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

1. Introduction

Human detection and pose estimation is a challenging problem in computer vision studies. The major difficulties come from that 1) human body is non-rigid and the configurations of the limbs have a large degree of freedom, 2) limbs vary greatly in appearance due to changes in clothing and body shape, as well as changes in viewpoint manifested, and 3) human body parts can be occluded or self-occluded due to the performed activity or interaction with other objects. However, a success in human detection and pose estimation can be very critical, since it will help a large number of tasks such as object detection, activity recognition, and robot nevigation. After decades of efforts, significant progress has been made by the computer vision researchers, but the problem still remains unsolved, and many related papers are published every year.

Recently there has been a number of outstanding published works in addressing human detection and pose estimation [2, 1, 3, 7, 6]. A common strategy of these works is to represent human body by local parts, and build a relational model to perform full body detection. Part-based models can be viewed as 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. An immediate and fundamental question is using what kind of part representation will be helpful for human detection. In [2, 1, 3], human parts are taken to be discriminant local patches in the images of human body, while in [7], parts are explicitly modeled by a list of predefined body joints. Different design rationales capture different strengths in the human detection task, but they are addressing the same challenge coming from the large variability of human appearance due to articulation.

In this work, we proposed to replicate Yang and Ramanan's paper [7] on human detection and pose estimation using RGB images. We have seen one state-of-the-art object detection method, namely the deformable part-based model (DPM) [3], in the lecture. Yang and Ramanan's work differs from the DPM model in several important ways: 1) while the DPM paper addresses generic object detection, Yang and Ramanan focus specifically on human detection, 2) DPM treats the parts as latent variables and the model is learnt only given the annotations of object full extents, while in Yang and Ramanan's paper, the annotations of body parts are given as a stronger supervision during training, and 3) DPM models significant appearnce variance (e.g. viewpoint changes) by using a set of global mixture models, while Yang and Ramanan use local mixture models (i.e. multiple templates for each body part), and allow a larger variabilitiy in capturing appearance changes. Given these advantages and the state-of-the-art performance reported in the paper, we conclude that this paper has significant contributions and is worth a replication study. An illustration of Yang and Ranaman's model and samples of human detection results are shown in Fig. 2 and Fig 1, respectively.

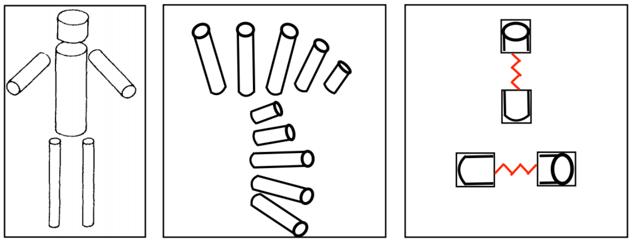

Figure 1: Illustration of the pictorial structure and mixture of local parts in Yang and Ramanan's paper [7].

2. Replication Overview

Our plan for this replication study can be divided into three main blocks. In short, we aim to first replicate the experimental results by following the techniques presented in the paper, and then study the proposed method further by experimenting with different settings, as well as understand how well it can be generalized to more realistic images.

- Re-implement the inference algorithm, in paricular,

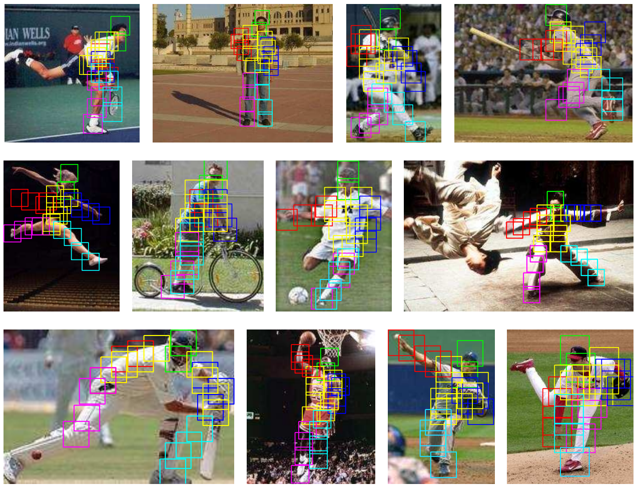

Figure 2: Illustration of Yang and Ramanan's work on human detection and pose estimation [7]. Different color of boxes identify different local parts of human body.

the dymamic programming algorithm. This is a good starting point, since the authors have provided the code for learning and inference, and a fixed training and test split on the benchmark dataset. This will be helpful for us to verify the implementaion, since we can fix the trained model and focus only on the inference evaluation of the test set.

- Re-implement the learning algorithm, in particular, solving the structural SVM problem using the dual coordinate descent solver. The paper uses a fixed training set provided by the Image Parse dataset [4]. We plan to re-train the model by changing the size of the training set. The goal is to better understand how sensitive the trained model is to the size of the training set, and whether the current training set contains enough training samples. The training code is also provided by the author, so it can be used to verify our implementation.
- Build our own human dataset. We aim to build a small but more challenging dataset with more realistic images, and see how far can the target method go. This will also tell us 1) whether there is a bias in the dataset the paper evaluates on, and 2) how much gain could we get by adding new training image into the current training set.

We want to note that the authors have released their implementation for research purposes [1]. This serves as a great resource for our replication study by the following reasons: 1) The information provided in the paper is insufficient from the full replcation aspect. For example, the paper describes their dual coordinate descent algorithm for learning only briefly and point to another reference. The reference does detail the algorithm better, but it could not provide some critical information in the context of the first paper such as initialization schemes, hyperparameters, and convergence criterions. Having the provided code will help us pick up the missing information. 2) The released code will provide a good amount of help (but not complete) to our debugging process. Instead of checking the whole framework due to incoherent outputs, we could compare the intermediate results to better understand what the mistakes are.

3. Inference

We first briefly review the overall pictorial structure model for human detection proposed in [7] and the inference problem in Sec. 3.1. Next we evaluate our implementation of the inference algorithm on the the Image Parse dataset [4] in Sec. 3.2.

3.1. Algorithm Overview

Yang and Ramanan 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 correspends to finding the maximum scoring locaions (l, t) given the im-

[1] http://www.ics.uci.edu/ yyang8/research/pose/

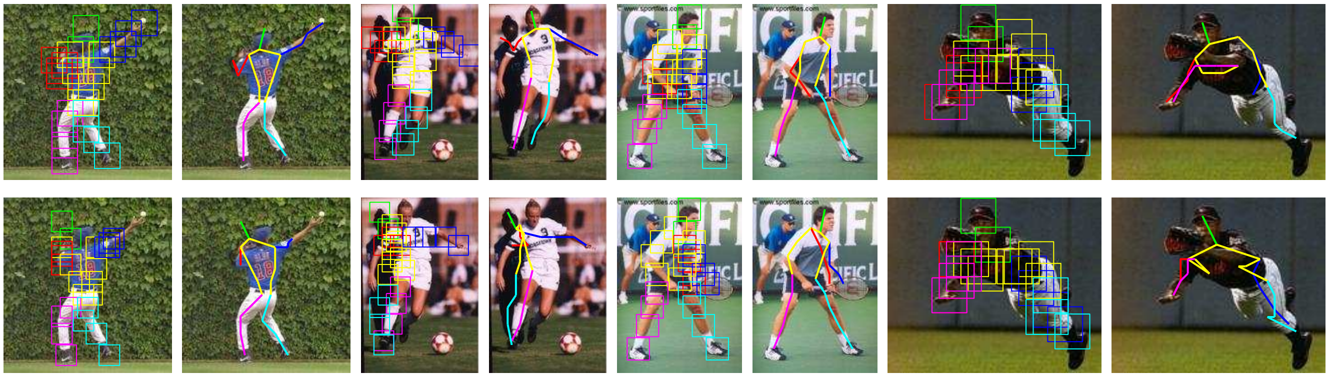

Figure 3: Qualitative samples of the inference replication on the Parse dataset. The pre-trained model is composed of 26 human parts. Each human detection is visualized by the detected bounding boxes of parts (left) and the skeletons (right). The top row show the result obtain from the author's implementation. The bottom row shows our implementation. Our implementation performs comparable to the authors' globally while there are local errors occasionally.

age 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. Replication Result

We verify our implementation by first comparing the result on the same benchmark dataset, the 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 fix the trained model and use only the test images for evaluation. As [4] suggests, we evaluate the result based on the accuracy of body joint localization. We follow the paper and use two different metrics to measure such accuracy: 1) probability of a correct keypoint (PCK), and 2) average precision of keypoints (APK). Note that for PCK, the detector is ran within the annotated bounding box of the person. This metric highlights how well the body joints are localized given the ground-truth human bounding boxes.

We start by comparing our implementation and the implementation provided by the authors [2]. For the quantitative evaluation, we observe a significant drop (over 10%) in our implementation compared to the authors' implementation. We also show some qualitative examples in 3, where the top rows are the results of the authors' implementation, while the bottom rows are ours. There are occationally errors in only a local part of the detection, for instance, the right arm on the leftmost image, and very few cases where there are global errors, for instance, the second left image. Since in general our implementation works in a reasonable level from a qualitative perspective, it suggests that the cause of the gap between our result and the original paper is less likely to be on the algorithmic side. Unfortunately, the question remains at the due of this report. We suggest a few possible explains on the result difference: 1) There could be a bug caused by a typo in our implementation, and it has a local affect which does not reflect significantly in the

	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[7]	88.3	81.7	62.2	42.0	71.0	68.5	61.2	67.8
Ours	78.0	70.0	40.5	28.3	58.8	56.3	52.4	54.9

(a) Probability of correct keypoints (PCK)

	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[7]	83.9	77.6	45.7	24.2	60.4	53.8	46.5	56.0
Ours	69.3	60.7	24.5	10.4	42.3	36.6	33.6	39.6

(b) Average precision of keypoints (APK)

Table 1: We compare the inference result between the paper's implementation [7] and our implementation (Ours). The trained model is fixed to be the same. There is a accuracy drop in our implementation compared to the authors'.

[2] Note that there is a slight difference between the result reported in the paper and the result we obtain from their code. We decide to use the output of the code for all the comparision in this study

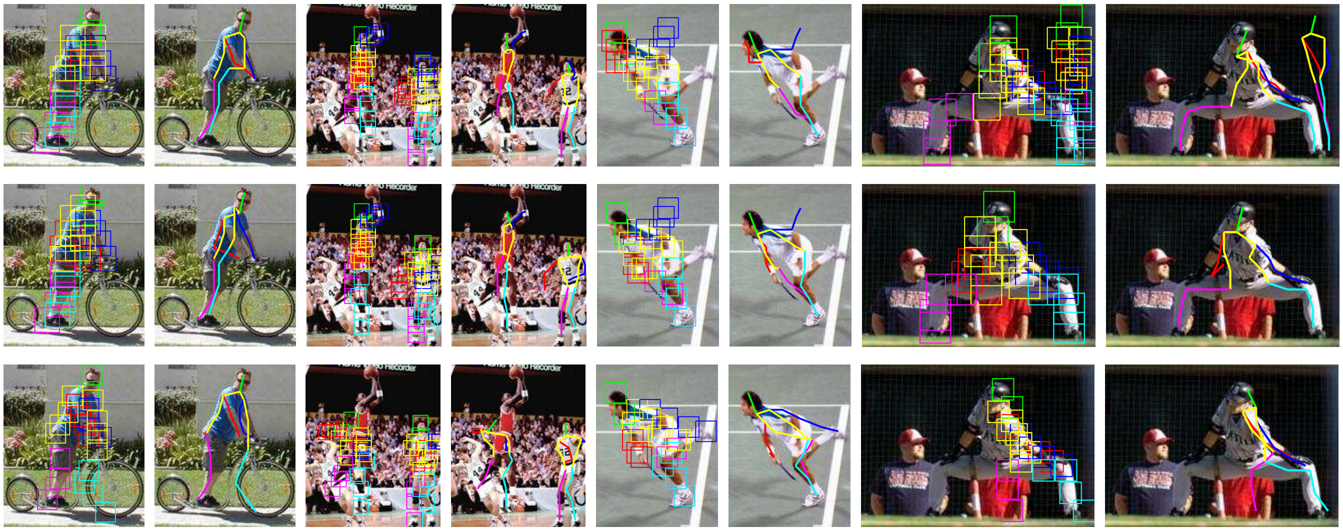

Figure 4: Qualitative examples of the learning replication on the Parse dataset. The first row is using the authors' implementation for both learning and inference. The second row is using our implementation for learning and the authors' implementation for inference. The third row is using our implementation for both learning and inference. We obtain comparable results on learning when using the same inference implementation.

global detection result. 2) There could possibly be a numerical issue on the computation, which suggests that the implementation is not numerically stable. From the debugging experience, we learned that the debugging can still be difficult and time-consuming even the original code is available.

4. Learning

The goal is to learn the model parameters in 3 that maximizes the discrimination power. The paper formulates a structural SVM learning problem and solves it using an efficient algorithm based on dual coordinate descent (DCD). The primary focus of our replication study is to re-implement the dual coordinate descent solver. We first review the learning problem and the DCD solver in Sec 4.1. The replication result is presented in Sec 4.2.

4.1. Algorithm Overview

The paper adopts a supervised learning paradign. Given labeled positive examples $\{I_n, l_n, t_n\}$ (images with labeled human body parts) and negative examples $\{I_n\}$, the goal is to learn a discriminative model that will score high on test regions when a person is presented and score low otherwise. Since the score function 3 is linear in model parameters $\beta = (\omega, b)$, we can write $S(I, z) = \beta \cdot \Phi(I, z)$. The model parameters is then learnt by solving the following optimization problem:

$$\begin{aligned} \arg\min_{\beta,\xi_n\geq 0} \quad & \frac{1}{2}||\beta||^2 + C\sum_n \xi_n \\ \text{s.t. } \forall n \in \text{pos} \quad & \beta^T\Phi(I_n, z_n) \geq 1 - \xi_n \\ \forall n \in \text{neg}, \forall z \quad & \beta^T\Phi(I_n, z) \leq -1 + \xi_n, \end{aligned} \tag{4}$$

where C is the hyperparameter and ξ_n is the slack variable corresponded to sample n.

The form of the learning problem above is known as a structural SVM (SSVM), an extension of the binary SVM to handle the prediction of structured outputs. There has been many powerful techniques proposed to solve the SSVM problem, for instance, the cutting plane algorithm and the stochastic gradient descent (SGD). This paper uses a *dual coordinate descent* solver, which is detailed in [5] by one of the co-authors. As [5] mentioned, in large scale learning problems, batch algorithms are often impractical due to the difficulty of fitting the entire dataset into the memory, while the online algorithms alleviate such problem but often converge slowly. The proposed dual coordinate descent solver serves as an intermediate solution: it keeps a relatively small set of active constraints (support vectors) in the memory, which leads to a better performance on the convergence, while keeping it as stable as the batch algorithms.

The proposed dual coordinate descent solver has many practical advantages in large scale learning problems, and it could potentially be benefitial to our later research studies. Thus we decide to focus primarily on implementing the

	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[7] / [7]	88.3	81.7	62.2	42.0	71.0	68.5	61.2	67.8
O / [7]	87.3	79.8	58.8	43.7	70.0	63.9	58.3	66.0
O / O	79.0	71.7	40.7	30.0	57.8	58.3	54.1	56.0

Probability of correct keypoints (PCK).

	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[7] / [7]	83.9	77.6	45.7	24.2	60.4	53.8	46.5	56.0
O / [7]	83.2	75.7	44.9	26.5	57.0	49.2	41.9	54.1
O / O	72.5	63.6	25.1	13.8	42.7	43.0	37.1	42.5

Average precision of keypoints (APK).

Table 2: We compare the result of learning between the paper's implementation and our implementation. A / B represents using A for learning and B for inference ([7] for the paper and O for Ours). We achive comparable results on learning when using the same inference implementation. Our inference implementation causes a performance drop as in Tab. 1.

proposed DCD algorithm for the training part of this replication study.

Now we highlight the idea of the proposed DCD solver (refer to [5] for the complete details). The optimization problem 4 is a quadratic programming (QP), and we can derive the associated Lagrangian:

$$L(\beta,\xi,\alpha,\mu) = \frac{1}{2}||\beta||^2 + C\sum_i \xi_i - \sum_{ij}\alpha_{ij}(\beta^T\Phi - 1 + \xi_i) - \sum_i \mu_i\xi_i, \tag{5}$$

where α and μ are the Lagrange multipliers. We can take the derivatives of the Lagragian with respect to the primal variables, and get the dual of the QP in 4:

$$\begin{aligned} \max_{\alpha\geq 0,\mu\geq 0} \quad & F(\alpha) = -\frac{1}{2}\sum_{ij,kl}\alpha_{ij}x_{ij}^T x_{kl}\alpha_{kl} + \sum_{ij}\alpha_{ij} \\ \text{s.t. } \forall i, \quad & \sum_j \alpha_{ij} \leq 1 \end{aligned} \tag{6}$$

By strong dualality, the solution of problem 4 (primal) and problem 6 (dual) are equivalent. One important property about 6 is that when we minimize the objective with respect to a single α_{ij} while keeping other fixed, the objective function is quadratic over α_i, and there exists an analytical solution. The efficiency of the proposed DCD algorithm is thus based on this fact. We randomly pick a single α_{ij} and optimize 6, and this will garuantee to increase the object function. It is also shown in [5] that the lower-bound and an approximated upper-bound can be tracked efficiently during the run of the optimization. We can determine the convergence of the learning when the gap between the upper-bound and the lower-bound is below a fixed threshold.

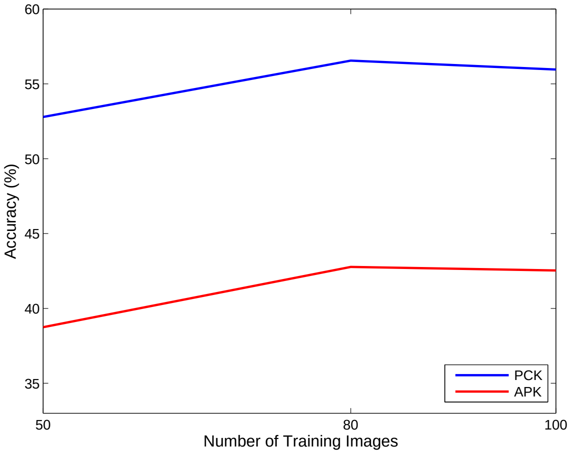

Figure 5: Evaluation on Parse test set using different number of training images.

4.2. Replication Result

We use the standard training and test split for the Image Parse dataset. We compare the result obtained by 1) using the authors' implementation for both learning and inference, 2) using our implementation for learning and the authors' implementation for inference, and 3) use our implementation for both learning and training. We report the body joint localization accuracies and qualitative examples in Tab. 1 and Fig. 3, respectively. As shown in Tab. 1, we obtain comparable result with the authors' implementation when the inference choice is fixed. This indicates a success in the replication on the learning side, since our implementation performs comparable to the authors' implementation. However, we still see an accuracy drop when applying our inference implementation. This is expected from the result in Sec 3.2. We note that the success of the replication is contributed significantly from the availability of the paper's code. The code provides large amount of additional information to the paper, especially on the learning side, such as the initialization scheme, loading control of data sequence, and hyperparameters.

We also investigate the affect of the size of the training set on the Parse dataet. We change the number of number of training images by randomly picking a subset of the original training set, and run the test experiment on the test set. The results on PCK and APK are shown in Fig. 5. We observe that the accuracy converges around using 80 training images. The reason could be that the Parse dataset is relatively small and the statistics of the training and test set

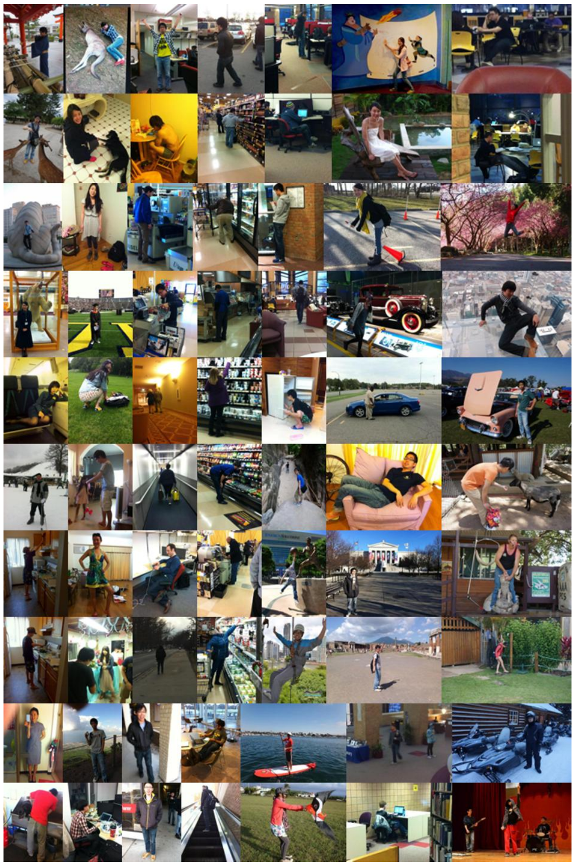

Figure 6: Our proposed new dataset. It contains 70 images taken from our daily life.

are similar. Therefore, using a small amount of training images can reach the upper bound of the test accuracy. As we will shown in Sec 5, the statistics of the Parse dataset is very different from the general images we take everyday. To increase the generalization power of the trained model, we need to add the training images from different sources to balance the statistics of the dataset, in order to increase the generalization power.

5. New Dataset

To better understand the result on the parse dataset, as well as to gain more insights on the generalization of the trained model, we propose a new dataset for further evaluations. As shown in Fig. 6, this dataset contains 70 images which mostly came from the photos we took in our past daily life. Compared to the images in the Parse dataset, which are carefully selected for evaluation purposes, our new dataset are more common and realistic. We annotate the human body joints for the new dataset in the same way as the parse dataset.

	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[7] / [7]	67.9	47.9	29.3	20.7	33.6	31.4	30.0	37.2
[7] / O	60.0	39.3	17.9	19.3	25.0	25.7	30.7	31.1
O / [7]	67.9	51.4	22.9	18.6	34.3	30.7	31.4	36.7
O / O	57.9	38.6	20.7	15.0	24.3	27.1	32.1	30.8

Probability of correct keypoints (PCK).

	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Avg
[7] / [7]	54.8	35.0	19.4	10.5	22.0	15.7	19.1	25.2
[7] / O	46.2	29.0	9.6	6.5	19.8	15.0	17.3	20.5
O / [7]	49.6	34.5	15.7	7.6	22.8	13.9	18.0	23.2
O / O	47.9	32.3	8.0	4.5	15.1	14.6	19.6	20.3

Average precision of keypoints (APK).

Table 3: The inference result on the new dataset using the model trainined on the Parse dataset. The result of different implemention choices are reported. A / B represents using A for learning and B for inference ([7] for the paper and O for Ours). The performance drops significantly compared to the Parse test set.

To see how well the model trained on the Parse dataset generalizes to images from a different source, we take our trained model from Sec 4.2 and run inference on the new dataset. The accuracies of body part localization (PCK and APK) are shown in Table 3. In addition to the same drop of performance for our inference implementation, the overall accuracies drops significantly compare to the evaluation on the Parse dataset (Table 1 and 2). This indicates there is a bias between the Parse dataset and our new dataset. We point out four major issues that explains the accuracy drop on the new dataset as follows. 1) **Poses:** Most images in the Parse dataset are people playing sports, e.g. basketball, baseball, tennis, etc. This causes a bias in the pose distribution. For example, the model learnt from the Parse dataset can not identify the squating pose shown in the lower left image in Fig 7 because of lacking similar training examples. 2) **Focus of Image:** People are often the dominant objects of the images in the Parse dataset. This is not necessarily true for general images, as seen in Fig 7. The problem becomes significant when the scale of the person is too small such that the extracted feature can not discriminate difference poses. 3) **Occlusion:** Images in Parse have been selected to limit the amount of occlusions and self-occlusions. The limbs can be observed most of the time possibly due to the performed activity (sports). However, in a more realisitic scenario, human in an image suffers more from occlusions and self-occlusions. In many cases, half of the body may not be visible at all. Unfortunately the proposed method are not capable of handling these cases yet. 4)

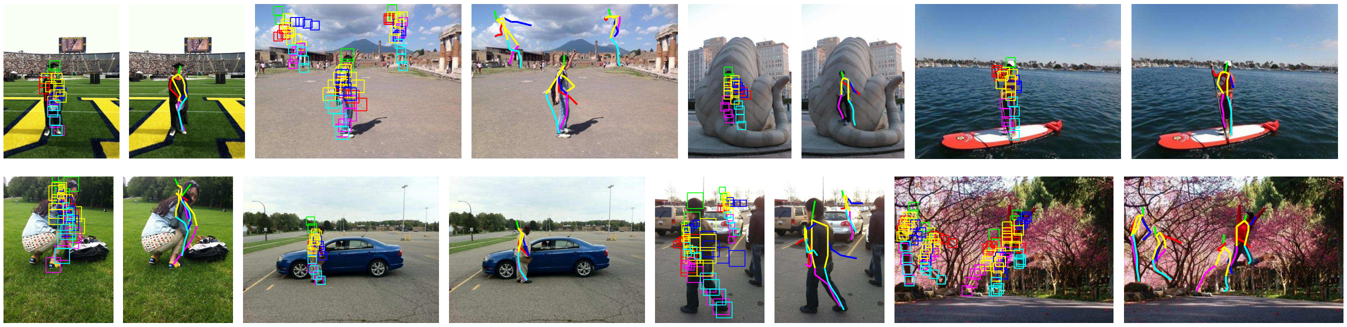

Figure 7: Qualitative examples on the new dataset using the model trainined on the Parse dataset.

Noisy Background: The background of the human can be very cluttered in realistic images. The problem is even significant when the people are less dominant in the image. As shown in the lower right image of Fig 7, noisy background can cause false positives due to the noisy image features. From the analysis above, we conclude that there is still a gap between the Parse dataset and realistic images, and the later is still very challenging for several reasons that may cause the proposed method to fail.

6. Conclusion

In this replication study, we have re-implemented the inference algorithm and the learning algorithm. The replication of learning is successful, while the replication of inference still requires verfication. We show some analysis on how training on the Parse dataset is affected by the size of the training set. We also perform extensive studies on the generalizaion of the trained model by testing it on a new proposed dataset. We conclude that due to the large variability of human body poses, as well as different sources of noises in the real-world images, a large amount of training images is required to successfully increase the generalization power of the proposed method.

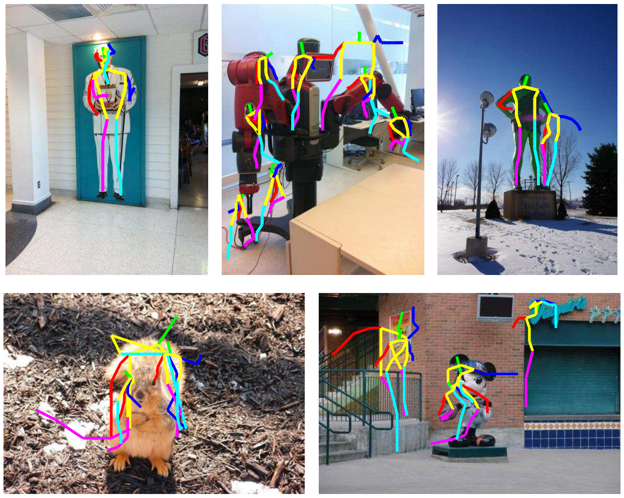

Figure 8: Anecdotal examples on non-human (or fake human) images.

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 *Proceedings 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. 2, 3

[5] D. Ramanan. Dual coordinate solvers for large-scale structural svms. Technical report, Univ. of California, Irvine, 2012. 4, 5

[6] 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

[7] 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, 5, 6