body joints from 2D images, human pose estimation plays an essential role in computer vision. In this paper, we propose a new learning-based method for estimating human pose from a single image, using Dual-Source Deep Convolutional Neural Networks (DS-CNN).

Most of the previous works on human pose estimation are based on the two-layer part-based model. The first layer focuses on local (body) part appearance and the second layer imposes the contextual relations between local parts. These pose estimation methods using part-based models are usually sensitive to noise and the graphical model lacks expressiveness to model complex human poses. Furthermore, most of these methods search for each local part independently and the local appearance may not be sufficiently discriminative for identifying each local part reliably.

Recently, due to deep convolutional neural network (CNN)'s large learning capacity and robustness to variations, there is a natural rise in the interest to directly learn high-level representations of human poses without using hand-crafted low-level features and even graphical models. Toshev et al. [3] present such a holistic-styled pose estimation method named DeepPose using DNN-based joint regressors. This method also uses a two-layer architecture: The first layer resolves ambiguity between body parts in a holistic way and provides an initial pose estimation and the second layer refines the joint locations in a local neighborhood around the initial estimation. From the experiments in [3], DeepPose can achieve better performance than several recent comparison methods. However, DeepPose does not consider local part appearance in initial pose estimation. As a result, it has difficulty in estimating complex human poses, even using the CNN architecture.

In this paper, we propose a dual-source CNN (DS-CNN) based method for human pose estimation, as illustrated in Fig. 1. This proposed method integrates both the local part appearance in image patches and the holistic view of each local part for more accurate human pose estimation. Following the region-CNN (R-CNN) that was developed for object detection [1], the proposed DS-CNN takes a set of category-independent object proposals detected from the input image for training. Compared to the sliding windows or the full image, that are used as the input in many previous human pose estimation methods, object proposals can capture the local body parts with better semantic meanings in multiple scales. Moreover, we extend the original single-source R-CNN to a dual-source model (DS-CNN) by including the full body and the holistic view of the local parts as a separate input, which provides a holistic view for human pose estimation. By taking both the local part object proposals and the full body as inputs in the training stage, the proposed DS-CNN performs a unified learning to achieve both joint detection, which determines whether an object proposal contains a body joint, and joint localization, which finds the exact location of the joint in the object proposal. In the testing stage, we use multi-scale sliding windows to provide local part information in order to avoid the performance degradation resulted by the uneven distribution of object proposals. Based on the DS-CNN outputs, we combine the joint detection results from all the sliding windows to construct a heatmap that reflects the joint location likelihood at each pixel and weightedly average the joint localization results at the high-likelihood regions of the heatmap to achieve the final estimation of each joint location. The proposed method is implemented using the open-source CNN library Caffe and therefore has good expandability.

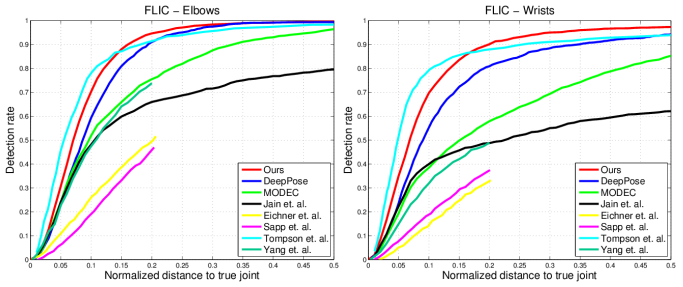

Figure 2: PDJ comparison on FLIC.

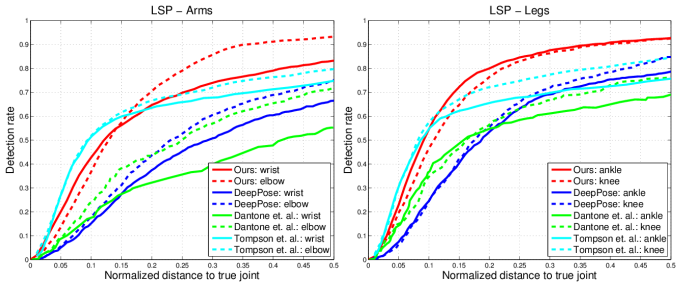

Figure 3: PDJ comparison on LSP.

In the experiments, we test the proposed method on two widely used datasets, FLIC and LSP, and compare its performance to several recently reported human pose estimation methods, including DeepPose, using two popular metrics: Percentage of Corrected Parts (PCP) and Percentage of Detected Joints (PDJ). Figure 2 and 3 show the PDJ curves of the proposed method and seven comparison methods. The results show that the proposed method can produce superior performance against all comparison methods except for Tompson et al [2]. When compared with Tompson et al [2], the proposed method performs better when normalized distance is large and performs worse when normalized distance is small.

Moreover, we also found an interesting result, the average precision (AP) of the joint detection when taking only the full body as the input to CNN. In our method, the full body input actually includes local part information in the form of an alpha channel (holistic view). As a result, the use of only the full body can lead significantly better AP than the use of only the local part appearance (R-CNN).

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[2] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.

[3] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *CVPR*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.