Multi-person Pose Estimation with Mid-Points for Human Detection under Real-World Surveillance

Yadong Pan(✉) and Shoji Nishimura(✉)

NEC Biometrics Research Laboratories, Tokyo, Japan
{panyadong,nishimura}@nec.com

Abstract. This paper introduces the design and usage of a multi-person pose estimation system. The system is developed targeting some challenging issues in real-world surveillance such as (i) low image resolution, and (ii) people captured in crowded situation. Under such conditions, we evaluated the system's performance on human detection by comparing to other state-of-art algorithms. The leading results by using the proposed system are accomplished by several features in the system's design: (i) training and inference of mid-point, which is the center of two body region points defined in human pose, (ii) core-of-pose which is association of a plurality of body region points, and used as root of each individual person during parsing multiple people under crowded situation. The proposed system is also fast and has the potential for industrial use.

Keywords: Industrial image analysis · Pose estimation · Human detection · Real-world surveillance

1 Introduction

Human pose estimation is recently attracting a great attention and has been studied extensively in action recognition [2,3,25], online human tracking [16,23], person re-identification [20], human-object interaction [6] and human parsing [4]. As to human detection, which is an important task in real-world surveillance, using pose estimation to determine the human bounding area is becoming a more practical way, compared to directly using a human detector such as faster R-CNN [18], SSD [13] or YOLO [17]. This is because in real-world surveillance, especially in public spaces where people often appear in a crowd, (i) some people's bodies are under partial occlusion, and (ii) because of the distance between camera and people, and the requirements of real-time data processing, images and people captured are often with low resolution. These two facts would lead to inaccuracy when using human detector [7]. For example, in the task of recognizing suspicious persons near two countries' border (Fig. 1(left)), animals or even shaking trees are often recognized as human; In surveillance of a pedestrian crossing like Fig. 1(right), it happens a lot that multiple people in a crowd are

S. Palaiahnakote et al. (Eds.): ACPR 2019, LNCS 12046, pp. 239–253, 2020.
https://doi.org/10.1007/978-3-030-41404-7_17

Fig. 1. Examples of real-world surveillance: The left one shows a wide-range surveillance near two countries' border (Getty image). The right one is surveillance of a pedestrian crossing (MOT dataset [14]).

recognized as a single person. To solve such problems, a practical way is to implement bottom-up approach, which means to first detect body region points (each region point represents a certain part of human body), then to build association among those region points in order to get a pose vector for each individual person.

State-of-the-art bottom-up approach include recent works such as OpenPose [1], Art-Track [15] and Associative Embedding [10]. OpenPose uses a part-affinity-field to train the area between each pairwise body region points. Art-Track trains the geometric relationship between head and each of other body regions. Associative Embedding uses a neural network to estimate a person-index-number for each detected body region. In this paper, we designed a bottom-up pose estimation system called NeoPose, and compared NeoPose to those state-of-the-art algorithms for human detection task. The comparisons were conducted on MHP (Multi-Human Parsing) dataset [11], which contains many cases of dense people in the images. We resized all images to smaller size to make the test under low image resolutions. NeoPose gained leading results in the task.

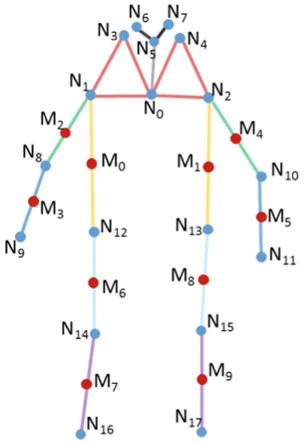
The design of NeoPose is featured by two concepts: mid-point and core-of-pose. Mid-point is the center of two body region points defined in human pose (Fig. 2). We trained and inferred mid-points to help in the association of body region points. In OpenPose [1], the idea of mid-point was mentioned but denied due to concerns of crowded situation. In this paper, we explained when and how to use mid-points. Firstly, we point out that mid-point would be suitable for pose estimation under low image resolution. This was referred to a problem of part-affinity-field, which was used in OpenPose. Second, we used mid-point after human parsing, and supplemented it with a reference of body size. The human parsing and the estimation of body size were realized by what we called core-of-pose.

Core-of-pose is the combination of a plurality of upper body's region points and the links among them (Fig. 5(a)). It is defined on each individual person, and used as root to associate other body region points of the person. What's more, body size of a person could be estimated by referring to the length of links in core-of-pose, and could be used as a criterion for other region points' association. Such criterion helps to reduce the region points' association that crosses different persons. In this paper, we explained the algorithm of building core-of-pose. Compared to a previous work [21] that used a single region point (the head point) as root of each person, and length of head as reference size of human body, our algorithm, using multiple links to build the core, and functioned with the help of mid-points, would thereby reduce the risk of errors under low image resolution and crowded situation.

Overall, targeting real-world surveillance, this paper provides two directions for the design of bottom-up pose estimation and human detection system. (i) training mid-points in order to better support the association of body region points under low image resolution, (ii) using core-of-pose which consists of upper body's region points to parse multiple persons, and to estimate a reference size of each person's body in order to supplement the region points' association among multi-person under crowded situation.

2 Methodology

In this research, human pose is defined as in Fig. 2. Totally 18 body region points are associated to build up one person's pose. During pose estimation, there might be some region points which are not detected, thereby we defined the pose vector of one person as a subset of the 18 body region points. The 10 mid-points are defined according to 10 pairs of region point, each pair of which are physically connected on human body. Mid-points are not involved in pose vector, but help to associate the body region points and to determine the pose vector.



N_1 ~ N_{17}:
body region points following COCO format[19]

N_0:
region point of neck

M_0 ~ M_9:
mid-points
(i.e. M_0 is the center of N_1 and N_{12})

Pose vector:
A subset of {N_0 ~ N_{17}}

Fig. 2. Human pose defined by body region points and mid-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.

Given an image that contains one or multiple persons, pose estimation and human detection by NeoPose are achieved through three steps: (i) generating body region points and mid-points, (ii) generating core-of-pose, and (iii) generating pose vectors and human bounding boxes.

2.1 Generating Body Region Points and Mid-Points

Based on COCO dataset [12], we trained 18 body region points, 10 mid-points as well as a background channel using the deep network defined in Fig. 3(a). Ground truth of body region points except the region point of neck (defined as center of two shoulders) were provided by COCO dataset, and they were used to calculate the ground truth for the region point of neck as well as 10 mid-points. The network starts with a pre-defined VGG-19 [19], followed by two branches, each of which consists of three stages. The output of each stage in the first branch includes 19 channels (18 region points and one for background), while the second branch generates 10 channels for the mid-points after each stage. Concatenation layers between the stages share the features from VGG to the first branch, and from the first to the second branch. After the second and the third stages, all 29 feature maps from two branches are concatenated and used to calculate loss:

$$Loss = \sum\nolimits_T \sum\nolimits_C \sum\nolimits_P W(P) \cdot \parallel S_P^T(P) - S_P^G(P) \parallel_2^2$$

In the loss function, T stands for the second and the third stages, C refers to the 29 channels, and P represents all pixels in the feature map. S_P^T is the score generated from the deep network and S_P^G is the ground truth. W is a binary weight, which returns a value 0 when the annotation 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 from the 29 feature maps.

Compared to the deep network of OpenPose, NeoPose made three changes: (i) The branch for training part-affinity-field (PAF) was replaced by training mid-points. This is a crucial change to make the network better support images with low resolution. Figure 4 shows that PAF would involve too much unreliable information when the image resolution is low. In such cases, utilizing a simple mid-point would help in reducing the risk of errors. (ii) In the design of concatenation layers of NeoPose, the feature sharing goes along a single direction from the branch of region points to that of mid-points, compared to the interactive structure in OpenPose that PAF's features are also shared with body region points. Such a design was made because mid-points were calculated based on region points, and sharing mid-points' feature with region points would lead to multiple detections on each body region. (iii) The number of stages after VGG-19 was reduced from 6 in OpenPose to 3 in NeoPose in order to speed up the inference of region and mid-points. We also found that by reducing the number of stages, the network could recognize more region and mid-points under crowded situation. Such phenomenon was studied in [24], which suggested that repeating the process of convolution would make the network focus more on features of the whole scene rather than individual object/person.

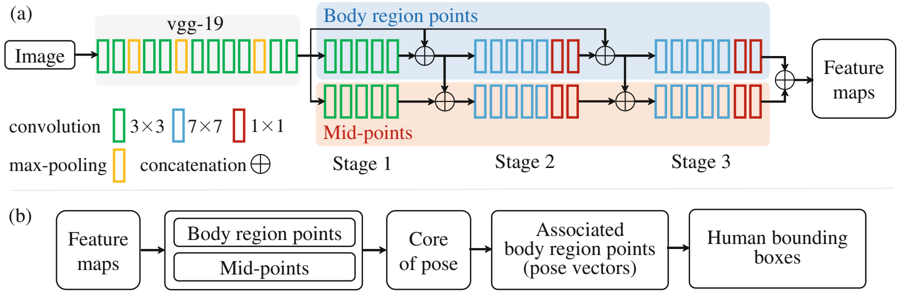


Fig. 3. Architecture of NeoPose. (a) the deep network, (b) the flowchart of data processing after the deep network.

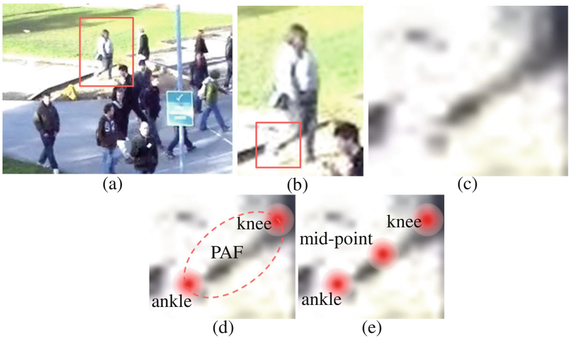


Fig. 4. Description of mid-point. (a, b, c) image, person and body region with low resolution, (d) part-affinity-field (PAF), which is used by OpenPose [1], (e) concept of mid-point used in this research.

2.2 Generating Core-of-Pose

Core-of-pose is defined based on region points of neck, shoulder and ear. These body parts are selected because they are highly spatially correlated on human body, and they are more likely to be captured in real-world surveillance even if people are in a crowd. Six types of link can be included in the core: neck and left shoulder, neck and left ear, left shoulder and left ear, neck and right shoulder, neck and right ear, right shoulder and right ear. Figure 5(a) shows the four types of core (TA, TB, TC, TD) and one midterm format (TE). TA is the full core which has two triangles corresponding to the neck. TB, TC and TD has one triangle. TE is a midterm format and can be converted to TB and TC by excluding the link in the middle of the path between two neck points.

The algorithm for generating core-of-pose (Fig. 5(b)) starts with a graph G and V. G includes all detected body region points of neck, shoulders and ears in the image. V is called full mapping links that consists of all allowable types of link among the region points in G. A pairwise matching algorithm (PMA) is then performed on G and V to filter each type of the link. Assuming that

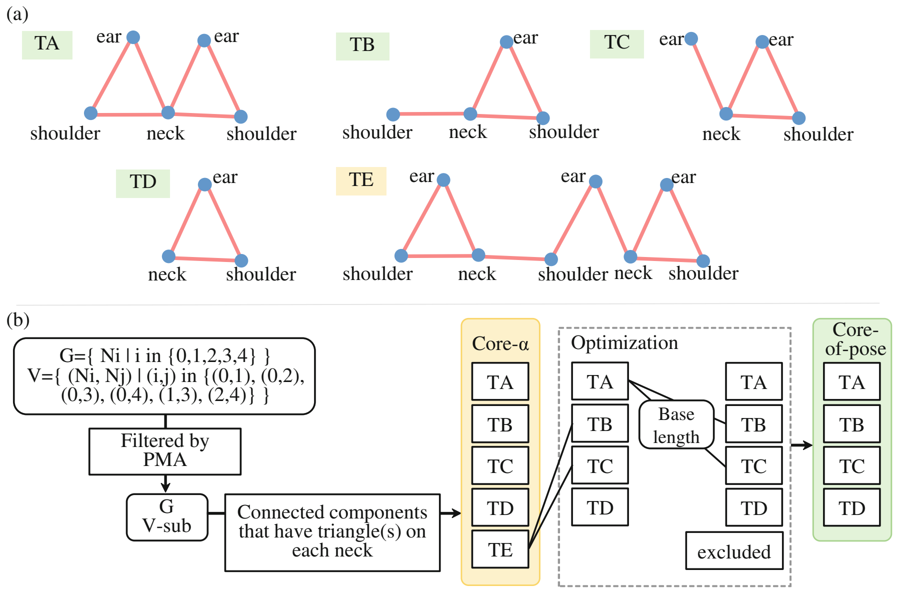


Fig. 5. (a) The four types of core-of-pose TA, TB, TC, TD, and one midterm format TE. (b) Algorithm of generating core-of-pose from the detections of neck, shoulder and ear region points.

a type of link is between two kinds of region points RX and RY, the PMA first accepts only the shortest link for each of RX among its multiple links and removes all other links. Then for each of RY, PMA also accepts the shortest link from its remaining links and removes other links. After performing PMA, the graph G with the filtered links (V-sub) will contain a plurality of connected components. Among all connected components, those which include at least one triangle according to each neck point are accepted and called core-α. TA, as well as TB, TC, TD and TE, are the only five possible types of core-α.

Core-α is not the completed format of core-of-pose. Two steps of optimization are performed on core-α. (Step 1) TE is converted to TB and TC by excluding the link in the middle of the path between two neck points. By doing this, all core-α are aligned with having one neck point. (Step 2) A base length for each core-α is calculated:

$$L_a = \begin{cases} min(|(N_0, N_1)|, |(N_0, N_2)|) & (N_1 \ and \ N_2 \ exist) \\ |(N_0, N_1)| & (N_2 \ does \ not \ exist) \\ |(N_0, N_2)| & (N_1 \ does \ not \ exist) \end{cases}$$

$$L_b = \begin{cases} min(|(N_0, N_3)|, |(N_0, N_4)|) & (N_3 \ and \ N_4 \ exist) \\ |(N_0, N_3)| & (N_4 \ does \ not \ exist) \\ |(N_0, N_4)| & (N_3 \ does \ not \ exist) \end{cases}$$

$$L_c = \begin{cases} min(|(N_0, M_0)|, |(N_0, M_1)|) & (M_0 \ and \ M_1 \ exist) \\ |(N_0, M_0)| & (M_1 \ does \ not \ exist) \\ |(N_0, M_1)| & (M_0 \ does \ not \ exist) \\ L_a + L_b + 1 & (M_0 \ and \ M_1 \ do \ not \ exist) \end{cases}$$

Among multiple detections of M_0 and M_1, we used the closest ones to N_0 for calculating L_c.

$$Base \ length = \begin{cases} L_c & (L_c \leq L_a + L_b, L_b \leq L_a \times 2) \\ L_c \times 1.17 & (L_c \leq L_a + L_b, L_b > L_a \times 2) \\ L_a + L_b & (L_c > L_a + L_b, L_b \leq L_a \times 2) \\ L_b \times 1.7 & (L_c > L_a + L_b, L_b > L_a \times 2) \end{cases}$$

1.17 and 1.7 are fixed referring to $1/\sin 60°$ and $\tan 60°$, assuming that in the core-of-pose of a front-view person, each triangle is an equilateral triangle.

The base length is used to exclude some region points and links in core-α. In a core-α, when the distance between a region point R and the neck point N is over the base length of the current core-α, point R as well as any link associated to it will be excluded from the current core-α. With such process, some of TA will be converted to TB or TC, and some of TB, TC and TD will lose their triangle. Those core-α without a triangle will be excluded and not be used anymore. Throughout two steps of optimization, a plurality of core-of-pose are obtained. We assume that all body region points included in the same core-of-pose are located on the same person.

2.3 Generating Pose Vectors and Human Bounding Boxes

Having core-of-pose, the next step is to associate other types of body region point detected in the image to each core. The association follows an order described in Table 1. Each step of association shares the same algorithm as shown in Fig. 6. Taking the "right shoulder-right elbow" link as an example, PX in Fig. 6 corresponds to right shoulder which is already associated (in core-of-pose), and PY represents right elbow which is to be associated. The algorithm first generates full mapping links between all associated region points of right shoulder and all detected region points of right elbow. The full mapping links are then filtered by two criteria. (i) Length of link should be no more than the allowable maximum length of the current type of link. The allowable maximum length is related to the base length calculated from core-of-pose, and varies according to each type of link (Table 1). The varied length according to type of link is determined based on human's body context [9]. (ii) A mid-point with its type corresponding to the link should be detected in a middle area of the link. As shown in Fig. 7, the middle area is an ellipse area centered on a mid-point M' between two region points N_i and N_j. In this research, R_{major} of the middle area is set to $|(N_i, N_j)| \times 0.35$, and the R_{minor} is set to $R_{major} \times 0.75$. The algorithm excludes the links in which no mid-point exists in the middle area. Note that for those links located on the head, filtering by mid-point is not required. After filtering the links by maximum length and mid-point, the pairwise matching algorithm is performed to optimize the association. As a result, the region points of right elbow on the remaining links are accepted and associated to the right shoulders.

Following the order of region points' association described in Table 1, the full association is completed after hands, feet and eyes are associated to core-of-pose. In case that some types of region point are not detected or not satisfying the proposed criteria, the full association may not be completed. Figure 8 shows some examples of multi-person pose estimation using NeoPose. For each person, body region points, mid-points and core-of-pose are rendered on the image. Having estimated the pose, the human bounding box could be created by enclosing all the region points of a person.

Table 1. Rules for association of body region points based on core-of-pose

Order	Association	Requiring mid-point	Maximum length
1	Shoulder and elbow	Yes	1.5 × base length
2	Elbow and wrist	Yes	1.5 × base length
3	Shoulder and hip	Yes	2.0 × base length
4	Hip and knee	Yes	2.0 × base length
5	Knee and ankle	Yes	2.0 × base length
6	Neck and nose	No	1.0 × base length
7	Nose and eye	No	0.5 × base length

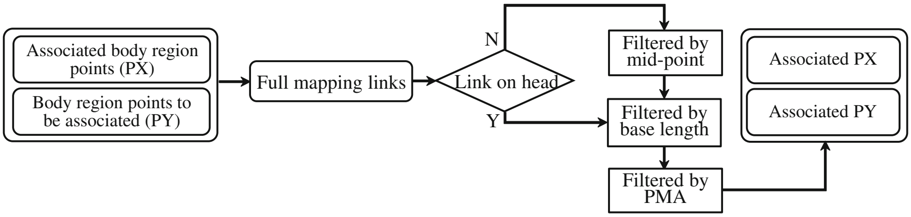


Fig. 6. Algorithm of associating two types of body region point which are physically connected on human body. (e.g. PX is right shoulder and PY is right hip)

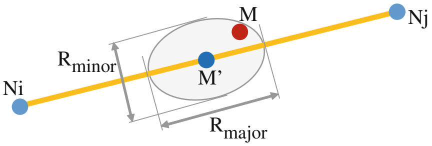


Fig. 7. Mid-point and the middle area. M': ground-truth of mid-point. M: detected mid-point. N_i and N_j: two body region points.

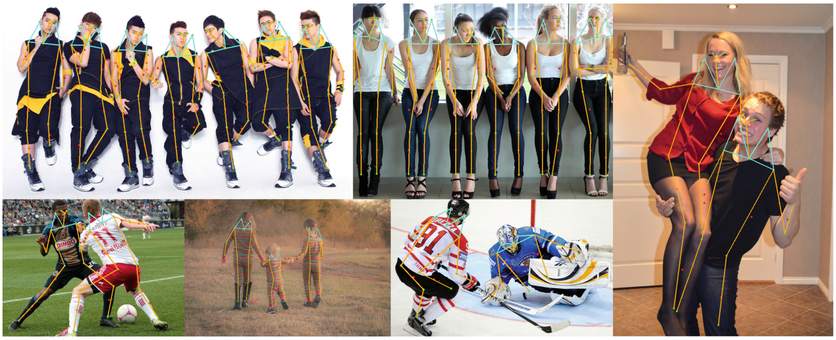


Fig. 8. Multi-person pose estimation by NeoPose on images in MHP dataset [11]. Body region points, mid-points and core-of-pose are rendered for each person.

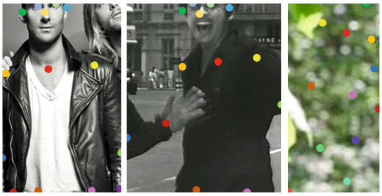


Fig. 9. Three categories of human detection: (left) correct association, (middle) false association, (right) ghost association.

3 Evaluation

To evaluate the quality of NeoPose's human detection function, we performed a quantitative analysis on MHP dataset [11]. MHP dataset contains many cases of dense people in the images, and a variety of different poses in real-life scenes. Its original mission is for testing human parsing algorithms. We considered that MHP is a good dataset to simulate the situation of crowded people in real-world surveillance. What's more, before evaluation, we resized all images in the dataset to a fixed height (120 pixels) without changing the aspect ratio, and used the resized images as input to NeoPose's deep network. By doing this, we simulated the situation of analyzing low resolution images.

We used NeoPose to perform pose estimation on all images in MHP dataset (merging training set with validation set). Figure 8 shows some images in MHP dataset rendered with estimated poses. To evaluate human detection function, we extracted those estimated pose vectors which have at least 10 body region points associated (including the region points in core-of-pose), rendered the region points on the image with color according to different type, and extracted the person's image along with his/her bounding box. We asked two data annotators to manually check those extracted images and classify them into three categories as shown in Fig. 9: (i) correct association, which means all associated region points are located on one person's body without an obvious position error,

(ii) false association, which means associated region points are located on different persons, or some region points are located on the background rather than human body, and (iii) ghost association, which stands for the situation that all associated region points are located on background rather than human body.

Assuming that the position error (PE) of a region point is defined like:

$$PE = |(P_{DT}, P_{GT})|/Hp$$

DT stands for detection and GT for ground-truth. Hp is height of the person.

Since the fluctuation of PE on person with low resolution would be more violent than that on person with larger resolution, it is difficult to fix a threshold of position error for evaluation. In this research, we ask data annotators to manually judge whether the body region points are correctly located or not.

We also performed the same evaluation using state-of-art algorithms including OpenPose, Art-Track [10] and Associative Embedding (AE) [15] under the same criteria. Based on the results of three categories, we computed precision and recall for each algorithm. For NeoPose and OpenPose, we also compared the system's processing speed on MHP since they were implemented under the same framework (assuming OpenPose's speed is 1). Table 2 summarizes the results. The results suggests that OpenPose and Art-Track's precisions are close to NeoPose (difference within 1%). However, NeoPose's recall is much higher than OpenPose and Art-Track. On the other hand, AE and NeoPose's recalls are on the same level (with a 0.7% difference), but AE's precision is 2.4% less than NeoPose. Overall, NeoPose performs the best in the evaluation.

Table 2. Results of human detection on MHP dataset using different algorithms

	GT	Correct	False	Ghost	Precision	Recall	Speed
OpenPose	12319	8762	1499	5	85.3%	71.1%	1
Art-Track	12319	6878	1190	2	85.2%	55.8%	–
AE	12319	9372	1738	125	83.4%	76.0%	–
NeoPose	12319	9284	1516	9	85.8%	75.3%	1.6

4 Discussion

4.1 Associating Parts Rather Than Detecting the Whole Target

In the evaluation of NeoPose, we focused on how it can succeed in correctly associating more than 10 region points for each individual person. This is a practical way of evaluation especially for industrial use. Taking the task in Fig. 1(left) as an example, when the task is to recognize suspicious persons near two countries' border, what is the most important is to confirm that the target recognized is indeed a person. With low resolution of person in the image and a complex environment around the person, human detection directly using a human detector often fails because of the existence of animals, the texture of ground, or even the shaking trees. For such tasks, to first recognize parts of human body as a plurality of distributed evidence, and to check whether they could be associated together, helps in generating more reliable detection result.

4.2 Training Mid-Points

In the deep network of NeoPose, mid-points use the shared features from body region points, which include information of both appearance and location in the image. Considering that the appearance of a mid-point may not have significant difference compared to its nearby points, we assume that the location information of body region points is the dominant factor learnt by the network for inference of mid-points. Such theory of training might have extensions on machine learning of human-object/human-human interaction.

4.3 Triangles in Core-of-Pose

The process of generating core-of-pose contains a criterion that at least one triangle corresponding to the region point of neck should exist. Since each link in the triangle is obtained by performing pairwise matching algorithm that scans all the links of that type in the image, a triangle could thereby suggest that each pair of the three region points are spatially close to each other. Based on such strong spatial correlation, we could assume that the three region points are located on a same person, and use such spatial correlation to parse multiple persons before associating other body region points.

5 Conclusion and Future Work

In this paper, we introduced the design, field-of-use and evaluation of NeoPose. The design of NeoPose - implementing mid-points and core-of-pose - helps the system deal with the difficulties under real-world surveillance. Firstly, training mid-points reduces the risk of errors compared to training part-affinity-field under low image resolution. Secondly, core-of-pose benefits in two ways: (i) using upper body's information - which is more likely to be captured even under crowded situation - to parse multiple persons, (ii) providing a reference size of each individual person's body, and utilizing the size to supplement the association of body region points. These features of NeoPose provide good directions in designing systems for pose estimation and human detection under real-world surveillance.

Using pose estimation for human detection is a practical way for industrial problems. In this paper, the evaluation of human detection on MHP dataset attempted to simulate a variety of different poses under low image resolution and crowded situation. The leading results by using NeoPose compared to other state-of-art methods also suggests that NeoPose has potential industrial use. Currently the use of NeoPose is limited to general road surveillance cameras which are not 360-degree vision or drone-based.

Future work is to test the system in a wide scope of industrial issues depending on customers' need, such as recognizing suspicious person/behavior in public spaces, sports training, worker's skill assessment on production line, animals' behavior analysis on the farm, etc.

Appendix: Human Detection and Pose Estimation on Industrial Scene

As an early result of future work, we tested NeoPose's performance on the surveillance of a pedestrian crossing(from MOT dataset [14]), and compared it to using OpenPose and a human detector Faster-RCNN. Besides human detection, we compared the quality of pose estimation by using the three algorithms as well. For pose estimation, Faster R-CNN was used together with a pose detector called MS Pose [22]. Such kind of approach that estimates pose based on human detection was implemented in some recent researches [5,8]. For both the tasks, the resolution of the image input to the deep network was (width: 768 pixels, height: 576 pixels), and the height of each person was less than 120 pixel as we did in the evaluation based on MHP dataset.

Figure 10 shows some typical examples of the results. Regarding human detection, there happened a lot in cases of using OpenPose or Faster-RCNN that multiple persons who were in a crowd were recognized as an individual person, or an object that occluded a person was recognized as part of the person (Fig. 10(a)), while such mistakes were much fewer in case of NeoPose. A second issue is that OpenPose generated lots of false detection of region point especially on the ground (Fig. 10(b)). Those false detections would make the association of region points much slower. Another issue in MS Pose is that under crowded situation and with low resolution, many body region points were not correctly located (Fig. 10(b)). Similar issues also happened in other surveillance scenes in MOT dataset. These early results reveal NeoPose's potential in solving both human detection and pose estimation under low resolution and crowded situation.

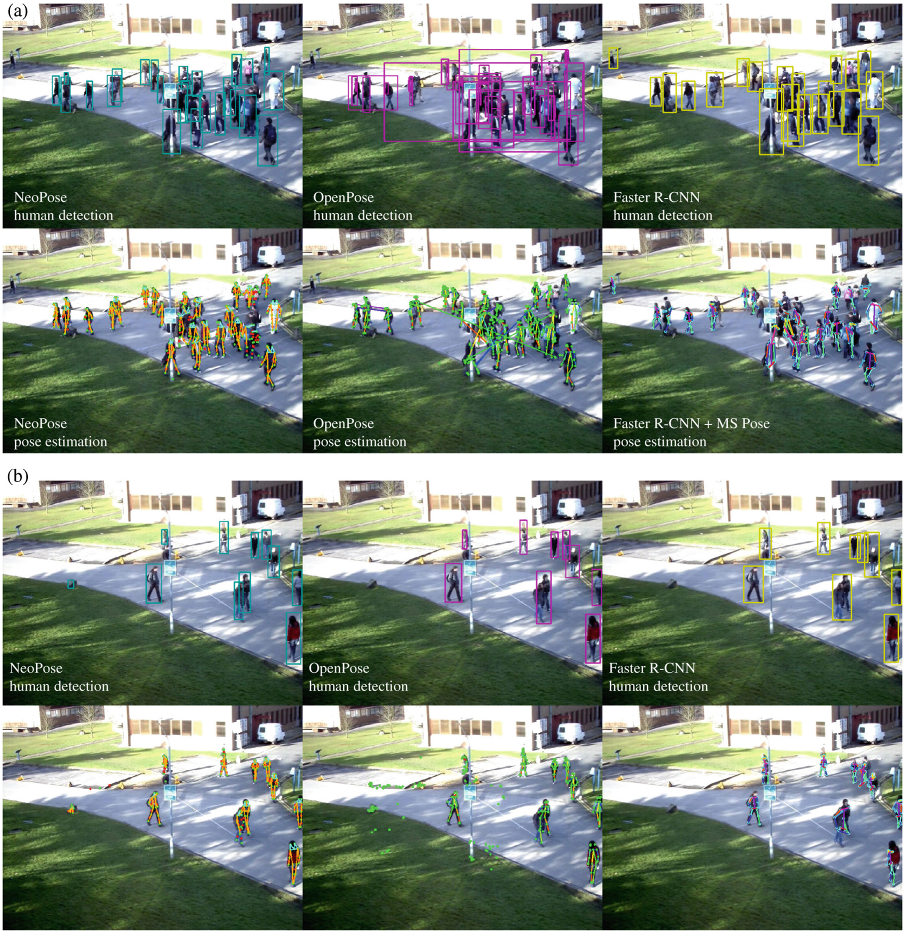


Fig. 10. Comparisons of human detection and pose estimation on surveillance images of a pedestrian crossing (MOT dataset) using different algorithms. (a) and (b) are two representative moments.

References

1. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
2. Choutas, V., Weinzaepfel, P., Revaud, J., Schmid, C.: Potion: pose motion representation for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7024–7033 (2018)
3. Demisse, G.G., Papadopoulos, K., Aouada, D., Ottersten, B.: Pose encoding for robust skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 188–194 (2018)

4. Fang, H.S., Lu, G., Fang, X., Xie, J., Tai, Y.W., Lu, C.: Weakly and semi supervised human body part parsing via pose-guided knowledge transfer, pp. 70–78 (2018)
5. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2334–2343 (2017)
6. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8359–8367 (2018)
7. Gkioxari, G., Hariharan, B., Girshick, R., Malik, J.: Using k-poselets for detecting people and localizing their keypoints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3582–3589 (2014)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
9. Herman, I.P.: Physics of the Human Body. BMPBE. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-23932-3
10. Insafutdinov, E., et al.: Arttrack: articulated multi-person tracking in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6457–6465 (2017)
11. Li, J., et al.: Multiple-human parsing in the wild (2017)
12. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
13. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
14. Milan, A., Leal-Taixe, L., Reid, I., Roth, S., Schindler, K.: MOT16: a benchmark for multi-object tracking (2016)
15. Newell, A., Huang, Z., Deng, J.: Associative embedding: end-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems, pp. 2277–2287 (2017)
16. Raaj, Y., Idrees, H., Hidalgo, G., Sheikh, Y.: Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4620–4628 (2019)
17. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
20. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3960–3969 (2017)
21. Varadarajan, S., Datta, P., Tickoo, O.: A greedy part assignment algorithm for real-time multi-person 2D pose estimation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 418–426 (2018)
22. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of European Conference on Computer Vision (2018)

23. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose flow: efficient online pose tracking. In: Proceedings of British Machine Vision Conference (2018)
24. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2131–2145 (2019)
25. Zolfaghari, M., Oliveira, G.L., Sedaghat, N., Brox, T.: Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2904–2913 (2017)