Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/13/20230728

The comparison of top-down and bottom-up methods in multi-person pose estimation

Ruijie Li

Dalhousie University, Halifax, B3H 4R2, Canada

Rj569393@dal.ca

Abstract. 2D multi-person pose estimation is a process to detect all bodies in a two-dimensional picture. The main purpose of this report is to discuss the difference between top-down and bottom- up method in multi-person human pose estimation. This estimation will focus on many people in one picture. There are two popular methods in this area. The first is top-down method, which is to find people firstly and then to detect the body of each people. On the other hand, the bottom-up method is to detect body parts and then find each person. Their comparison and analyse in algorithm, speed and accuracy may help researchers to find more suitable methods when they research in human pose estimation. Generally, top-down methods have higher accuracy than bottom-up methods but have higher speed.

Keywords: multi-person pose estimation, top-down methods, bottom-up methods.

1. Introduction

Multi-person pose estimation is one of the newest and basic fields in computer vision. 2D multi-person estimation is to find all persons in an image and detect their key joints. Single person estimation is the prerequisite of the multi-person estimation. Obviously, there is only one person in an image, and we must detect this person and body parts, such as head, hands and feet. If multi-person estimations are developed in the video, human pose tracking will be realized. Moreover, inputting RGB image and add another coordinate for the multi-person estimation will create 3D human post estimation. Multi-person estimation implements convolution neural network (CNN) to detect body parts and deep learning network can also represent feature map of images. For the 2D multi-person estimation, the MPII and COCO databases are most popular. There are number of methods in 2D multi-person estimation, but generally they can be divided into 2 types: bottom-up methods and top-down methods. Bottom-up methods usually detect all persons in an image and then detect body joints of each person. Top-down methods will detect all key joints initially and then link them to their person. More specifically, this report will discuss the bottom-up and top-down methods in multi-person human pose estimation as well as compare them in different aspect. Furthermore, the clear comparison will assent to develop pose estimation and solve current problems in 2D multi-person estimation. The following parts will analyse 5 main methods in bottom-up and top-down methods. They are deepcut, ArtTrack, G-RMI, Part Affinity fields (PAF) and end-to-end joint detection. Their comparisons analyses will be represented in the final part.

DOI: 10.54254/2755-2721/13/20230728

2. Analysis

2.1. Top-down method

Multi-person pose estimation top-down method is detecting all people in a picture and using boarding-box to circle each person and finally detect their body parts. Usually, MPII Multi-Person Dataset and MSCOCO datasets are the main datasets in top-down method.

In the deep cut method, convolution neural network (CNN) will be used to detect body parts in a picture. "DeepCut is a state-of-the-art approach to multi-person pose estimation based on integer linear programming (ILP) that jointly estimates poses of all people present in an image by minimizing a joint objective [1]." Deeper cut has higher accuracy and efficiency than deep-cut methods. In the Deeper-cut methods, the two variables X and Y used in ILP separate different classes of body parts to the distinct humans. Insafutdinov et.al proposed that they improved estimation methods by using branch-and-cut algorithm, saving 4-5 times running time [1]. Moreover, the image-Conditioned Pairwise Terms will reduce the number of joints and increase the speed of detecting [1]. Even though Integer linear program (ILP) refines the human pose modules, there are some shortcomings in this method, running fast R-CNN and ILP at the same time is a difficult and slow work. In the testing, some body parts are too close to be detected: geometrically is consistent, but appearance is not connected well. Modelling consistency will be focused on the future to solve this problem [1]. Another problem is that body parts cannot be linked to the correct comments when several bodies parts from one person are clustered into different groups or some backgrounds are included in body parts [1].

ArtTrack method is based on the architectures point for ingle-fame pose estimation and implementing simplified body-part as well as move a mass of computing task in the front of convolution neural network to reduce the processing time [2]. In the process of articulated multi-person tracking, local search methods solve the subgraph multi-cut problems [2]. The simplified connection with the fewer edges can increase the detecting speed. A new convolutional body-part detector was created to detect people's location and the tasks of associating body parts is moved to the proposal machine [2]. Thus, if the structure clutters, the system can still detect the joints of each person. ArtTrack method focuses on complex real environment. First, this method will count the number of people in the environment. Second, detect each joint and rigid parts of each person. Specifically, head detection has fewer errors in implementation, so head and neck detecting are first. At the same time, the cross-entropy binary classification loss can predict the heatmaps of body parts. And a single person will be detected initially from neck to all body parts. Finally, after the convolution network, all body parts are detected and connected joints in each person [2]. Although this method is efficient in the real word, there are some problems. When some people are covered by other people or persons on the edge of the photo, they will not be recognized, since geometric image-condition pairwise cannot tell each over-lapping body parts and misuses the post-CNN to link to body poses. The combination of multiple image cues and the joint modeling of multiple individuals and reasoning in the multi-perspectives may solve these problems [2].

G-RMI team improves the convolution method in the multi-person post estimation and won the first place in 2016 MS COCO test change. They emphatically implemented 2 steps in optimization. Firstly, they estimate the location and size of box by using a fast-RCNN [3]. Secondly, the key points of each people with a bounding box are predicted and the dense heatmaps and offsets are estimated by a fully convolution ResNet [3]. For the purpose to combine them together, they also implement a brand-new aggregation procedure and they also use the key Point – based Non-Maximum-Suppression (NMS) to estimate bounding boxes [3]. In the person box cropping part, they increase the size of the bounding box with the original proportion to make sure persons in the bounding box. After that, they use ResNet-101 backbone to get a denser output feature map [3]. In order to detect more than one person in the graph, they implement the combination of classification and regression methods: finding heatmap initially and then predicting a 2-D local offset vector. Moreover, they add two new "offsets" in the output of network, instead of using the original "heatmap" [3]. Therefore, this method can connect joint and rigid by combining these three links, which improve the accuracy of estimation. G-RMI method implements a

DOI: 10.54254/2755-2721/13/20230728

dichotomy loss function and a regression loss function, and the dichotomy loss is used to recognize key points. Thus, G-RMI method improve the accuracy of joint recognition in multi-person estimation.

2.2. Bottom-up method

Multi-person bottom-up method is to detect body parts first, and then link these body parts as well as finally combine to people.

One of the most famous methods of bottom-up is using Part Affinity Fields (PAFs) in Realtime multi-person 2D estimation, which has high accuracy to estimate most of all the people in the photo by using the bottom-up greedy algorithm. For the parts detection, they use neural network and convolution network to detect key points of bodies in the heat map. In the convolution network, Cao et.al implemented many learning stages to decrees errors from each previous stage. The set of 2D vector of each key points as well as the greedy inference accurately calculate direction and position of each person [4]. Traditionally, for each confidence map, each body part will be shown as a pixel location and each pixel has one confidence coefficient. When a person appears in the photo, the confidence map will get to the peak. However, when many people are close with each other, this confidence map will have some errors, since they only detect locations of body-parts detected but lose directions [4]. Therefore, Cao et.al implement the part affinity field, which is "a set of flow fields that encodes unstructured pairwise relationships between body parts" and can keep location and direction of each body-part by 2D vector fields [4]. When they combine body parts, deep learning network and parts association run at the same time and help each other. In the test, errors occur in estimating multi person cross body parts. When some people are cross with each other, the detector will produce unclear link with key points. However, this error may be corrected in the real environment, so 3D estimation and addition of a depth can solve this problem. In the future, the techniques of full body pose estimation and 3D human shape estimation of multiple people will continue to improve.

End-to-end joint detection method implements end-to-end associative embedding, which is signing a vector embedding tag on body point to detect their group. Predicting an embedding for each candidate and detecting score work together to group points with same tags together. In this process, pixels of images will be assigned into their regions. After scanning a picture, each picture is allowed to input its detection score and vector embedding of each pixel. At the same time, stacked hourglass structure are used in main network: all features will be processed by a standard set of convolutional and pooling layers and combined their results with higher resolution until they get predicted results. Stacked hourglass models initially are implemented in single person estimation, detecting heatmap of each joint of people. However, Papandreou et.al increases the capacity and network to implement this method in detecting multi person, so they add more peaks and a detection loss on the output heatmaps, which can computer errors between predicted result and "ground truth" heatmap [5]. When they get peaks on the heatmap, they detect their tags and those same tags are linked together to the same person by using associative embeddings. Papandreou et.al. iterate detecting process until each part are assigned to a person. They used the first result from detecting body part to make a first detecting pool, other body parts are defined by their scores and tags [5]. However, the resolution rate decreases after pooling. Therefore, the Multiscale Evaluation accessed pictures by averaging value and modifying heatmaps. Finally, they also add a vector to associate tags of each pixel to process tags in each group. End-to-end joint detection method combines associative embeddings with convolutional neural network to realize multi-person detection. This method can be used in instance segmentation and multi-objects tracking in videos.

2.3. Compare parts

Top-down methods have more computing tasks and easier algorithm than bottom-up method. Top-down method will firstly implement target detection to find all people and analyse each of them separately, which is repeating single person estimation. Accordingly, for top-down method, detecting single person needs more focus. Each key join of a person will be detected initially by heatmap, and then connect them to a complete skeleton following the predefined order [6]. Usually, key joints are detected from heatmap

DOI: 10.54254/2755-2721/13/20230728

directly, and there are two methods that can still find key points in pictures. The first one is Graph-PCNN, which adds a regression network on the original heatmap to get more accurate location output [7]. The other method is LCR-Net, assigning a skeleton to a person and comparing the difference between ground truth and the assigned skeleton to get a better result than the standard non maximum suppression algorithm [8]. These two methods improved some detecting part of the traditional methods. In conclusion, top-down methods have more focuses on methods of single-person estimation. On the other hand, the most important part for bottom-up methods is connecting each joint and assigned them to people correctly. Bottom-up methods will not only detect location of each joint, but also estimate connections between these joints. Since detecting joints can be realized by the single person estimation, connecting joints have much research. One of the most popular methods is PAFs, which can track 2D vectors from each key points and finally get the accurate location and direction of each joint. The other approach in bottom-up method is making attaching the same ID to the joints of the same person, which is from end-to-end associative embedding method. However, in the real lives, the number of people in the picture is unclear, so initially, close joints from the same person have similar ID and far joints have different ID. Accordingly, if the bottom-up methods wanted to be improved, the algorithms of connection joints needed to be optimized. However, when researchers add more detectors to catch people in the picture, the top-down methods will be improved. To sum up, the algorithms of top-down methods are easier than bottom-up methods' algorithms.

Comparing with bottom-up methods, top-down methods have higher accuracy. A traditional approach for top-down method is person detection followed by pose estimation. Since person detection has a relatively mature technology, it can detect most of all the people in a picture. There are two approaches for most of the person detecting methods: artificial feature exaction and classifier as well as deep learning methods. Take G-RMI method for instance, they implemented R-CNN system to detect person in a picture [3]. Firstly, an input picture will be exacted a feature map by convolution neural network and pooling. And following Region Proposal Network (RPN) separates feature map as several small parts to tell the foreground and get coordinates of these foreground. After getting location of each anchor, they train data to get an exact position regression. Finally, they use the exact position to extract the relative target from feature map and pooling into fixed-length data [9]. By choosing faster R-CNN, G-RMI method has a high accuracy in detecting people of photos. For most of bottom-up methods, body joints will be detected firstly and then connect them and assign them to each person correctly. Take PAFs method for instance, PAFs is based on open the Convolutional Pose Machines (CPM). CPM proposes an order structure of CNN which can directly run in the belief map and return the joint points of each person appearing in the image [10]. Unrelative points will be removed following by center map. CPM also offers an intermediate supervision study to solve vanishing gradient problem during training. The traditional estimation process, image feature model and context information generation in the pose machines are replaced by the deep convolution structure. The connecting method, OpenPose, in Cao's article is comparing each vector between 2 key points and then tell whether those vectors are on the same direction with the line connecting those key points. If their directions are same, this connecting is correct. By implementing CPM and OpenOpse, PAFs method output an excellent result in detecting accuracy. However, if there are many overlaps condition on the picture, bottom-up methods have lower accuracy than top-down. According to the data from G-RMI method, the final average precision is 0.649 on the coco development dataset and 0.643 on the benchmark dataset, which is better than other methods in COCO Key point challenge [3]. One of the reasons is that bottom-up methods must detect and connect all the key points in an image, but top- down method can rescale the patch of each detected region to a larger size, which suffers less degradation at a smaller scale. In addition, the result of top-down method has more dependence on pedestrian detector. Some shortcomings of the bottom-up method are unclear orientation, meaning that there are some improvements in capturing spatial dependencies [4].

One of the solutions about scale change problems in the bottom-up method is Scale-Aware High-Resolution Network (HigherHRNet). HigherHRNet creates a higher resolution ratio heatmap by a new pyramid model. This new model starts at the highest resolution ratio of backbone concentrator nodes and improves the resolution by Deconvolution feature maps. In addition, Cheng et.al also proposes a

DOI: 10.54254/2755-2721/13/20230728

brand-new Multi-Resolution Supervision strategy, assigning different training objects to the relative levels of pyramid model. Finally, implementing a multi-resolution heat map aggregation strategy to generate high resolution heat maps with scale sensing [11]. This method can solve the problem about scale change in multi-person attitude estimation by Multi-resolution supervision for training and multi-resolution aggregation for reasoning.

Generally, the most bottom-up methods have less running time than top-down methods. It is acknowledged that the top-down methods have to repeat single person detection many times, but bottom-up methods process all people in an image together. Especially for the multi-persons, top-down methods are slower than bottom-up methods. For instance, Deeper-Cut method implements CNN as the body detection first and then assign each joint to their person. The ArtTrack method also counts the number of people in the image and then detect each body joints. Therefore, the process of detecting all person in the top-down method spends more time than the process of finding all key joints directly in the bottom-up method. In the end-to-end associative embedding method, network is taught to output both detection and packet assignment, saving much time in processing detection. In addition, the whole estimation is done in one session, eliminating the complex post-processing steps. Thus, for the running speed, bottom-up methods are quicker than top-down method.

3. Conclusion

Accordingly, top-down methods have more accuracy and easier algorithms but lower speed. However, bottom-up methods can process images more quickly but often miss persons in multi-person images. It is considered that the top-down methods are combining many single person estimations, and they implement CNN in semantic segmentation, so they hardly miss people in an image, but their processing time will increase with number of people increasing. On the other hand, the bottom-up methods detect all key joints initially and when they assign these body parts to relative persons in an image with many people, misconnection will happen. Thus, bottom-up methods have lower accuracy than top-down methods. Currently, increasing the detecting speed of top-down methods and decreasing misconnection of bottom-up methods are hardest problems to solve. In the future, implementing with 3D multi-person estimation in bottom-up methods may solve body overlap problem. 3D detectors will tell body joints more clearly than 2D detectors. Therefore, increasing the accuracy in the multi-person image and detecting speed is the direction of future development of human pose estimation. 2D human pose estimation can be implemented in animation. To track people in real world can simulate more natural behaviours in animation, even realize interaction with humans. Moreover, implementing human pose estimation in tracking body behaviours can also assist to improve gestures in dance and fitness. Similarly, human estimation and analyzing can also be used in tracking suspects.

References

[1] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, & B. Schiele. (2016). DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In European Conference on Computer Vision (ECCV), May 2016. 4321, 4322, 4327

[2] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, & B. Schiele. (2017). Arttrack: Articulated multi-person tracking in the wild. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[3] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, & K. Murphy. (2017). Towards accurate multiperson pose estimation in the wild. arXiv preprint arXiv:1701.01779, 8, 2017. 4328

[4] Z. Cao, T. Simon, S.-E. Wei, & Y. Sheikh. (2017). Realtime multiperson 2d pose. estimation using part affinity fields. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 4327, 4328

[5] A. Newell, Z. Huang, & J. Deng. (2017). Associative embedding: End-to-end learning for. joint detection and grouping. In Advances in Neural Information Processing Systems, pages 2274–2284, 2017. 4327

[6] Luo, X., & Li, F. (2021). Stacked hourglass networks based on polarized self-attention for human pose estimation. Second IYSF Academic Symposium on Artificial Intelligence and Computer Engineering. https://doi.org/10.1117/12.2622889

[7] Wang, J., Long, X., Gao, Y., Ding, E., & Wen, S. (2020). Graph-PCNN: Two stage human pose estimation with graph pose refinement. Computer Vision – ECCV 2020, 492–508. https://doi.org/10.1007/978-3-030-58621-8_29

[8] Rogez, G., Weinzaepfel, P., & Schmid, C. (2017). LCR-net: Localization-classification-regression for human pose. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.134

[9] Zhang, L., Lin, L., Liang, X., & He, K. (2016). Is faster R-CNN doing well. for pedestrian detection? Computer Vision – ECCV 2016, 443–457. https://doi.org/10.1007/978-3-319-46475-6_28

[10] Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional Pose Machines. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.511

[11] Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr42600.2020.00543