Replication Studies on a State-of-the-art Part-based Human Detector

Yu-Wei Chao
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109, USA
ywchao@umich

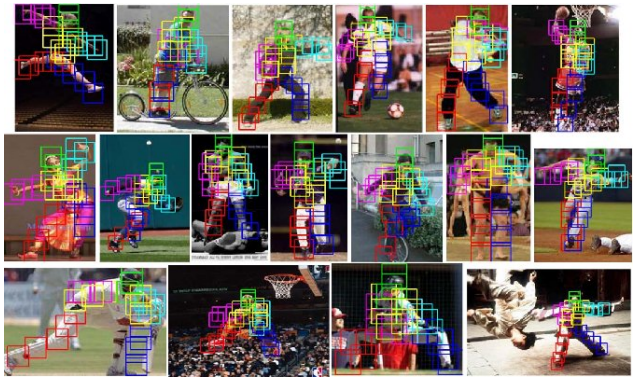

Figure 1: Illustration of Yang and Ramanan's work on human detection and pose estimation [6]. Different color of boxes identify different local parts of human body.

1. Introduction

Human detection and pose estimation is a challenging problem in computer vision studies. Recently there has been many outstanding works published in addressing the these problems [2, 1, 3, 6, 5]. A very interesting line of work is the application of part-based models [3, 6]. Part-based models can be viewed an extension of the rigid template models in the way that the target objects are represented by local parts, and the locations of these local parts have some amount of flexibility to capture the uncertainly in real-world data.

In this work, we proposed to replicate Yang and Ramanan's paper [6] on human detection and pose estimation. Unlike the famous deformable part-based model [3], which addressed generic object detection, Yang and Ramanan focus specifically on human detection in RGB images. They carefully designed a system that can locate the body parts of the human, as illustrated in Fig. 1. They used a mixtures of templates for each parts so they can effetively modeled human poses.

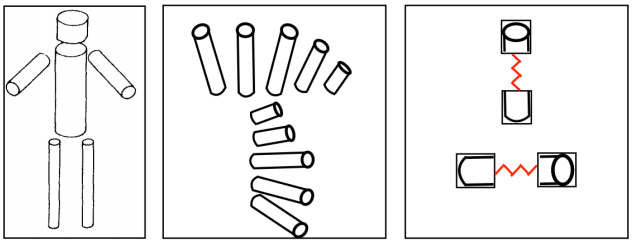

Figure 2: Illustration of the pictorial structure and mixture of local parts.

2. Replication Plan

We list our plans for this replication study as following:

- Re-implement the inference algorithm. We plan to investigate the efficiency of the inference algorithm, which is not presented in the paper.
- Re-do the training process. The paper uses a fixed training set provided by the Image Parse dataset and the Buffy Stickman dataset. We plan to re-train the model with different settings, either using a subset or on a new dataset. The goal is to better understand how the trained model changes by different training settings.
- Build on own human dataset. We can build a small but more challenging dataset, and see how far can the target method go.

3. Inference

We first briefly review the overall pictorial structure model for human detection proposed in [6] and the inference problem in Sec. 3.1. Next we evaluate our implementation of the inference algorithm on the the Image Parse data [4] in Sec. 3.2.

3.1. Algorithm Overview

Yang *et al.* proposed a pictorial structure representation to robustly model the human poses caused by articulated body configurations. They have a flexible design in the way that each local part is a mixture of different component (i.e. can be chosen from a template set). Let I denote the image, $l_i = (x_i, y_i)$ be the location of part i, and t_i be the mixture component of part i (using the t_ith template for part i), where $i \in \{1, \ldots, K\}$, $l_i \in \{1, \ldots, L\}$, and $t_i \in \{1, \ldots, T\}$. Assuming the human body is a tree structured graph $G = (V, E)$, the score function of a human is defined as following:

$$S(I, l, t) = \sum_{i \in V} \omega_i^{t_i} \cdot \phi(I, l_i) + \sum_{i,j \in E} \omega_{ij}^{t_i, t_j} \cdot \psi(l_i - l_j) + S(t). \tag{1}$$

The three terms in eq. 1 correspond to the appearance model, deformation model, and the co-occurence model. Appreance model computes the local scores by placing a part template $w_i^{t_i}$ (of type t_i) at the location i. Deformation model controls the relative position of part i and j. Note that $\psi(l_i - l_j) = [dx \;\; dx^2 \;\; dy \;\; dy^2]^\top$. The co-occurrence model

$$S(t) = \sum_{i \in V} b_i^{t_i} + \sum_{i,j \in E} b_{ij}^{t_i, t_j} \tag{2}$$

captures the occurrence likelihood of part i with type t_i and the co-occurrence likelihood of part i with type t_i and part j with type j.

Given the model parameter, inference corresponds to finding the maximum scoring locaions (l, t) given the image I. Denote $z_i = (t_i, l_i)$, the eq 1 can be written as

$$\begin{aligned} S(I, z) &= \sum_{i \in V} \phi_i(I, z_i) + \sum_{i,j \in E} \psi_{ij}(z_i, z_j), \\ \text{where} \quad \phi_i(I, z_i) &= \omega_i^{t_i} \cdot \phi(I, l_i) + b_i^{t_i} \\ \psi_{ij}(z_i, z_j) &= \omega_{ij}^{t_i, t_j} \cdot \psi(l_i - l_j) + b_{ij}^{t_i, t_j} \end{aligned} \tag{3}$$

One nice consequence of the tree structure assumption on human body representation is enabling us to solve the optimization problem efficiently by dynamic programming. Our first focus of this replication study is on re-implementing this inference algorithm.

3.2. Evaluation

We verify our implementation by first comparing the result on the same benchmark dataset, Image Parse Dataset [4]. The dataset consists of 100 training images and 205 test images. Since we only want to verify our inference implementation, we use only the test images for experiment. As [4], we evaluate the result using two different metrics: 1) probability of a correct keypoint (PCK), and 2) average precision of keypoints (APK). The comparison between our implementation and the original paper is reported in Table 1 and 2. Note that the authors have made their code public online. Since we observe a difference in the result obtained by running their released code and the number they reported in [4], we use the numbers obtained by their release code for comparison here.

Parse	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[6]	88.3	82.7	61.2	44.1	68.8	66.8	61.0	67.6
Mine	77.3	71.2	47.3	31.7	58.8	56.8	49.3	56.1

Table 1: Probability of correct keypoints (PCK) on the Image Parse Dataset.

Parse	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[6]	83.9	77.0	49.6	26.8	56.1	52.3	43.7	55.6
Mine	71.2	62.8	32.4	15.8	42.9	38.8	33.1	42.4

Table 2: Average precision of keypoints (APK) on the Image Parse Dataset.

For the quantitative evaluation, we observe a significant gap between our implementation and the author's implementation. This indicates that our implementation of inference clearly still contain mistakes. We also showed the qualitative examples for detected human in Fig. 3. Looking through the example image, we see that the implementation actually works in a reasonbly good level. This suggests that the mistake might be minor. A good strategy for debugging the implementation will be comparing with the author's code by fixing the module of the algorithm one at a time. We plan to carry out that as our next step.

4. Training

In progress. Aim to finish before 03.18.

5. New Dataset

In progress. Aim to finish before 04.08.

6. Milestone

The milestones are shown in Table 3

References

[1] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *Proceedings of the 11th European Conference on Computer Vision*, 2010. 1

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Proceed-*

Parse	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[6] Tr + [6] Inf	88.3	81.7	62.2	42.0	71.0	68.5	61.2	67.8
[6] Tr + My Inf	78.0	70.0	40.5	28.3	58.8	56.3	52.4	54.9
My Tr + [6] Inf	87.3	79.8	58.8	43.7	70.0	63.9	58.3	66.0
My Tr + My Inf	79.0	71.7	40.7	30.0	57.8	58.3	54.1	56.0

Table 4: Probability of correct keypoints (PCK).

Parse	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[6] Tr + [6] Inf	83.9	77.6	45.7	24.2	60.4	53.8	46.5	56.0
[6] Tr + My Inf	69.3	60.7	24.5	10.4	42.3	36.6	33.6	39.6
My Tr + [6] Inf	83.2	75.7	44.9	26.5	57.0	49.2	41.9	54.1
My Tr + [6] Inf	72.5	63.6	25.1	13.8	42.7	43.0	37.1	42.5

Table 5: Average precision of keypoints (APK).

Date	Plan
02.05 - 02.11	re-implement the inference algorithm
02.12 - 02.18	re-implement the inference algorithm
02.19 - 02.25	re-implement the inference algorithm compare the result with the paper analyze the result
02.25 - 03.04	re-do training
03.05 - 03.11	re-do training
03.12 - 03.18	collect new dataset
03.19 - 03.25	collect new dataset
03.26 - 04.01	run experiment on new dataset
04.02 - 04.08	run experiment on new dataset
04.09 - 04.15	Buffer time
04.16 - 04.21	Prepare for presentation

Table 3: Project milestones

ings of the IEEE International Conference on Computer Vision, 2009. 1

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 1

[4] D. Ramanan. Learning to parse images of articulated bodies. In *Advances in Neural Information Processing Systems 19*, pages 1129–1136. 2006. 1, 2

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1

[6] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2011. 1, 2, 3, 4

	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[6] Tr + [6] Inf	67.9	47.9	29.3	20.7	33.6	31.4	30.0	37.2
[6] Tr + My Inf	60.0	39.3	17.9	19.3	25.0	25.7	30.7	31.1
My Tr + [6] Inf	67.9	51.4	22.9	18.6	34.3	30.7	31.4	36.7
My Tr + My Inf	57.9	38.6	20.7	15.0	24.3	27.1	32.1	30.8

Table 6: Probability of correct keypoints (PCK).

	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[6] Tr + [6] Inf	54.8	35.0	19.4	10.5	22.0	15.7	19.1	25.2
[6] Tr + My Inf	46.2	29.0	9.6	6.5	19.8	15.0	17.3	20.5
My Tr + [6] Inf	49.6	34.5	15.7	7.6	22.8	13.9	18.0	23.2
My Tr + [6] Inf	47.9	32.3	8.0	4.5	15.1	14.6	19.6	20.3

Table 7: Average precision of keypoints (APK).

Figure 3: Replication results on the Parse dataset. The pre-trained model is composed of 26 human parts. Each human detection is visualized by the detected bounding boxes of parts (left) and the skeletons (right). The first two rows show the correct detected human poses. The last row shows failure examples. The failure can be caused by 1) false alarming parts with high score, 2) double counting, and 3) exceptional pose configurations.