A Comparative analysis for Single person and Multi person Pose estimation using Deep learning algorithms

Manisha Patel
Research scholar, E. C. Department
R.K.University
Rajkot, India.
mpatel708@rku.ac.in

Nilesh Kalani
Director IQAC
R.K. University
Rajkot, India
nilesh.kalani@rku.ac.in

***Abstract* - Due to numerous application of Human posture application in the real world, it has recently gained a lot of attention. This paper gives a complete assessment of deep learning-based human pose estimation methods and analyses the methodology used, because deep learning can improve the performance of state-of-the-art human pose estimation methods. Pipelines for single people and multi-person are examined individually first. The Alpha pose and Open pose deep learning frameworks used in these pipelines and, are then compared and analyzed.**

***Index Terms* - Pose Estimation, Alpha Pose, Open Pose.**

I. HUMAN POSE ESTIMATION

A. Overview

The purpose of computer vision is to simulate and model human eyesight. This incorporates deductive reasoning concerning human-object interactions [7]. A machine must first have access to postural interpretations generated from visual input in order to perform such high-level reasoning. Pose estimate is the process of determining a person's postures in images or videos [9].

The articulated posture estimation is formulated as shown in figure 1: Given an image of a human body and a skeleton based model of a human body will generated, which is made up of a collection of joints (key points) such as ankles, knees, shoulders, elbows, wrists, and limb orientations.

B. Top-down approach [11]

In this approach, the network first find out humans and uses an object detector to draw a bounding box around it. After that, it will estimate the key points of each bounding box.

Input Image Output Pose
Figure 1. Human Pose Estimation

.

Figure. 2 Top- down approach

978-1-6654-0203-3/21/$31.00 ©2021 IEEE

C. Bottom-up approach [12]

In this approach the model detects, all available key points of all humans, and then it will assemble a group of key points (with the heat map) into skeletons, to show distinct objects.

Figure 3. Bottom- Up approach

The quality of the detected bounding boxes determines perfection of Pose Estimation in Top-down approach. When two or more people are too near together, the assembled human poses in the Bottom up framework are confusing.

D. Single Person Pose Estimation and Multi Person Pose Estimation [8]

Posture estimation is divided into two types based on the number of persons being tracked: single-person and multi-person pose estimate. Single person is easy in compare of multi person, as there is no issue of inter-person occlusion.

II. POSE ESTIMATION WITH ALPHA POSE FRAMEWORK [8]

A. Overview

Alpha Pose is based on a Top-down framework. It can recognize correct human poses even when the boundary boxes are incorrect. To overcome the problems of bottom-up approach, Object detector Faster R-CNN and SPPE Stacked Hourglass model [2] is used in this framework. Even if the bounding boxes are regarded valid with an IoU > 0.5, the identified human poses may be incorrect.

RMPE system is suggested to address the aforementioned issues. SPPE-based human pose estimation algorithms perform better with this framework. To takeout human figure from erroneous selection, SSTN is linked with SPPE. To optimize this network, a new parallel SPPE branch is introduced. A parametric posture NMS (Non-Maximum- Suppression) is developed to overcome problem of duplicated detection. Parametric pose NMS reduces superfluous poses by comparing pose similarity using a unique pose distance measure. The posture distance parameters are optimised using a data-driven technique. Finally, to supplement training data, a unique pose-guided human proposal generator (PGPG) is proposed.

B. RMPE (Regional Multi-Person Pose Estimation) Framework

The framework is collected of 3 sections:
(1) SSTN stands for Symmetric Spatial Transformer Network, extracting single-person regions from erroneous bounding boxes.
(2) Non-Maximum-Suppression (NMS) in Parametric Pose, Non-Maximum Suppression of Parametric Attitude: Resolving Redundancy
(3) Enhance Training Data with the Pose-Guided Proposals Generator – PGPG Attitude Guide Area Box Generator On the MPII dataset [14], this approach is capable of handling faulty bounding boxes and redundant detection, with a score of 76.7mAP.

The target detection method obtains the target area of the human body, as shown in Figure 4. Then use the STN+SPPE module to automatically detect human body position using the area frame. Refine it by PP-NMS. Parallel SPPE is utilized during the training phase to avoid local optimization and improve the SSTN effect. Create the PGPG structure to enhance the current training sets.

For a comparative study and analysis of Alpha Pose[8] and Open pose[5] algorithms, , we have implemented both said algorithms using python on 120Hz Intel Core i7-10750H 10th Gen GPU, GTX 1660Ti and tested the performance with results on our own real time images.

1) Symmetric STN and Parallel SPPE

SPPE is not well-suited to human proposals provided by human detectors. This is due to the fact that SPPE was developed specifically for single-person images and is extremely sensitive to localization problems. Small translations or cropping of human ideas have been demonstrated to have a significant impact on SPPE performance. When given unsatisfactory human proposals, symmetric STN + parallel SPPE was proposed to improve SPPE

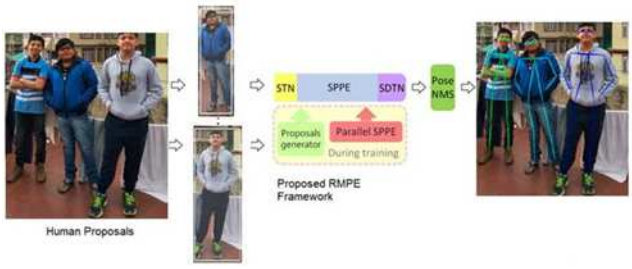

Figure 4. Pipe line of RMPE Frame work

As per the result shown in figure 4, Pipe line of RMPE Frame work with SSTN and parallel SPPE Module has removed occlusion problem in multi person image and has given appropriate pose estimation.

The spatial transformer network (STN) has shown to be quite good at automatically picking regions of interest. They employ the STN in this framework to extract high-

978-1-6654-0203-3/21/$31.00 ©2021 IEEE

quality remarkable human suggestions. SPPE can be used for error less posture prediction after isolating high-quality dominating human proposal zones. The SSTN is fine-tuned in conjunction with SPPE during training. SPPE in parallel STN includes a parallel SPPE branch in the training phrase to aid in the extraction of good human dominant areas. This branch has the same STN as the original SPPE, but it doesn't have the spatial de-transformer (SDTN). This branch's human posture label is specified to be centered.

During the training phase, all of the layers of this parallel SPPE are frozen. The weights of parallel SPPE are fixed along with feedback path, is provided to STN for detecting accurate human poses. The parallel branch will propagate substantial errors if the extracted posture of STN is not located in center. STN will be able to focus on the right location along with clearly identify human regions in given images. The parallel SPPE is rejected during testing process.

2) Parametric Pose NMS

Human detectors are prone to producing unnecessary detections, which result in unnecessary pose estimations. To overcome the above problem, pose non-maximum suppression (NMS) is necessary. First, the best confident posture is chosen as a reference, along with remaining poses that are near to it are removed using an elimination algorithm. This technique is repeated for remaining positions till only unique poses are recorded and no redundant poses exist.

3) Pose guided Proposals Generator

Proper data augmentation allows SSTN+SPPE adapt to incorrect body area placement results in two-Stage gesture recognition (first locates the region, then pose point positioning). Otherwise, while running during the testing phase, the framework may not be used to odd human positioning findings. Using the detected area box during the training phase is an intuitive way. Target detection, on the other hand, simply creates one location area for a person. A certain impact can be achieved by utilizing the created human body placement. We can construct a large sample training set using samples that are consistent with human body detection findings because we already have each person's true position and the detected positioning frame.

C. Evaluation Datasets and results

Alpha pose framework was tested on entire MPII multi-person test set. They identify difficult joints like wrists, elbows, ankles, and knees with an average accuracy of 72 mAP. Alpha pose was also trained, tested and validated on MSCOCO dataset.

III. POSE ESTIMATION WITH OPEN POSE FRAMEWORK [5]

A. Overview

The algorithm accepts a colour image as an input and produces 2D positions of anatomical key-points for every human in the image as an output. The architecture depicted in Figure 5, predicts both detection confidence maps and affinity fields, which shows complete relationship.

B. Open pose framework

The network is divided into 2 sections: the top section, depicted in beige, predicts confidence maps, while the bottom section depicted in blue, predicts affinity fields. Following Wei et al. [3], each section is an iterative prediction architecture that refines predictions over successive phases, tϵ {1,….T}, with intermediate supervision at each stage. A convolutional network (initialized using the first 10 layers of VGG-19 and fine- tuned) analyses the image first, generating a collection of feature mappings F that are fed into the first step of each section. The network generates collection of detection confidence maps S1 = ρ1 (F) and set of part affinity fields L1 =φ1 (F), in the first stage, where ρ1and φ1 are the CNNs for inference. The predictions from both sections in last step, as well as the original picture features F, are concatenated and used to construct refined predictions in each following stage.

$$S^t = \rho^t (F, S^{t-1}, L^{t-1}), \text{ for all } t \geq 2$$

$$L^t = \varphi^t (F, S^{t-1}, L^{t-1}), \text{ for all} t \leq 2$$

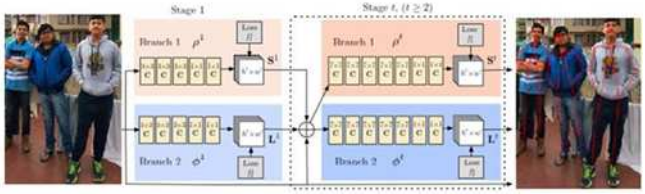

Figure 5. Open pose Framework

Between the estimated forecasts and the ground truth maps and fields, they apply an L2 loss. The confidence maps are subjected to non-maximum suppression to produce a discrete set of body part candidate locations.

They provide numerous significant innovations in this framework, such as a representation of important points that incorporates both the location and orientation of human limbs, and a second architecture that learns part detection and association simultaneously. High-quality body posture parses may be demonstrated regardless of the amount of persons in the image.

978-1-6654-0203-3/21/$31.00 ©2021 IEEE

C. Evaluation Datasets and results

The MPII human multi-person dataset and the COCO 2016 key-points challenge dataset were used to assess the Open pose framework for multi person posture estimation. These two datasets collect photos in a variety of settings that include crowding, scale variation, occlusion, and touch, among other real-world obstacles. The time required for inference is 6 orders of magnitude less.

IV. RESULTS AND CONCLUSION

We present qualitative and quantitative results by implementation of Pose Estimation algorithms Alpha pose and Open pose on single person image (with far and near person)as per Figure 6, Multi person image (with far and near person) as per figure 7 and right key point location on SPPE as per figure 8.

Figure 6 Comparison of Far and near person (Single person)

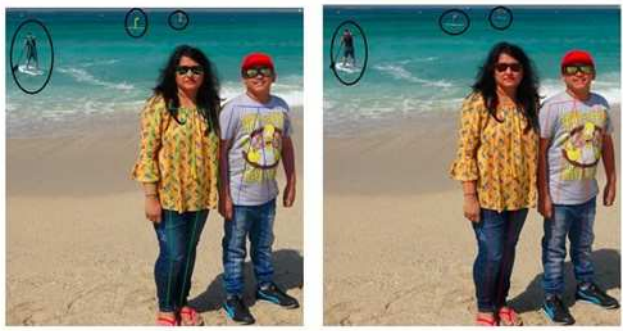

Figure 7 Comparison of Far and near person (Multi person)

The performance of both framework can be analysed in figure 6 , 7 and 8. In figure 6 and 7, we can clearly say that Alpha pose framework can identify far person and can present pose estimation of far person. In figure 8, specifically key points of eyes, nose, elbows and wrists are clearly estimated in Alpha pose.

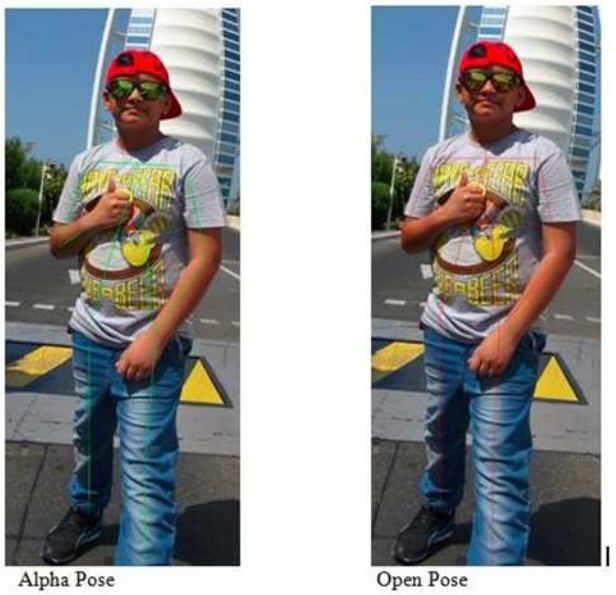

Figure 8. Comparison of Key point location on single person

V. FUTURE SCOPE

Deep neural network (DNN) models are frequently employed in applications that need video stream content analysis. These models are usually trained on servers with powerful GPUs. More processing resources are required for training of larger and deeper neural networks. Deployments of such deeper neural network are difficult in some applications like mobile phone, embedded networks etc. Deep neural networks consume a lot of energy and have a significant carbon footprint.

We may use neural network compression methods to create lightweight, efficient neural networks may be placed in appliances with restricted computational abilities. Many redundant parameters in large neural networks have little effect on the network's performance. Pruning or eliminating these unnecessary parameters reduces the complexity of networks.

REFERENCES

[1] A. Toshev and C. Szegedy, "Deep pose: Human Pose estimation via Deepneural networks", 2014 IEEE conference on Computer visionand pattern recognization.

[2] Alegjandro Newell, KAIYU Yang and Jia Deng" Stacked Hourglass network for Human pose Estimation", European conference on computer vision- ECCV 2016, pp. 483-499, Springer

[3] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh "Convolutional pose machines" In CVPR, 2016 IEEE.

[4] Jonathan Tompson, Ross Goroshin, Arjun Jain, Y ann LeCun, ChristophBregler, Efficient Object Localization Using Convolutional Networks, CVPR 2015, open access and available at IEEE Xplore.

[5] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Real time multi-person 2d pose estimation using part affinity fields. In CVPR, 2017

978-1-6654-0203-3/21/$31.00 ©2021 IEEE

[6] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Mur-phy. Towards accurate multi-person pose estimation in the wild. In CVPR, 2017
[7] Sakshi Indolia, Anil Kumar Goswasmi, S.P.Mishra, Pooja Asopa, " Conceptual Under-standing of Convolutional Neural Network_ A Deep Learning Approach" ICCIDS 2018, 679-688
[8] Fang, H., Xie, S., Ti., Y., C.: RMPE: Regional multi person pose estimation. In 2017 IEEE international conference on computer vision (ICCV) Venice, PP 2353-2362(2017)
[9] Anubhav Singh, Shruti Agarwal, Preeti Nagrath, Anmol saxena, Narina Thakur, " Human Pose estimation using Convolutional neural networks" 2019 Amity international conference (AICIA)
[10] Szegedy, C., Toshev, A., and Erhan, D. (2013) “Deep neural networks for object detection.” In Advances in neural information processing systems (pp. 2553-2561).
[11] Guanghan Ning, Ping Liu, Xiaochuan Fan, Chi Zhang : A Top-Down Approach to Articulated Human Pose Estimation and Tracking, Springer International Publishing
[12] Miaopeng Li , Zimeng Zhou, Jie Li, Xinguo Liu : Bottom-up Pose Estimation of Multiple Person with Bounding BoxConstraint, 2018 24th International Conference on Pattern Recognition (ICPR), IEEE
[13] Dario Pavllo, Christoph Feichtenhofer, David Grangier, Michael Auli, 3D human pose es- timation in video with temporal convolutions and semi-supervised training, Computer Vision and Pattern Recognition, 29 Mar 2019.
[14] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele,” 2D Human Pose Estimation: New Benchmark and State of the Art Analysis” Max Planck Institute for Informatics, Germany

978-1-6654-0203-3/21/$31.00 ©2021 IEEE