SN Computer Science (2020) 1:208
https://doi.org/10.1007/s42979-020-00217-9

ORIGINAL RESEARCH

Training Physical and Geometrical Mid-Points for Multi-person Pose Estimation and Human Detection Under Congestion and Low Resolution

Yadong Pan[1] · Ryo Kawai[1] · Noboru Yoshida[1] · Hiroo Ikeda[1] · Shoji Nishimura[1]

Received: 1 April 2020 / Accepted: 5 June 2020 / Published online: 21 June 2020

Abstract
This paper introduces the design and evaluation of NeoPose which is developed for multi-person pose estimation and human detection. The design of NeoPose is targeting the issue of human detection under congested situation and with low resolution in the image. Under such situations, we compared the performance of different versions of NeoPose as well as other existing algorithms in a human detection task. Throughout the task, the usefulness of two kinds of mid-point (physical and geometrical mid-points) and a deconvolution structure was discussed. Experiment results indicated that NeoPose which applied geometrical mid-points and deconvolution structure performed the best in terms of both precision and recall in the evaluation.

Keywords Mid-points · Multi-person pose estimation · Human detection · Low resolution

Introduction

Human detection and pose estimation are two joint issues in recent artificial intelligence researches. They can be used for recognition of human action [1–3], tracking [4, 5] and re-identification [6] of human in online surveillance and human-object interaction [7]. Most of the latest algorithms on pose estimation tend to embed a human detector at the beginning of its data processing unit, such as [9–11] which ranked top three in COCO key-point challenge 2019. Those human detection-based algorithms are called top-down methods. However, as to solving real industrial problems, such as the outdoor surveillance where people are often in a congestion (Fig. 1a) or monitoring suspicious people near two countries' border line where people captured are with low resolution (Fig. 1b), human detector tends to fail. Such problem has been pointed out by Gkioxari et al. in their research [8].

To solve real industrial problems, compared to applying top-down method for recognizing human and human behavior, we consider than human key-point-based method which is called bottom-up method would be a better solution. A bottom-up method first recognizes human key-points (also called body region points) on visible area of human body in the whole image, then associates those visible key-points into individual persons and generates human bounding boxes. As a result, human bounding boxes may not enclose a person's full body but the detection itself is reasonable. Figure 2 shows an example of comparison between top-down- and bottom-up-based human detection under congestion and low image resolution. From the example, we could see how bottom-up method benefits in such outdoor scene. Therefore, the scope of this paper is to develop a better bottom-up approach in order to solve real industrial problems.

In this paper, we proposed a bottom-up pose estimation system called NeoPose. NeoPose detects different types of body region points in the image and associates them into different individuals. Each group of associated body region points is a called a pose vector, which represents the pose of an individual person in the image. Human detection can be realized by calculating several bounding boxes that enclose each pose vector. Compared to some previous researches such as OpenPose [12], Art-Track [14] and Associative Embedding [13], NeoPose performed better in a human

This article is part of the topical collection "Machine Learning in Pattern Analysis" guest edited by Reinhard Klette, Brendan McCane, Gabriella Sanniti di Baja, Palaiahnakote Shivakumara and Liang Wang.

✉ Yadong Pan
panyadong@nec.com

[1] NEC Biometrics Research Laboratories, Kawasaki, Japan

Fig. 1 Industrial problems that require human sensing: **a** surveillance of a pedestrian crossing (MOT dataset [18], **b** a wide-range surveillance near two countries' border (Getty image)

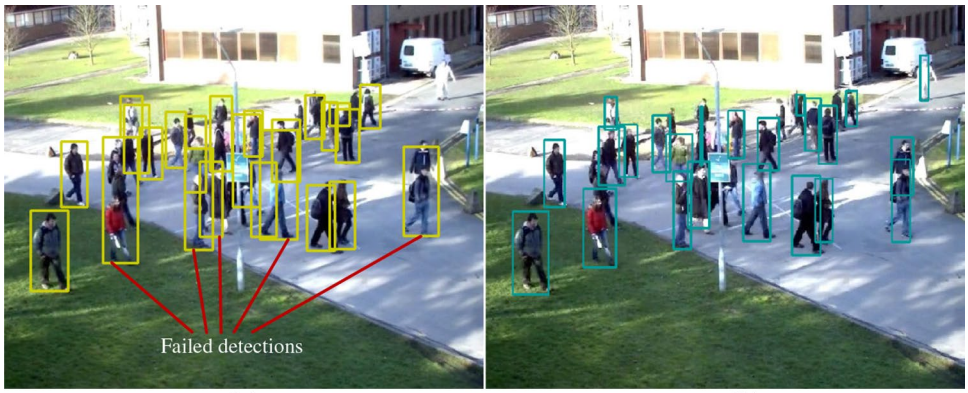

Fig. 2 Comparison between **a** Xiao et al.'s top-down method [21] and **b** our bottom-up method [15] on human detection under real-world surveillance

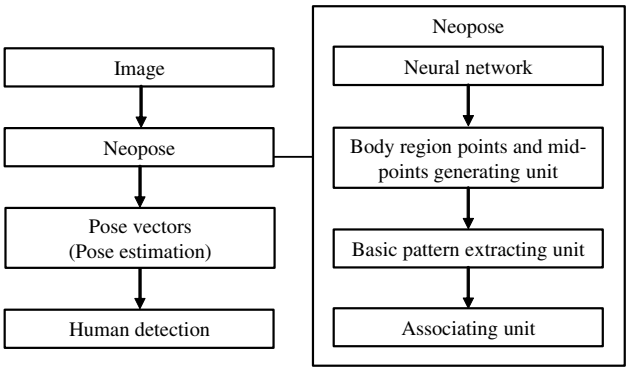

Fig. 3 Data flowchart and architecture of NeoPose

detection task under low image resolution. The task will be explained in a following section of this paper.

The data flow in NeoPose (Fig. 3) follows our previous work [15], where a structure called basic pattern is generated for each person in the image after different types of body region points are detected through a deep neural network. A basic pattern is a set of body region points including a person's shoulders, ears and neck. After generating basic patterns, each body region point with other types is associated with one of the basic pattern or ignored as false detection. Mid-points (middle of two body region points) which are also detected through the deep neural network are used as reference to give a judgment in associating each body region point to a specific base pattern.

To extend our previous research, in this paper, we made some additional design on NeoPose. (i) We extended the design of mid-points from physical mid-points to geometrical mid-points. Both physical and geometrical mid-points are defined based on general body region points as shown in Fig. 4a. Physical mid-points (Fig. 4b) are those mid-points which physically locate on human body, such as mid-point between a person's right shoulder and right waist. On the other hand, a geometrical mid-points is defined as a mid-point between any two kinds of body region points, which may not locate on but around a person's body according to specific pose (Fig. 4c). According to the definition, geometrical mid-points include physical mid-points but represent more types of mid-points. (ii) We enhanced the deep neural network for training both general body region points and mid-points. In this paper, we compared the quality of human detection under different design of NeoPose and discussed the usefulness of training mid-points under low image resolution as well as its usage in solving real industrial issues.

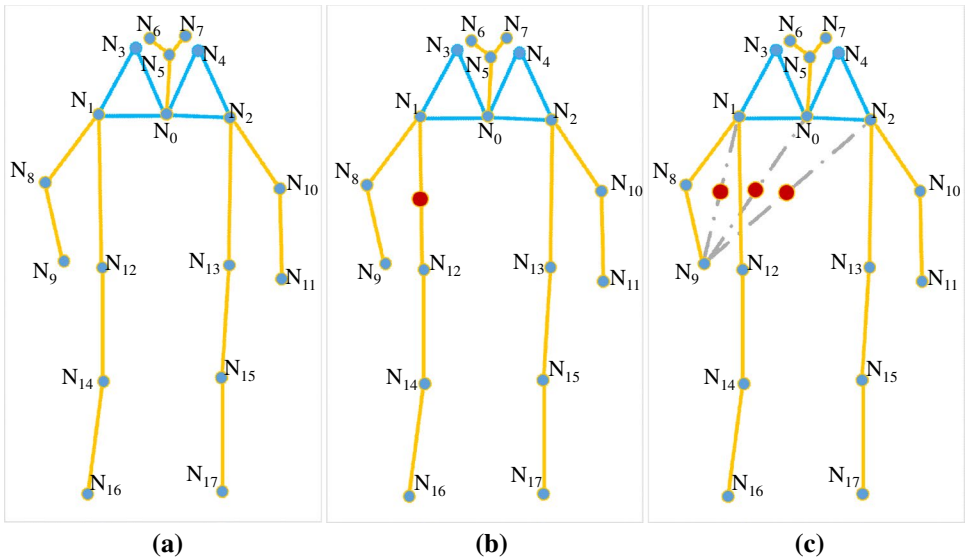

Fig. 4 **a** Definition of human body region points: N_0:neck, N_1 :right shoulder, N_2:left shoulder, N_3:right ear, N_4:left ear, N_5 :nose, N_6:right eye, N_7:left eye, N_8:right elbow, N_9:right wrist, N_{10}:left elbow, N_{11}:left wrist, N_{12}:right hip, N_{13}:left hip, N_{14} :right knee, N_{15}:left knee, N_{16} :right ankle, N_{17}:left ankle. **b** An example of physical mid-point. **c** Examples of geometrical mid-point

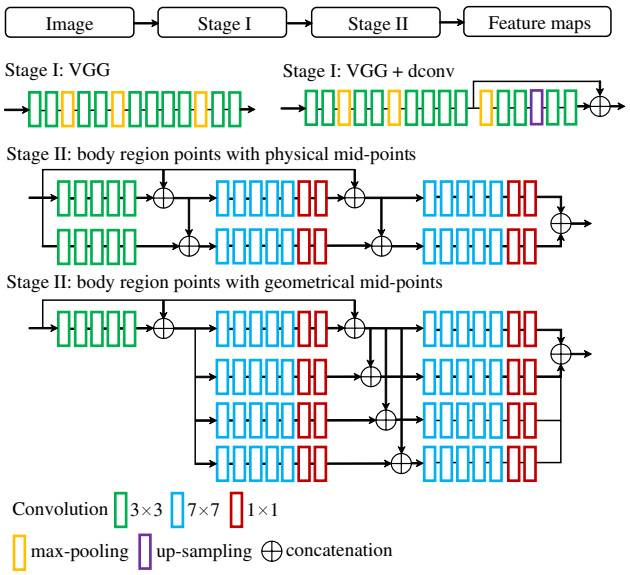

Fig. 5 Deep neural network in different versions of NeoPose

Methodology

Generating Body Region Points and Mid-points

NeoPose applies a deep neural network which is trained on COCO dataset [16]. The neural network (Fig. 5) which consists of two stages trains/infers the body region points we well as mid-points. The resolution of image as input to the network can be adjusted but should be multiple of 8. Throughout the network, a feature map is generated for each type of body region point and mid-point. The first stage is the backbone of the neural network. In our previous research, we applied a vgg for the first stage, while in this paper, we modified the vgg by adding in a deconvolution (dconv) structure including an up-sampling layers, two convolution layers and a concatenation layer. This is referring to some recent researches [20, 21] where deconvolution was applied to reduce false detections.

As to the second stage, in our previous research we trained 19 channels for 18 body region points and the background in one branch, and 10 physical mid-points in another branch. The mid-term feature from the body region points' branch was shared with the mid-points' branch. In this paper, we modified this part of network to a four-branch structure. The main branch that receives the data from the first stage (vgg/vgg + dconv) consists of 19 channels for training 18 body region points and the background. Other three branches were designed for 30 types of geometrical mid-points that are defined according to a body region point in $S = \{N_0, N_1, N_2\}$ and one in $D = \{N_i \mid 8 \leq i \leq 17\}$, where the geometrical mid-points corresponding to the same item in S were trained in the same branch. For example, the middle of N_8 and N_0 and the middle of N_9 and N_0 were trained in the same branch. In the network, the mid-term feature from the main branch was shared with the other three branches. The output of the network includes 49 feature maps. In case of training geometrical mid-points, the loss was calculated as:

$$Loss = \sum_{C} \sum_{P} W(P) \cdot \parallel S_P^T(P) - S_P^G(P) \parallel_2^2$$

where C stands for the 49 channels, and P represents all pixels in the feature map. S_P^T is the score calculated by the deep network and S_P^G is the ground truth. The ground truth for geometrical mid-points was calculated based on that of body

Fig. 6 Detections of basic patterns, body region points of left knee and geometrical mid-points corresponding to left shoulder and left knee. **a** original image from COCO, **b** rendered image

region points in COCO dataset. W is a binary weight, which gives a value 0 when ground truth is missing at the current location in an image. After training the deep network, body region points and mid-points in an image can be extracted by searching the local peaks in the feature maps.

Associating the Detected Body Region Points

The process of generating basic patterns follows the method described in our previous research [15]. As to associating the detected body region points, in this paper, we proposed a novel method. The novel method is developed based on the training of 30 types of geometrical mid-points. For each detected body region point with type $N_i (8 \leq i \leq 17)$ (also called a N_i point), the presence of its corresponding types of geometrical mid-points was checked to determine which basic pattern it should be associated to. Figure 6 is an example that shows the detection of basic patterns, body region points of left knee and the geometrical mid-points that correspond to left knee and left shoulder. From the figure, we can know that even though such geometrical mid-point may not locate on a person's body but around the body (e.g., the person on the right), it can be detected through the deep neural network.

To associate a N_i point $(8 \leq i \leq 17)$ to a basic pattern. NeoPose first builds links from the N_i point to the N_0, N_1 and N_2 points of each basic pattern. In case of missing any of the three points in a core, the link to that point would not be built. However, the N_0 point must exist according to the definition of basic pattern. Next, for each basic pattern, NeoPose counts the number of valid links where a corresponding type of geometrical mid-point is detected within an ellipse area between the two terminals of the link. Figure 7 shows an example of ellipse area. Following our previous research [15], in this research, R_{major} of the ellipse area is set to $|(N_i, N_j)| \times 0.35$, and R_{minor} is set to $R_{major} \times 0.75$. The basic pattern(s) having the most number of valid links will be considered as candidate basic pattern(s). Then, the candidate basic pattern with the minimum distance from its N_0 point to the N_i point will be

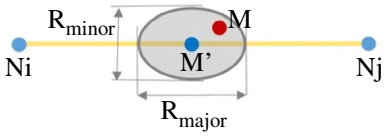

Fig. 7 Mid-point and the ellipse area. M': ground-truth of mid-point. M: detected mid-point. N_i and N_j: two types of body region point

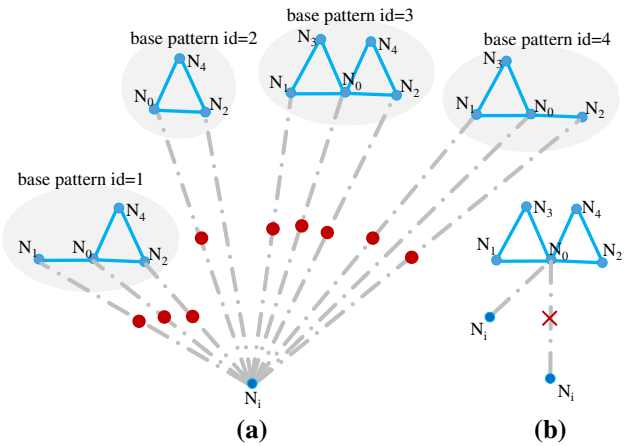

Fig. 8 Example of body region points' association. **a** Associating a body region point to a basic pattern. **b** For a basic pattern, selecting one from multiple body region points with the same type

accepted as the basic pattern for the N_i point to be associated to. Taking Fig. 8a as an example, for the N_i point, basic pattern with id 1 and 3 have the most number of valid links. In such case, the N_i point should be associated to basic pattern with id 1 since the distance from the N_i point to N_0 point in that basic pattern (id 1) is shorter than that of the other one (id 3). For other types of N_i points $(5 \leq i \leq 7)$, the basic pattern was selected by having the minimum distance from its N_0 point to the N_i point. Overall, the process of associating all different types of body region points to basic patterns can be done in parallel, which helps to improve NeoPose's processing speed.

After the process that associates those N_i points to specific basic patterns, for each basic pattern, NeoPose drops the number of N_i points that are associated to it. If a basic pattern is associated with multiple N_i points, the Ni point with the minimum distance to the N_2 point in the basic pattern will be accepted and others will be excluded (Fig. 8b). As a result, each basic pattern can associate to no more than one body region point with any specific type. Figure 9 shows some examples of images rendered with pose analyzed by NeoPose with deconvolution structure and applying geometrical mid-points.

Evaluation

In this section, we tested the performance of different design of NeoPose:

vgg + physical mid-points
vgg + deconvolution + physical mid-points
vgg + geometrical mid-points
vgg + deconvolution + geometrical mid-points

The test we conducted is a human detection test on images from MHP dataset [17]. The dataset contains around 4000 images with multiple people captured in each image and a variety of different poses in outdoor scenes, which meets the needs to investigate NeoPose for industrial usage (e.g., outdoor surveillance). To simulate the situation in many industrial problems, we resized all images in the dataset to a fixed and smaller height (120 pixels) and a smaller width according to the image's aspect ratio and used the resized images as input to NeoPose's deep neural network. After executing pose estimation on all images, we extracted those pose vectors which associates at least 10 body region points. According to our experience in solving industrial problems, 10 associated body region points can be considered as a practical level to suggest that the human detection is successfully done.

For each pose vector with at least 10 body region points associated, a minimum bonding box that encloses all body region points in the pose vector was calculated (Fig. 10). Sub-images aligned with those bounding boxes and rendered with a pose vector's all body region points were segmented from the original images. The segmented images were checked frame-by-frame by experts on image sensing and categorized into three classes: (i) Correct detection, which means all body region points in this pose vector are located on the same person's body without a fatal error. The criterion for a fatal error is that it does not mislead the understanding of a person's pose, which is judged with the experts' experience. (ii) False detection, which means in this pose vector, the body region points are located on different persons, or some are located on the background rather than

Fig. 9 Pose estimation on images from MHP dataset [17] using NeoPose that applies deconvolution structure and geometrical mid-points

Fig. 10 Three categories of pose vector in the human detection task. **a** Correct detection, **b** false detection, **c** ghost detection

human body. (iii) Ghost detection, where all body region points are located on background rather than on human body. The evaluation explained above were conducted for different kinds of bottom-up approach including OpenPose [12], ArtTrack [14] Associative Embedding(AE) [13] and the four versions of NeoPose. The evaluation results are shown in Table 1.

The results suggested that: (i) NeoPose with deconvolution and geometrical mid-points performed the best in terms of both precision (88.0%) and recall (80.7%). (ii) Whatever applying physical or geometrical mid-points, the deconvolution structure helped to reduce the ghost detections. Figure 11 shows some more details on how the deconvolution structure benefits in reducing the false positive detections of body region points. (iii) Compared to applying physical mid-points, networks with geometrical mid-points gained more correct detections, and therefore raised the level of recall.

Discussion

Training of Physical and Geometrical Mid-points

In OpenPose [12], the association of body regions is realized through part-affinity-field(PAF), which is surface integral of those pixels between two body regions. However, in images with low resolution, the reliability of PAF drops and leads to the errors on pose estimation. The design of mid-points (both physical and geometrical ones) is targeting such issue by simplifying the representation of two body regions' correlation. Compared to OpenPose, NeoPose that applied either physical or geometrical mid-points showed better performance in the human detection task.

As to the association of body region points, networks with geometrical mid-points gained a larger number of correct association. One of the important reasons is that the

Table 1 Results of human detection on MHP dataset using different algorithms

	GT	Correct	False	Ghost	Precision (%)	Recall (%)
OpenPose	12319	8762	1499	5	85.3	71.1
Art-Track	12319	6878	1190	2	85.2	55.8
AE	12319	9372	1738	125	83.4	76.0
NeoPose (vgg + physical mid-points)	12319	9284	1516	9	85.8	75.3
NeoPose (vgg + deconvolution + physical mid-points)	12319	9356	1465	0	86.4	75.9
NeoPose (vgg + geometrical mid-points)	12319	9890	1417	7	87.4	80.2
NeoPose (vgg + deconvolution + geometrical mid-points)	12319	9943	1352	1	88.0	80.7

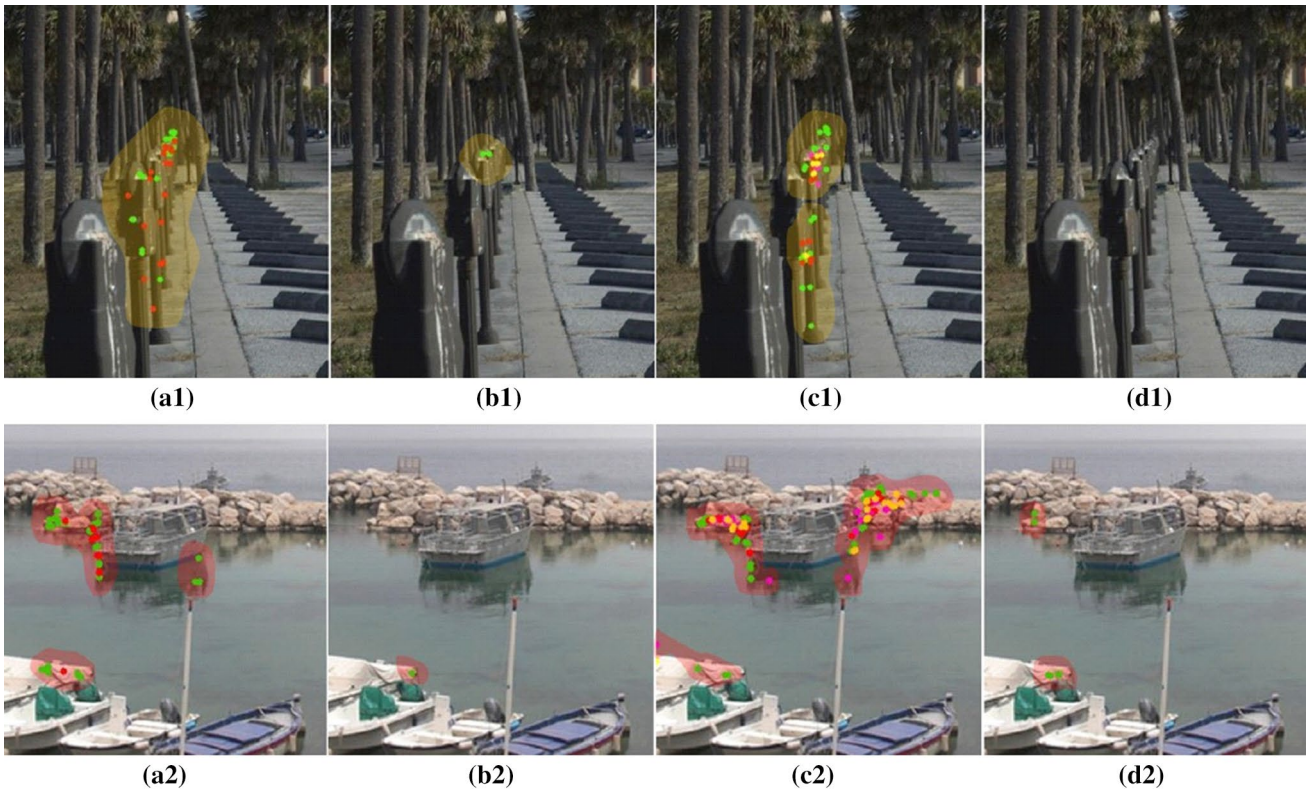

Fig. 11 False positive detections of human body region points using different versions of NeoPose. **a1**, **a2** vgg + physical mid-points, **b1**, **b2** vgg + deconvolution + physical mid-points, **c1**, **c2** vgg + geometrical mid-points, **d1**, **d2** vgg + deconvolution + geometrical mid-points

body regions' association based on geometrical mid-points was done in parallel. In such case, any type of body region point can be associated to a basic pattern without relying on other types of body region points. While in case of applying physical mid-points, the association is done is sequence (e.g., right waist, right knee, right ankle), where missing the association of any body region will cause that the following ones could not be successfully associated to the basic pattern.

What's more, geometrical mid-points can be located out of a person's body, which means what we trained is not only the visible information but also the geometrical correlation among different body region points. Figure 6 shows that geometrical mid-points can be successfully detected. Such way of training the correlation is recently studied in some researches such as training the center point of a person [19]. We also considered that such theory of training could be used in future researches on human-object/human-human interaction.

Deconvolution Under Low Image Resolution

Deconvolution is recently considered as a practical way to enhance the robustness of sensing algorithms [20, 21]. In this research, we evaluated its usage under low resolution images which is not studied in previous researches, and found that the false positive detections of body region points significantly decreased by using the networks with deconvolution structure.

Associating Body Regions Rather Than Conducting Top-Down Human Detection

The task for evaluating different algorithms in this paper targets on the real industrial problems where people captured are congested and with low resolution. One research by Gkioxari et al. [8] and our previous work [15] indicated that under such situations, top-down human detectors tend to fail in both human detection and pose estimation. Instead of top-down methods, in this paper, we investigated how bottom-up methods could be used in solving human detection problems. In the task on MHP dataset where all images were resized to a smaller resolution (height = 120 pixels), those networks with deconvolution structure gained almost no ghost detections, which inferred that such designs of networks could be suitable for real industrial problems such as the tasks shown in Fig. 1. In those tasks, what the most important is to confirm that the target recognized is indeed a person. Throughout associating multiple body regions, the human detection would be

more reliable in low resolution images, and even though some parts of human body are under occlusion, it is possible to determine the presence of a person by associating the visible body regions. Such claim is also considered to be extended to solve the recognition of other objects rather than human.

Conclusion and Future Work

This research is an extended version of our previous work [15]. In order to find solutions for real industrial problems, we have been focusing on following ways of thinking: (i) Using pose estimation for human detection in order to deal with occlusion in congested situation and to improve the confidence of detected person under low image resolution. (ii) For those images with low resolution, simplifying the deep neural network from training area (e.g., PAF in OpenPose) to training some mid-points for the association of body regions. (iii) What's more, in this research, we discussed how physical and geometrical mid-points performed in a task on MHP dataset and evaluated the usage of a deconvolution structure.

Although our experiment suggested that NeoPose performed better compared to other recent bottom-up approaches, both precision and recall does not reach 90 percent. This is a challenging issue because in the real industrial problems, the environment, the quality of image and the status of captured persons are usually in a complex representation. Also, such complexity is the reason why many good algorithms in academy studies are still not appropriate for releasing industrial products. We will be continuing looking for better theory and better algorithms to fulfill the gap between academic and industrial usages of AI algorithms as our constant value in future work.

We are also interested in seeing our algorithms being applied into many industrial scenes, such as autonomous driving buses, automatic sports behavior analysis, gesture-based human-computer interface for touchless systems in post-covid-19 world, which are all considered as our future publications.

Compliance with Ethical Standards

Conflict of interest Yadong Pan is working at NEC. Before he started his career at NEC, he was supervised by Prof. Kenji Suzuki at University of Tsukuba, Japan. Ryo Kawai is working at NEC. Before he started his career at NEC, he was supervised by Prof. Yasushi Yagi at Osaka University, Japan. Noboru Yoshida, Hiroo Ikeda and Shoji Nishimura are also working at NEC. Before they joined NEC, their majors were not directly related to image processing or pattern recognition.

References

1. Choutas V, Weinzaepfel P, Revaud J, Schmid C. Potion: pose motion representation for action recognition. In: IEEE conference on computer vision and pattern recognition, 2018;7024–7033
2. Demisse G G, Papadopoulos K, Aouada D, Ottersten B. Pose encoding for robust skeleton-based action recognition. In: IEEE conference on computer vision and pattern recognition workshops, 2018;188–194
3. Zolfaghari M, Oliveira GL, Sedaghat N, Brox T. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In: IEEE international conference on computer vision, 2017;2904–2913
4. Raaj Y et al. Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields. In: IEEE conference on computer vision and pattern recognition, 2019;4620–4628
5. Xiu Y, Li J, Wang H, Fang Y, Lu C. Pose flow: efficient online pose tracking. In: British machine vision conference, 2018;
6. Su C et al. Pose-driven deep convolutional model for person re-identification. In: IEEE international conference on computer vision, 2017;3960–3969
7. Gkioxari G et al. Detecting and recognizing human-object interactions. In: IEEE conference on computer vision and pattern recognition, 2018; 8359–8367
8. Gkioxari G, Hariharan B, Girshick R, Malik J. Using k-poselets for detecting people and localizing their keypoints. In: IEEE conference on computer vision and pattern recognition, 2014; 3582–3589
9. Li W et al. Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148. 2019
10. Zhang F, Zhu X, Dai H, Ye M, Zhu C. Distribution-Aware coordinate representation for human pose estimation. arXiv preprint arXiv:1910.06278. 2019
11. Su K, Yu D, Xu Z, Geng X, Wang C. Multi-person pose estimation with enhanced channel-wise and spatial information. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2019;5674–5682
12. Cao Z et al. Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE conference on computer vision and pattern recognition, 2017;7291–7299
13. Newell A, Huang Z, Deng J. Associative embedding: end-to-end learning for joint detection and grouping. Adv Neural Inf Process Syst. 2017;2277–2287
14. Insafutdinov E et al. Arttrack: articulated multi-person tracking in the wild. In: IEEE conference on computer vision and pattern recognition, 2017;6457–6465
15. Pan Y, Nishimura S. Multi-person pose estimation with mid-points for human detection under real-world surveillance. In: Asian conference on pattern recognition, 2019;239–253
16. The COCO dataset: cocodataset.org/. Accessed 1 June 2019
17. The MHP (Multi-Human Parsing) dataset: https://lv-mhp.github.io/. Accessed 1 July 2019
18. The MOT (Multiple Object Tracking) dataset: https://motchallenge.net/. Accessed 1 July 2019
19. Zhou X, et al. Objects as points. arXiv preprint arXiv:1904.07850. 2019
20. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2019;5693–5703
21. Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking. In: European conference on computer vision, 2018