ISSN: 2472-3703 DOI: 10.6911/WSRJ.202001_6(1).0028

Overview of 2D Human Pose Estimation

Ruijun Ma[1, a, *], Jian Zhang[1] and Yuming Qi[1]

[1]Tianjin University of Technology and Education, China

[a]Mrj-01@outlook.com

Abstract

2D human pose estimation is the basic content of human pose estimation, mainly by processing and extracting features in the image, then detecting potential human joint points, and finally clustering and modeling the detected joint points. This paper introduces three methods based on coordinate regression method, heat map-based detection method and regression and detection mixed model, and presents a clear and clear presentation of the current research path of 2D human pose estimation.

Keywords

2d, human pose estimation, CNN, deep learning.

1. INTRODUCTION

The development of human pose estimation has become more and more close to reality. For example, in the fields of gait analysis, human-computer interaction, and video surveillance, human pose estimation has a broad application prospect. The current mainstream 2D human pose estimation algorithms are based on deep learning methods, using neural networks to extract more accurate and robust convolution features than artificial features to predict more complex poses. Human pose estimation methods based on deep learning mainly use convolutional neural networks (CNN)[1]Extracting human pose features from the image can not only get richer semantic information features, but also get multi-scale and multi-type human joint point feature vectors and the full contextuality of features under different receptive fields, and get rid of the component model. Depending on the structural design, coordinate regression is performed on these feature vectors to reflect the current pose, so that the pose information can be applied to specific situations. The main methods are coordinate regression method, heat map detection method and regression and detection mixed model.

2. COORDINATE REGRESSION

The method based on joint point coordinate regression and constructing a label model (Ground Truth) takes the two-dimensional coordinates of the joint points as the labels, and trains the network to directly obtain the coordinates of each joint point. This problem model is referred to asCoordinate Net.

Deep pose [2] it is a method for single person pose estimation based on deep learning. It uses a multi-stage regression approach to design CNNs, and uses coordinates as the optimization target to directly return to the two-dimensional coordinates of human bone joint points. This method obtains the approximate position of the joint point in the initial stage, and then continuously optimizes the coordinates of the joint point in subsequent stages. Before entering the next stage of regression, using the currently obtained joint point coordinate center and cutting small-sized sub-images in its surrounding neighborhood as the input for this stage of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202001_6(1).0028

regression, providing the network with more details of the joint point image to continuously modify Coordinate value. The specific algorithm flowchart is shown in Figure 1.

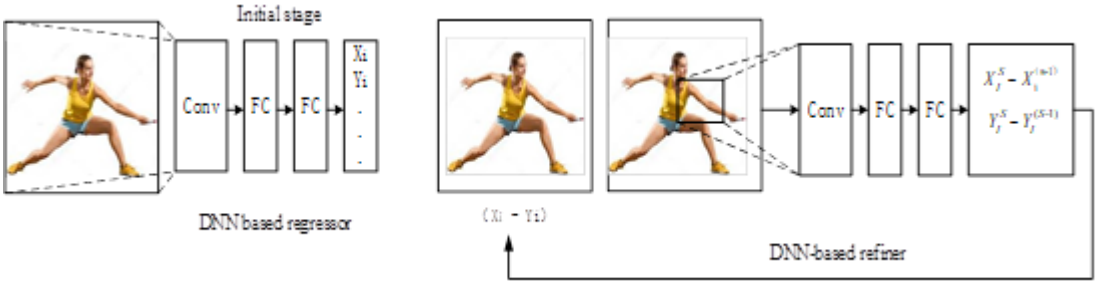

Fig 1. Deep pose structure

The table 1 shows the arrangement of rotary kiln with accessories. Kiln can be used as a rotary dryer to remove water and moisture content from solid substances by introducing hot gases into a drying chamber. Kiln shell should be structurally strong with non-conductor lining and designed to withstand high temperature and prevent the thermal losses of the kiln. Construction and position alignment of the kiln is very important for all the process. In thermal processing of residual materials with a various origin and predominantly for fire treatment of hazardous wastes rotary kiln are employed.

3. METHOD BASED ON HEAT MAP DETECTION

The early deep learning network needed to return is an offset of each key point relative to the picture, and the long distance offset is difficult to return in the actual learning process, the error is large, and the In the process, less supervision information is provided, and the convergence speed of the entire network is slower. Therefore, the researchers made a heat map based (Heatmap) detection model, referred to as Heatmap Net. Heatmap Net not only builds ground truth based on probability distribution, but also adds some structural information between human parts. Heatmap uses a probability map to represent each type of coordinates, giving a probability to each pixel position in the picture, indicating the probability that the point belongs to the key point of the corresponding category, and the closer the probability of the pixel point closer to the position of the key point, the closer 1. The probability of a pixel that is farther away from a key point is closer to 0. A multi-dimensional Gaussian model can be used to convert two-dimensional coordinates into a label in the form of a heat map. To a certain extent, each point provides supervision information and the network can be faster Convergence and prediction of each pixel position at the same time can improve the positioning accuracy of key points. In terms of visualization, the heat map method is also better than the method using only two-dimensional coordinates. The processing of the two methods is shown in Figures 2 and 3 below. Show. The current model based on heat map detection achieves the best results and is also the most widely used. Different ways to learn the structure of joint points according to the network model. Heatmap Net can be divided into two types: explicit addition of structural priors and implicit learning of structural information.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202001_6(1).0028

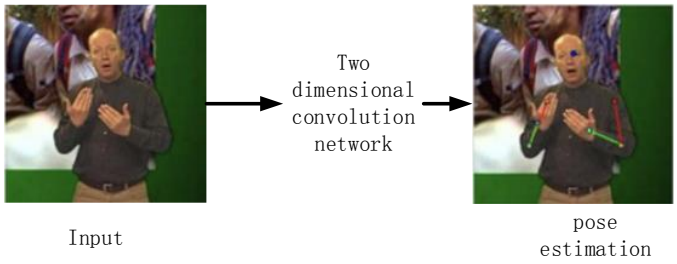

Fig 2. Coordinate Net detection process

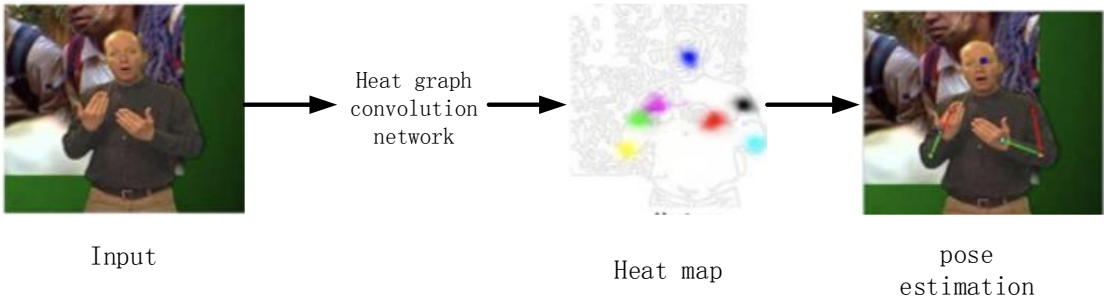

Fig 3. Heatmap Net detection process

3.1. Explicitly Adding Structural Priors

After generating a heatmap for each joint, Heatmap Net can explicitly build the prior network connection relationship of the graph structure or tree structure based on the probability map model and referring to the connectivity of human joint points, so that the network model is available before training The prior information of the human body structure, the training process is equivalent to artificially controlling the flow of information of each joint point in the network, and improving the network's sensitivity to these characteristic information flows. Through the integrated learning of the characteristic information flows, train each joint point Component detector for node components.

Thompson [3] Et al. jointly trained CNN and graph structure models. The CNN of this method uses a multi-resolution mechanism when extracting image features, and integrates local details and global information of joint point features when constructing the heatmap, ensuring the subsequent pixel classification accuracy and coordinate positioning accuracy of the heatmap using the graph structure. Meanwhile, the author uses Markov Random Field (MRF) [4] filter those abnormal nodes. Get the probability distribution of each node variable based on the heatmap, and then use MRF to model the pair-wise node pair consisting of all adjacent joint points, and construct the corresponding network structure to calculate the conditions in each pair-wise The probability distribution allows nodes within a pair-wise to modify the heatmap of neighboring nodes with each other. This effect between nodes is called affinity.

Correspondingly, the network model will finally get the heat-map of each joint point, and all the heatmaps based on conditional probability, that is, the affinity map. Cascade the two heatmaps to get the heatmap modified by the affinity map. The final heatmap actually expresses the joint probability distribution of each pair-wise. Therefore, each pair-wise relationship needs to build 4 sub-network modules as component detectors for training, 2 for training heatmaps, and 2 for training affinity maps. When the overall MRF is trained, redundant pair-wise joint points with smaller joint probability are continuously deleted, and the joint probability distribution of the relevant nodes of the human body is optimized to generate a relatively accurate complete human posture joint point position distribution map. The

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202001_6(1).0028

processing process is shown in the figure as shown in 4, it can be seen that the MRF model based on the pair-wise relationship requires a large number of component detectors for training, which makes the network structure more complicated.

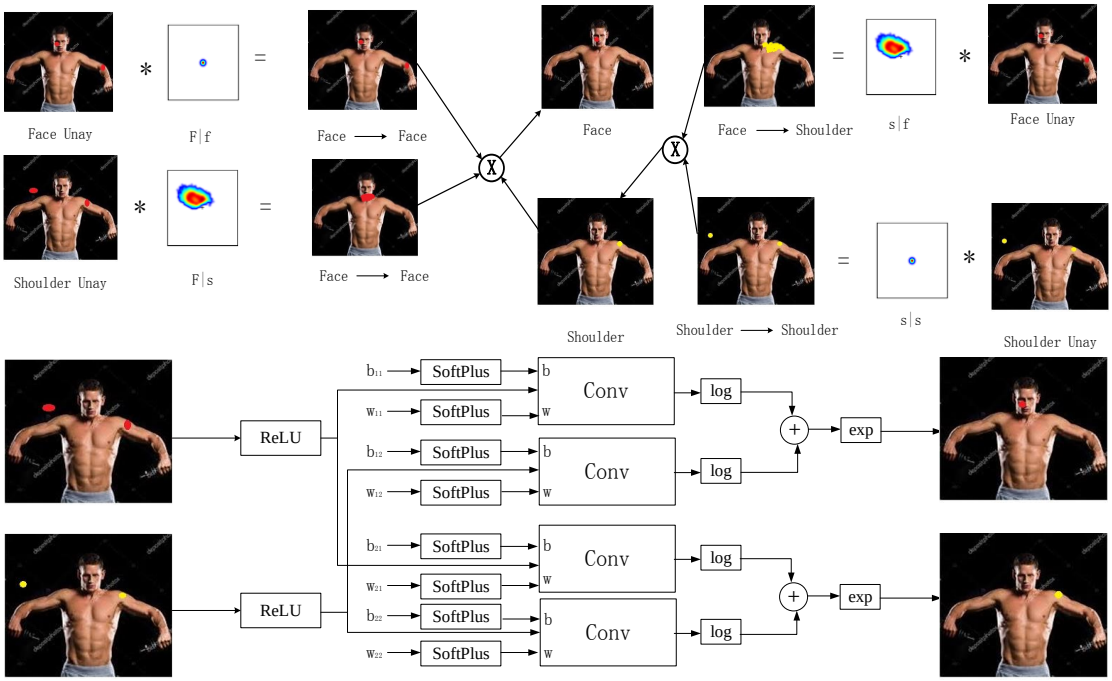

Fig 4. Structure of Tompson et al.

3.2. Implicit Learning Structure Information

At present, most methods are mainly based on the large receptive field mechanism to learn body structure information implicitly. The receptive field is defined as the size of the area mapped by each pixel on the original image in the feature map output by each layer of the CNN. The larger the value of the receptive field, the larger the range of features of the original image that the network can learn. It also means that it may contain more global and higher semantic features. On the contrary, the smaller the value, the more features it contains tends to be local and detailed. Therefore, by expanding the receptive field of the heatmap, the network can learn the distant joint point connection features and obtain the joint point structure information with a higher semantic level.

The method of implicitly learning structural information based on large receptive fields has strong generalization and robustness, and many methods are improved based on this. There are roughly three ways to increase the receptive field: expanding the pool layer, increasing the convolution kernel, and accumulating the convolution layer. However, all three methods have their own design flaws: excessively large pool operations will sacrifice accuracy, and if the accuracy is restored by deconvolution, a large amount of additional information will be added; increasing the convolution kernel is equivalent to increasing the amount of parameters, which is too expensive for computing resources; The cost of continually accumulating convolutional layers is the disappearance of gradients. The current mainstream network structure is based on these three methods, and is based on solving the side effects of increasing the receptive field.

Classic Convolutional Pose Machines (CPM) [5] Borrowed from the multi-stage convolution structure in Coordinate Net, a multi-stage cascaded deep network is constructed, and a 11 × 11 large convolution kernel is used to continuously accumulate convolution operations, so that each sub-network follows the order from front to back. Obtaining different receptive fields from small to large, and finally achieve the purpose of implicitly learning the structure information

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202001_6(1).0028

of human joints using contextual on the image. At the same time, in order to promote pixel level fusion and gradient conduction of multi-scale receptive fields, CPM uses full convolution (FCN)[6]The basic network structure is shown in Figure 5.

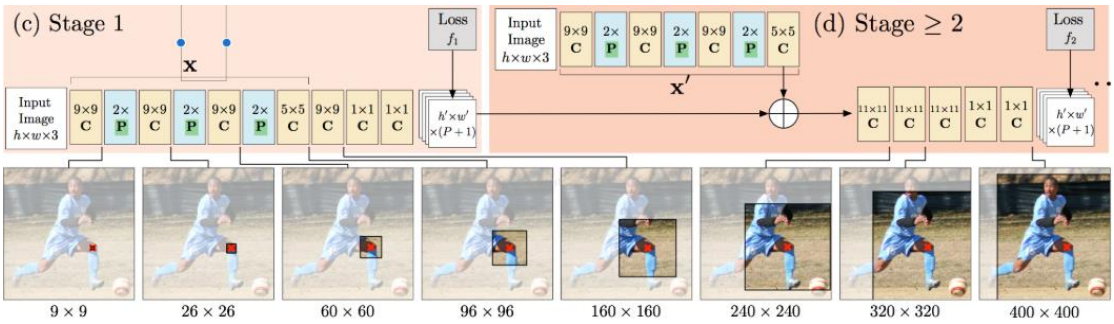

Fig 5. CPM Structure

In the initial stage (stage 1), CPM only convolves the input picture and outputs the heatmap of the relevant nodes. In each subsequent stage, CPM first designed a feature extractor (FeatureExtractor) to cascade the heatmap output from the previous stage and the feature map of the original image, and then input this preprocessed fusion feature map into this stage. The FCN is processed and finally a new joint point heatmap is obtained. However, changing the receptive field by continuously increasing the convolutional layer will cause a large training burden on the network, causing problems such as gradient disappearance. In order to avoid increasing the side effects of receptive fields, CPM uses relay supervised training, which adds up the heatmap generated in each stage and the errors generated by Ground Truth as the total error and iterates, and reverses the gradient from the output layer of the network in each stage. Propagate, avoid the disappearance of the gradient, and finally get a revised feature map (ie, a confidence map) of the response at each stage. The CPM is based on the response graph output of the last stage during testing.

4. HYBRID MODEL OF REGRESSION AND DETECTION

The hybrid model of regression and detection combines the characteristics of the above two, to construct a combined Ground Truth, and to perform detection and regression tasks at the same time. The network model contains the substructures of Coordinate Net and Heatmap Net.

Bulat [7] Et al. proposed a tandem structure by performing Coordinate regression on joint point heatmaps. As shown in Figure6, the network as a whole is divided into a component detector and a regression network. The problem of occlusion of joint points is better solved through the regression of heatmap. The author considers that the response value of the heatmap of the occluded part is relatively low. After the regression module, these low-confidence heatmaps have less impact on the subsequent coordinate correction. Therefore, the tandem regression network module can learn more joint points after the component detector the interdependent semantic information, that is, the structural information of other parts is used to correctly predict the positions of occluded joint points. In the regression module, convolution kernels of different sizes are also designed and an hour-glass regression method is tried.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202001_6(1).0028

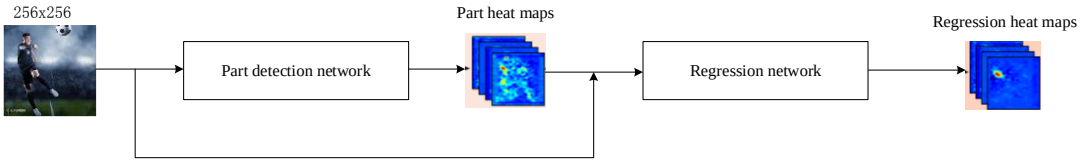

Fig 6. Bulat et al.'S tandem structure network

Fan et al[8]The proposed dual source convolutional network (DS-CNN) builds two parallel network modules, namely joint detec-tion and joint localizition. The specific structure is shown in Figure7. Joint detection is used to detect the local joint point category information contained in the image patch (PartPatch); the joint localiztion module returns the joint point position coordinates by combining the binary mask of the entire image (Body Patch) and the Part Patch. The two sub-networks perform interactive auxiliary training. For example, the joint detection module uses the global features of the Body Patch to determine the left and right attributes of the wrist joint. At the same time, the joint localiztion module normalizes the position coordinates based on the local information of the Part Patch.

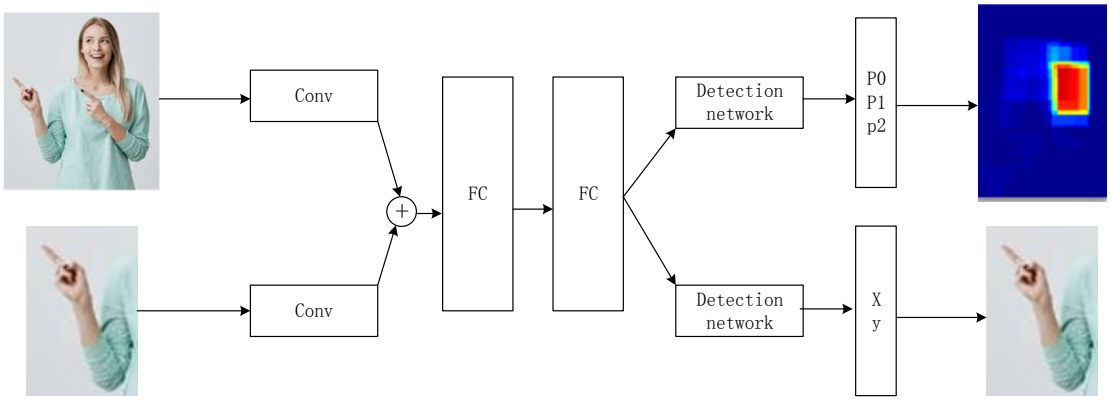

Fig 7. DS-CNN network structure

Heatmap + Coordinate's composite network model combines the advantages of regression and detection models, but from the perspective of feature extraction, the improvement of the composite model is essentially derived from the supervised nature of local joint features to global human features. The model usually uses the regression network module to process the features of the entire image, but a large number of background features will interfere with the processing of human pose features. At this time, the detection module of this method can use local joint features with less noise to emphasize human pose features, and then weaken the effects of background features.

5. CONCLUSION

The 2D human pose estimation method based on deep learning is essentially using CNN to detect the 2D coordinates of each joint point of the human body from the image. In order to get better training results, it is necessary to design a variety of training tricks and use a large amount of data to train the network. The 2D human pose dataset often used in current research is MPII [9], MSCOCO [10] And AI Challenger [11] Wait. However, due to the flexible and diverse methods of human pose estimation, in order to obtain a highly functional and effective network model, it is necessary to construct specialized ground poses for human networks that implement different functions. Specific methods mostly use combined Ground Truth, such as

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202001_6(1).0028

Heatmap + Coordinate, Heatmap + Part Mask, Heatmap + PAFs, etc. In terms of network structure design, some methods use a multi-branch cascade and multi-stage network framework to improve detection accuracy. They can also be designed as stacked hourglass structures to obtain image information of receptive fields of different scales and different resolutions. When performing multi-person pose estimation, target detection algorithms, instance segmentation algorithms, and clustering algorithms can be used to distinguish different human bodies, and a single person pose estimation method is used to obtain the final pose of each human body.

ACKNOWLEDGEMENTS

This paper was supported by Graduate Innovation Fund Project (YC19-13).

REFERENCES

[1] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition [J]. Neural computation, 1989, 1(4): 541-551.

[2] Girshick R.Fast R- CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[3] Tompson J J, Jain A, LeCun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[C]//Advances in neural information processing systems. 2014: 1799-1807.

[4] Li S Z. Markov random field modeling in image analysis [M]. Springer Science & Business Media, 2009.

[5] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.

[6] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[7] Bulat A, Tzimiropoulos G. Human pose estimation via convolutional part heatmap regression [C]//European Conference on Computer Vision. Springer, Cham, 2016: 717-732.

[8] Fan X, Zheng K, Lin Y, et al. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1347-1355.

[9] Andriluka M, Pishchulin L, Gehler P, et al.2D human pose estimation: New benchmark and state of the art analysis[C]//Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2014: 3686-3693.

[10] Lin T Y, Maire M, Belongie S, et al.Microsoft coco: Common objects in context[C]//Proceedings of the European Conference on Computer Vision, 2014: 740-755.

[11] Wu J, Zheng H, Zhao B, et al.Ai challenger: A large-scale dataset for going deeper in image understanding [J].arXiv: 1711.06475, 2017.