Turkish Journal of Computer and Mathematics Education Vol.12 No.13 (2021), 466-471

Research Article

Spatial Transformation Based 2D Image Segmentation for Single Human Pose Estimation with Improved Efficiency

Madhumitha.R[a], Premi.J.P[b], N.R.Raajan[c]

[a] Student, School of Electronics and Electrical Engineering, SASTRA Deemed To Be University, Thanjavur.
[b] Senior Associate Professor , School of Electronics and Electrical Engineering, SASTRA Deemed To Be University, Thanjavur.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: Human pose evaluation on multi-person is challenging than human pose evaluation on single-person . Even when the human detectors have given their best results, small errors are available in localization part and recognition. These errors in detection lead to failures in human pose estimation. So, we propose an architecture for human pose evaluation although when the errors are present in the human detection. The model consists of two main steps: first step is to detect the human and next step is to evaluate the human pose for the detected human. The components used primarily for human pose estimation is the STN(Spatial Transformer Network), SPPE(Single Person Pose Estimator) and Pose-NMS(Non-Maximum Suppression). This method is tested on the MPII(multi-person) dataset.

Keywords: Spatial Transformer Network, Non-Maximum Suppression, Pose Estimation

1. Introduction

Human pose evaluation is more difficult process in the region of image processing. Practically, to recognize a human in a crowded area is much more difficult when we compare with the detection of single person. There are two ways for human pose evaluation in a crowd, part-based framework and two step framework. In the first one we detect the parts of the human, and with that we estimate the pose of the human. In the second we first draw bounding boxes around the person and detect the person inside the bounding boxes. In the first approach if the persons are very nearer to each other then the pose estimation is very difficult. In the second approach the accuracy mainly depends on the detected bounding boxes. In both approaches the disadvantages and advantages are there. The most occurring issue in the two step framework is that localization error and redundant detection. Redundant detection is the due to the result of redundant bounding boxes.

Incase to solve the above issues, a multi-person pose evaluation model for the region is proposed. This method reduces errors and increases the performance of human pose estimation algorithms. In above method a spatial transformer is used to get the region closer to the human. It is attached before the single person pose estimator and it helps in producing high quality output. Along with the localization problem the redundant detection also leads to inaccurate human pose estimation. So to reduce the redundant detections a parametric pose NMS is introduced. It is attached at the last stage of the process. It uses a distance metric to remove the redundant detections.

This paper introduces something new called pose machines. This pose machine consists of two types of modules. The two modules can be image feature processor module and a prediction module. These devices architecture is with multiple stages where it can be processed from end to end and these machines can be differentiable. To learn rich implicit spatial models a sequential prediction framework is provided. This paper shows that long range spatial relationships can be completed by using huge receptive fields. The receptive fields can be increased. CPM is used here. The receptive fields can be increased by increasing the number of stages of CPM. Stage 1 does not change and greater than stage 2 are only same values of stage 2. It introduces novel framework for the convolutional pose machine. It also uses supervision in between and coming next stages to remove the issue of destroying gradients.

2. Literature Survey

Graph structures, the examples can be pictorial structures [5]-[7] and loopy structures [8]-[10], have usually used to design of the relationships in the spatial domain among the body parts. These models were developed using features such as HOG feature[11], and their efficiency and accuracy is based on the image pyramid. In recent trend, deep architecture have achieved highest level of results in human pose estimation[12]-[16]. Deep Pose[17] is the human pose estimation method first used technique formed on the deep convolutional neural network. It points out all the coordinates of the human, where it is difficult to learn the mapping of image-to-locations. So that designs after that using fully convolutional neural networks calculated the score maps. The part locations are marked as Gaussian peaks in score maps. Higher performance is achieved on image pyramids by using multi-scale testing which was frequently used and given the multi-scale representation. A model with strong

scale changes can be learned by a network having multiple branches implemented.

The visual attention model computationally efficiency is more and the efficiency in understanding images is also more, since it has given a enormous success in result in different works such as machine translation [19], object finding[20], image question answering [21], and human detection. Already existing approaches adopt repetitive neural networks to produce the attention map for an part of image at concurrent step, and collective data taken out in all steps are processed to produce the end result[22].

Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin Cordelia, Schmid Grégory Rogez[1] proposed Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images method in which Large-scale dataset of textured 3D meshes are used. In this paper it can get output of the detailed areas of the human. Nikos Kolotouros, Georgios Pavlakos, Kostas Daniilidis[2] proposed Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. The datasets used are Human3.6M and UP-3D. This method produces output as 3D shape and it returns to the parametric representation of a human. Hüseyin Temiz, Berk Gökberk, Lale Akarun[3] proposed Multi-view Reconstruction of 3D Human Pose with Procrustes Analysis. The datasets used are CMU Panoptic, Human3.6M. It proved that by implementing camera configuration with view from different points increases the 3D pose estimation. Kentaro Sakai, Yoshiaki Yasumura[4] proposed Three-dimensional shape reconstruction from a single image based on feature learning. The dataset used is One 2-D Image from BJUT-3D FaceDatabase. It produces result in the form of Three dimensional output which is the shape of the Human by implementing the Convolutional Neural Network and the MSE has decreased to a less amount when compared to half. Hwasup Lim, Ouk Choi[5] proposed CNN-Based Denoising, Completion, and Prediction of Whole-Body Human-Depth Images. The dataset used is synthetic depth image dataset It produces results in both back-view depth image and and the depth image of the front side in the human. It also refines the the depth image of the front side in the human.

Yasin[6] obtained final 3D pose as output. The main highlight is that projection error is minimized when compared to other papers. The datasets used are HumanEva-I and Human3.6M Kostrikov[7] predicted the joint positions in 3D relative to each other. To predict them, used a method called depth sweep regression forests. This method is trained with the help of three groups of features. The datasets used are Human3.6M and HumanEva-I. Wang[8] proposed a 3D pose by means of two frameworks. First to detect in 2D a 2D part detector is used then a sparse basis representation is used to obtained the output as 3D pose. Zhou[9] proposed Convex formulation in 3D. This is done by implementing the convex relaxation from the orthogonality constraint. The dataset used is CMU MoCap. Radwan[10] employed a 2D part detector and created multiple views. The two dimensional process of part detection is employed with a step which performs the identification of hidden parts. The different views are drawn with the help of twin-GPR performed in a stage-wise manner. Simo-Serra[11] proposed the two dimensional process of part detection. In addition to the detection of parts, a stochastic sampling is implemented in order to deep dive into each region. A set of hypotheses concentrates on the re-projection and drawbacks with the length.

3. Methodology

The Human proposals are send to the SSD detector where it detects the occurence of Human. The human proposals are generated by the single shot detector. Then the generated human proposals are given into the Spatial Tranformer Network(STN).It extracts the region located in close to the human. Then the result of STN is fed through the Single Person Pose Estimator(SPPE) which produces pose estimation of the human present in the localized region. The generated pose estimation is well refined by means of Non-Maximum Suppression(NMS) step at the final output. The parallel SPPE is used mainly to reduce errors during the training of images and regularize the framework.

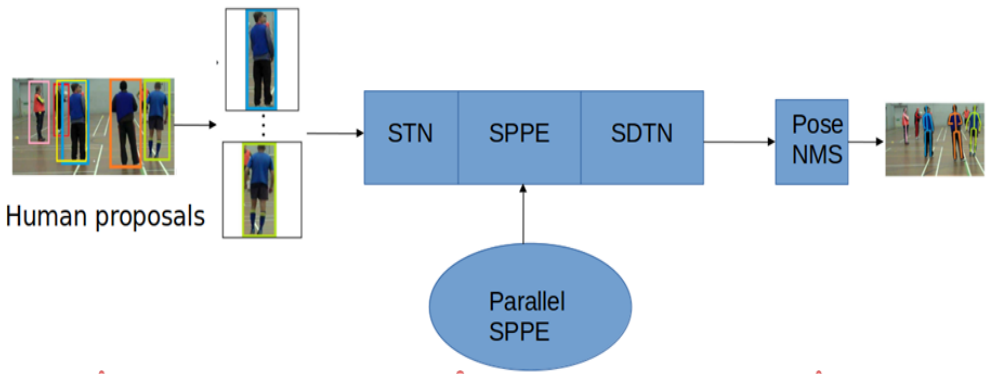

Fig:1 Block Diagram

Human detector

The human detector used here is the Single Shot Detector(SSD). SSD produces group of boxes around the person and score values for different types of objects or different instances present in these boxes. The final step includes Non-Maximum suppression(NMS) step is implemented mainly to reduce the unnecessary features in the image. Additional layers are included at the final layer of the basic structure to produce detections with additional features. The VGG-16 is used as the base network. The input convolutional layer is of size 224*224. The input is given through all the convolutional layers in which the filters were implemented with a receptive field of size 3*3. It uses 1*1 convolutional filters in one of the configurations In addition to the base network some layers are added at the ended of the network. These layers are mainly used to detect at different scales. The layers size decrease in size with decrease in layer number. The layers are provided with the filters and each layers can produce a prediction for each layer. Five extra layers are included at the fianl layer of the base network. In 5 layers 3 layers make 6 predictions and the other 2 layers make 4 predictions The default boxes are produced with different sizes, height and width. The default boxes are produced for each feature map cell. The total results obtained for a m*n feature map can be (c+4)kmn. Where c is the class scores for k at a given point and 4 offsets for each default box. The boundary boxes which are default are same to the anchors, which are implemented in Faster R-CNN. The ground truth boundary boxes are separated into different clusters. Then each different cluster is denoted by means of a default boundary box. The SSD predicted results can be divided into two types as positive matches and negative matches.

If the box is found with the IOU(Intersection Of Union) of greater than 0.5 than it can grouped into positive otherwise it will be negative. Then the human proposals are generated as the output of SSD based on the IOU and scores are stored in the local folder

Spatial Transformer Network(STN)

Spatial Transformer Network gives the feature map as output. It also takes the input as feature map. The output map is obtained as the transformation of the spatial information in the input. If there is multiple channel, and if it have the different inputs the same wrapping technique is applied. This mechanism is divided into 3 parts. The parts can be as follows

Localization network

The input is taken as feature map(U) with dimensions width(w) and height(h). The input have more number of channels c and the output will be given as theta , where θ = floc(U). The θ size depends upon the transformation. The localization function(floc) can take any network as input. It can be either any of the neural network. But the network needs a regression layer at its final to produce the output

Grid generator

The warping operation is performed by means of placing the kernel, which is used for sampling at a certain position in the input map, where it needs to be the centre of the input feature map. Here a 2D affine transformation can be performed to get the required spatial transformation. The transform allows different operations such as cropping, rotation,scaling, translation etc.

Sampler

The differential image sampling is performed here since it allows loss gradients to flow back. The spatial operation is performed when a sampler takes group of sampling points as input. It takes the points with the input feature map(U) and gives output as feature map(V). Any kernel can be used according to the theory. Either the integer sampling kernel or bi-linear sampling kernel can be used to give partial derivatives.

Single Person Pose Estimator(SPPE)

The pose evaluator for single person used in this flow is the stacked hourglass network. The hourglass network is basically a combination of encoder and decoder. The encoding process means to up sample the features and the decoding process refers to down sample the features. The hourglass architecture consists of two main layers . The layers can be the convolutional layer and the max-pooling layer. These layers bring down the image to a low resolution. During every step at max pooling the branches present in the network move forward and gives up added convolution at the resolution which is used before. Finally, reaching the resolution which is very low, it starts with a top-bottom sequence of up sampling and adds up the features of all scales To get the final network predictions, two repetitive rounds of convolutions are done, when we reach the output resolution λ, η of the

network. The highest resolution of hourglass is 64*64 since operating at full input resolution requires high amount of memory. The resolution does not change the efficiency of the network to produce joint predictions.

For the single person pose evaluator a stack of 8 hourglass network is used with intermediate supervision. A single hourglass network mainly consists of residual module, up sampling and the pooling layer. The structure of single hourglass network is repeated for other stacks. The blue block given is the residual block, the green block used is the max pooling layer and the red block given is the up sampling layer. The final results mainly depends on the residual layer. The dimension of filter 3*3 is never used. To do so bottle-necking of the residual layer is used to reduce the dimension and complexity.

In bottleneck architectures 1*1 layer are used to increase and reduce the dimensions and the middle 3*3 layer a bottleneck layer between dimensions present in both input and output. When the network reaches the lowest resolution, it starts with the process of top to down sequence of up sampling and add all features across all scales. To get the information between 2 adjacent resolutions, nearest neighbour up sampling is done with the combination of features along the scales.

The final predictions are obtained by means of two continuous rounds of 1*1 convolutions after reaching the required resultant resolution.

Parallel SPPE

The parallel SPPE(Single Person Pose Estimation) used mainly to back-propagate errors present at the center location. The Spatial Detransformer Network(SDTN) is not connected with the parallel SPPE. While training, the layers of this branch are stopped. The resukts obtained from this branch are focused with the labels of ground truth poses present at the center. The parallel SPPE used here is the 4 stack hourglass network. Only 4 stacks are used considering the memory location.

Spatial Detransformer Network (SDTN)

This network is the reverse process of spatial transformer network. This network computes the gamma to obtain the de-transformation of the network. The grids are formed based on the gamma factor.

$$[\gamma 1, \gamma 2] = (\theta 1, \theta 2) - 1 \quad (1)$$

$$\gamma 3 = -1 * [\gamma 1 \gamma 2] \theta 3 \quad (2)$$

The theta parameters are generated by the localization network in the Spatial Transformer Network(STN) based on the dimension. Here a three dimensional transformation to get the θ_1, θ_2, θ_3 parameters.

Pose NMS

Pose NMS is mainly needed by the network to remove the redundancies. The reference is selected as the most confidence pose. The poses which are similar to the confidence pose are the one which needs to be removed. The repetitive process of NMS is done until the redundancies are removed.

$$[f(P_i, P_j)|\lambda, \eta) = 1[d(P_i, P_j|\lambda, \square) \leq \eta] \quad (3)$$

The elimination criterion is set with the help of threshold λ ,η is the parameter set of the function d(.). The distance function is used to calculate the pose distance. By using the pose NMS only unique poses are reported. To remove the low confidence poses tanh operation is used.

4. Results

This network increases the accuracy of the human pose evaluation by implementing the spatial transformer network. Even though there is error in person detection the pose estimation is accurate by means of this network. Parallel SPPE is included mainly to back-propagate the errors present within the network. Pose NMS is included in the network to remove the redundancies.

Fig: 1 Input Image

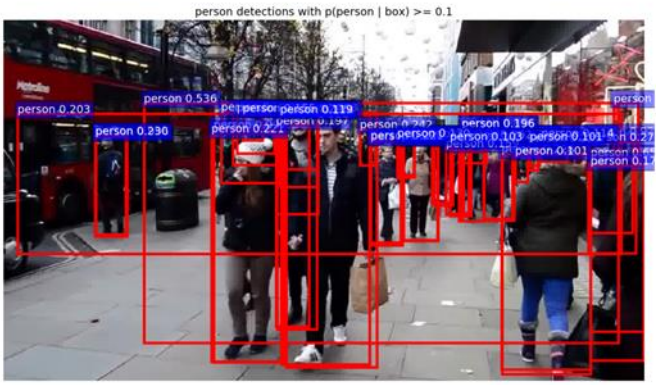

Fig:2 Human Detection

Fig: 3 Zoomed output of pose estimation of human

5. Conclusion

The pose estimation with multiple person is done here by using the Single person pose estimator. This network increases the accuracy of the pose estimator with the help of STN , which makes the network get focused in the localized region . It increases the performance in the two step framework and reduces the error occurred by the human detection network. The redundancies also reduced at the final stage of the network by means of using the Pose-NMS . The future work is to evalute the pose of the person in a video or in a image format where the human is moving or not in a stationary place.

References

1. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision 61(1) (2005)
2. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS. 2006
3. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR. 2009
4. Eichner, M., Ferrari, V.: Better appearance models for pictorial structures. 2009
5. M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. IEEE Transactions on Computers, 22(1):67–92,1973.
6. P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. IJCV, 61(1):55–79, 2005.
7. Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In CVPR, 2011.
8. X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In ICCV, 2005.
9. T.-P. Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In CVPR, 2010.
10. V. Ferrari, M. Mar´ın-Jimenez, and A. Zisserman. 2d human ´ pose estimation in tv shows. In Statistical and Geometrical Approaches to Visual Motion Analysis, pages 128–147. Springer, 2009.
11. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.
12. V. Belagiannis and A. Zisserman. Recurrent human pose estimation. FG, 2017.
13. E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multiperson pose estimation model. In ECCV. Springer, 2016
    A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In ECCV, 2016.
14. S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In CVPR, 2016.
    A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV. Springer, 2016.
15. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In CVPR, 2014.
16. J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In CVPR, 2015.
17. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In ICLR, 2015.
18. J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In ICLR, 2015.
19. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In CVPR, 2015
20. J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In CVPR, 2016.
21. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the