Human Pose Tracking based on Cascaded Pose Regression

Dengwei lv

National University of Defense Technology, Changsha, China

dengweilv10@163.com

Keywords: Deep learning, pose estimation, CPR, contextual information, track

Abstract: Human pose tracking is the first-step for videos in social and scientific applications. In this paper, we propose a method for human pose tracking based on Deep Neural Networks (DNNs) using Cascaded Pose Regression (CPR) framework and contextual information. We first introduce a cascade of DNN-based regressors to obtain precision human pose. Moreover, a context-based pose tracking strategy is proposed to improve the tracking rate. We analyze the performance of the proposed method with detailed evaluation metrics and challenging dataset, and obtain comparable or better performance to the state-of-the-art.

1. Introduction

The detection of keypoints in human bones is one of the basic algorithms in computer vision, which is essential for describing human posture and predicting human behavior. With the development of deep learning technology, the detection of keypoints of human bones has been continuously improved, and it has been widely used in related fields of computer vision, such as behavior recognition, character tracking, gait recognition and other related fields. However, the human body is quite flexible, and various postures and shapes vary greatly. Small changes in any part of the human body will produce a new posture, and the visibility of the keypoints is greatly affected by wearing, posture, and viewing angle [12, 15, and 16]. Moreover, it is also affected by the environment, such as occlusions [17, 18], light, fog, etc. In addition, the keypoints of 2D human body and key points of 3D human body will be visually different, and different parts of the body will have visual shortening effect, making the human body Bone keypoints detection become a challenging topic in the field of computer vision.

In recent years, researchers have proposed a number of human body posture detection programs. Among them, the 2D multi-person pose detection scheme can be mainly divided into top-down and bottom-up detection methods. The top-down human body posture detection method first detects all the people in the picture through the target detection algorithm, and uses the single-person attitude estimation method for the person in the detection frame, so that the posture information of all the people in the picture can be obtained, and the main algorithm has RMPE [19], Mask R-CNN [18]. The main disadvantage is that the demand for target detection algorithm is relatively high. If the person in the picture cannot be detected correctly, the human body posture detection cannot be achieved [14]. In addition, the calculation time of the algorithm is affected by the number of people in

Dengwei Lv, National University of Defense Technology, Changsha, China

The picture, and the more people, the more time it takes. The bottom-up human body pose detection method first detects all the keypoints of the human body in the picture, and then clusters all the key points into different human skeletons based on the related information between the keypoints through a certain correlation strategy, such as PAFs [14]. The main disadvantage is that the global information of the image is not well utilized.

Human body pose tracking applications, such as human-computer interaction, intelligent monitoring systems, motion capture, sports science and entertainment facilities [1, 2, 3, 13], have been hot research fields in recent years. Early dynamic models, such as the Markov model, were smooth, but they could not acquire the nonlinear features of the human body pose in the image, and

the tracking performance on the monocular camera video 3D human body posture was not good [4]. At present, according to different ways of defining the pose parameters, the existing methods can be roughly divided into three categories: a method based on the human skeleton, a method based on high and low dimensions, and an integrated method. Based on the human skeleton method, the human body posture is parameterized into a low-dimensional pose parameter space, but this method also requires a human body shape. The method based on high dimensional space can restore 3D human posture without action restriction, and the restoration of 3D human posture based on low dimensional space method is limited by training set [2]. The integrated approach is based on the advantages of the first two methods to implement human body pose tracking, such as [5].

This paper proposes a 2D human body pose tracking method based on cascaded pose regression (CPR). The CPR method in [5] has achieved many achievements in the fields of object pose estimation. The method first randomly generates an initial posture of the human body, and obtains the pose increments of the corresponding stage through the regressor of each stage, thereby optimizing the posture of the previous stage and completing the estimation of the human body posture. In order to improve the corresponding computational efficiency, this paper crops patches around the initial pose locations and inputs it into the regression network. In the subsequent stages, the same method is used to obtain the regression network input. In addition, based on the characteristics that human motion changes little between successive frames of a video, this paper adopts the method that initialize the human pose of the current frame based on the previous stage pose, which reduces the time used for initial posture search in each stage and improves the real-time performance of pose tracking.

The remainder of this paper is organized as follows: in the next section, we generally describe the CPR method, and then present our context-based CPR human 2D pose tracking method. In section 3, experimental settings and results are provided. Finally we conclude the paper in section 4.

2. Method

2.1 Cascaded Pose Regression

In order to express human pose, we often encode the locations of human body joints as a pose vector, and define the pose vector as $P = (p_1, ..., p_j, ... p_J) \in R^{2J}$, where p_j is a two-dimensional coordinate of the j th body joints. CPR framework consists of K stages, the method firstly obtains an initial pose and refines the pose from course to fine. Each stage refines the pose by producing an increment, and then the increment is add up to the current pose, that is,

$$P_k = P_{k-1} \circ \Delta P \tag{1}$$

The increment ΔP is generated by the stage regressor, which takes previous human pose P_{k-1} and the image feature I as inputs, that is,

$$\Delta P = R^k (P_{k-1}, I) \tag{2}$$

The main difference from the previous boosted regression approaches is that the CPR framework takes pose-indexed features as inputs [6, 7], therefore the output of features is related to the image data and the current pose estimation. As demonstrated in [6], the weak invariance assumption justifies the derivations and leads to strong convergence rates for the CPR method, shows an effective and practical performance.

The CPR method is proved to have several advantages in general pose estimation. Firstly, the pose-indexed feature re-calculation is practical in use. Secondly, the number of joints of the human pose has little impact on the testing efficiency [8]. Finally, CPR can be easily applied to data-lacking area and is prove to have effective performance in practice.

2.2 Context-based Pose Tracking

Different from classic methods that use the whole frame to predict the current pose in videos, we propose a context-based pose tracking approach, which is aimed to decrease the computational cost. It is known that the human pose is unlikely to change dramatically between adjacent frames. Therefore, we only need to predict the pose accurately for the first frame of the video. Then we take the accurate pose of the previous frame as a reference, more specifically the initial pose of the current frame. In this way, we can easily get the initial pose of each stage and extract the local patches around the J joints of the pose as the inputs of the network. Moreover, as the pose between adjacent frames changes a little, the tracking computational cost will decrease a lot in stage k.

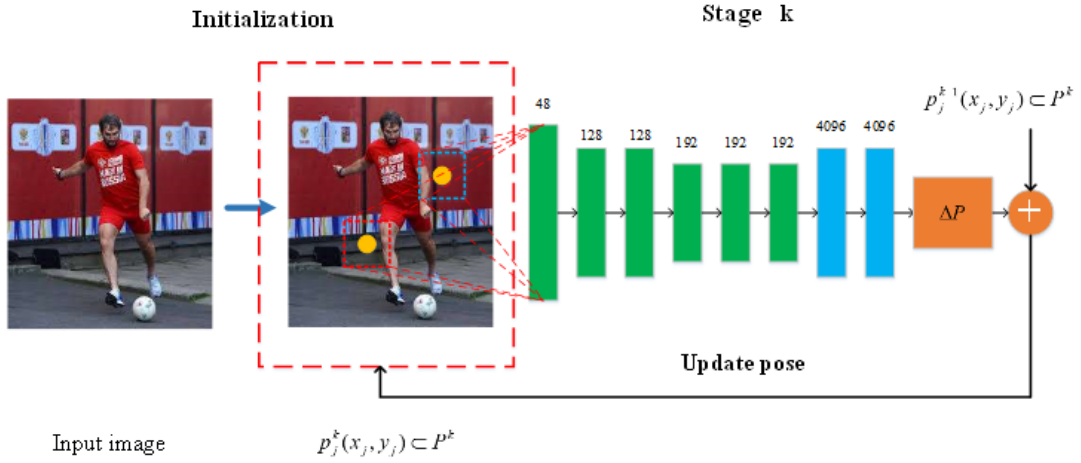

Figure 1. Architecture of the proposed cascaded pose regression

2.3 The Pose Tracking Framework

Based on the above methods, we propose the pose tracking method. The proposed pose tracking framework is shown in Fig. 1, which is inspired by DeepPose [9]. DeepPose is proposed to address human pose estimation, and it is the first time that DNN-based regressors were used for human body joints location regression.

The methods firstly obtain a frame from the object video, and automatically get the average pose of the human in the image. Then in the stage k, the patches around the initial pose joints are extracted and fed into the DNN-based network. In these blocks, six convolution layers are used to extract deep features of the image. In the first three convolution layers, a 11×11 and a 5×5 kernel are convolved with the input of the first three layers respectively, and for the rest three is 3×3.Both convolution and fully connected layers are followed by a Rectified Linear Unit (ReLU) [10] to generate the output.

Training The aim of the training procedure is to train a cascaded pose regressor R^k, where k represents for stage k. Here we assume that the training sets as $(I_i, P_i)_{i=1,\ldots,N}$, with I_i the training image and P_i the Ground Truths (GTs). Each stage the R^k is trained to predict the increment vector between GTs P and the current pose P_k, and the goal is to improve the error as follows:

$$\Delta P = \sum_{i=1}^{N} d(P_i^K, P_i) \tag{3}$$

Where P_i^K the final outputting is pose of the network, and P_i represents the GTs of the image i.

Each stage we start by generating the pose-indexed feature f_i^k and the current pose can be calculated as follows,

$$P_i^k = P_i^{k-1} \circ R^k(f_i^k(P_i^{k-1}, I_i)) \tag{4}$$

Therefore, we train the regressor R^k to minimize the loss in (3), that is,

$$R^k = \arg\min_R \{\sum_i d(R(f_i^k), \Delta P_i^k)\} \quad (5)$$

Based on the previous analyze, we first set the initial learning rate at 0.0001. As the proposed model parameter is large, proper data augment is utilized in our experiments. The dropout [27] regularization for the fully connected layers is set to 0.6. The training procedure completed in 400 epochs with the network converged. The trained parameters are then used for the human pose prediction

3. Experimental Results

3.1 Dataset

As human keypoints estimation has been a hot topic in computer vision recently, a lot of datasets are released for researchers to train the network and test the performance of their methods. Typical benchmarks, such as LSP [21], FLIC [20], MPII [23], and MS COCO [24], are listed in Table 1. LSP represents for LSP dataset and its extension [22], and MS COCO dataset is the 2014 version.

Due to facility restrictions, we choose MPII as the dataset for our experiments. MPII dataset contains almost 28000 training and 11000 test images with 40000 people annotated body joints, respectively. In order to establish the state-of-the-art human pose estimation benchmark, researchers systematically collect images from YouTube videos that covers a wide variety of every day human activities. The introduced dataset of images cover challenging conditions, such as occlusions, clothing types, and variability of imaging conditions. Rich annotations are provided for the images and each image is defined with 16 body joints.

Table 1. Typical human keypoint benchmarks

Benchmark	Training Sets	Test Sets
LSP	11000	1000
FLIC	4000	1000
MPII	28000	11000
MS COCO	82783	40775

3.2 Evaluation Metrics

Once the human pose is estimated, it is compared with the corresponding GTs. The evaluation metrics vary from benchmarks. The generally adopted metrics Percentage of Correct Parts (PCP) introduced by Ferrari at al. [16] measures the correctly localized body parts. More specifically, a body part is considered to be predicted correctly only if the estimated body part segment endpoints are within 50 percent of the GTs annotated segment locations. However, PCP is sensitive to the amount of the foreshortening of the limbs. To improve the previous PCP metrics, Yang at al. [25] proposed PCK metrics, which measures the accuracy of the localizations of the body parts. The PCK metrics use the fraction of the person bounding box size as the threshold for the corresponding of the estimated body part and the GTs [11]. Toshev at al. [9] introduced PDJ metrics, which calculates distance between the estimated joints and the GTs and the joints are considered to be correct if the distance is within a certain percent of the torso diameter. Andriluka at al. [26] compared the performance of the leading human pose estimation approaches on the benchmark with PCP, PCK, and PCKh evaluation metrics. Finally, we choose the mean Average Precision (mAP) of the joints based on PCK threshold to measure the performance of our methods and frames per second (fps) to evaluate the performance of the human pose tracking strategy.

3.3 Numerical and Visual Results

The experimental results on MPII benchmark for the proposed method and approach at al. [14] with the Percentage of Correct Parts (PCP) at 0.5. The runtime analyze is implemented on PC with one NVIDIA GeForce GTX-1060 GPU, and the operation system is Ubuntu 18.04.The experimental results of the proposed human pose estimation and tracking method on the test benchmark are demonstrated in Table 2. As shown in the table, the proposed method inference rate is up to 45 fps, while method in [14] is around 15 fps. This indicates that the proposed strategy improves the tracking efficiency. However, the mAP for the proposed method is 12.3% lower than the method at al. [14]. Some visualization of pose results on images from MPII benchmark are presented in Figure. 2.

Table 2. Comparison of different methods

Method	Inference Rate (fps)	mAP (%)
Proposed Method	**45**	62.3
Cao at al. [14]	15	**74.6**

Figure. 2 Visualization of human pose estimation results

4. Conclusion

In this paper a DNN-based cascaded pose regression framework using contextual information method is introduced to track human pose. We demonstrate the efficiency of the context-based CPR framework. We obtain comparable or closer performance of human pose estimation to the state-of-the-art, and achieve better tracking rate using contextual information. In our future work, we will focus on improving the estimation precision of the model and investigate suitable human pose initializing strategy to achieve robust prediction.

Acknowledgements

The corresponding anthor’s email: dengweilv10@163.com

References

[1] Qifei Wang, Gregorij Kurillo, Ferda Ofli,and Ruzena Bajcsy,“Evaluation of pose tracking accuracy in the first and second generations of microsoft kinect,” in Healthcare Informatics (ICHI), 2015 International Conference on. IEEE, 2015, pp.380–389.

[2] Y. Xu, J. Cui, H. Zhao, and H. Zha, "Tracking generic human motion via fusion of low-and high-dimensional approaches." IEEE transactions on systems, man, and cybernetics: systems 43.4 (2013): 996-1002.

[3] Huang, Chun-Hao, Edmond Boyer, and Slobodan Ilic, "Robust human body shape and pose tracking." 2013 International Conference on 3D Vision-3DV 2013. IEEE, 2013, pp. 287-294.

[4] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton, "Dynamical binary latent variable models for 3d human pose tracking." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010, pp. 631-638.

[5] M. Straka, S. Hauswiesner, M. Ruther, and H. Bischof. ¨Simultaneous shape and pose adaption of articulated modelsusing linear optimization. In ECCV, pp. 724–737. Springer, 2012.

[6] Dollár, Piotr, Peter Welinder, and Pietro Perona, "Cascaded pose regression." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010, pp. 1078-1085.

[7] Fleuret, Franois, and Donald Geman. "Stationary features and cat detection." Journal of Machine Learning Research 9.Nov (2008): 2549-2578.

[8] Yang, Heng, Changqing Zou, and Ioannis Patras., "Cascade of forests for face alignment." IET Computer Vision 9.3 (2014): 321-330.

[9] Toshev, Alexander, and Christian Szegedy. "Deeppose: Human pose estimation via deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, pp. 1653-1660.

[10] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." Proceedings of the 27th international conference on machine learning (ICML-10). 2010, pp. 807-814.

[11] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 4929-4937.

[12] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation." Advances in neural information processing systems. 2014, pp. 1799-1807.

[13] Lee, Mun Wai, and Ram Nevatia, "Body part detection for human pose estimation and tracking." 2007 IEEE Workshop on Motion and Video Computing (WMVC'07). IEEE, 2007.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 7291-7299.

[15] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang, "Multi-context attention for human pose estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 1831-1840.

[16] Ferrari, Vittorio, Manuel Marin-Jimenez, and Andrew Zisserman, "Progressive search space reduction for human pose estimation." 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008.

[17] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, C. Bregler, " Learning human pose estimation features with convolutional networks", in Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[18] He K, Gkioxari G, Dollár P, “Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017, pp. 2961-2969.

[19] Fang H S, Xie S, Tai Y W, "Rmpe: Regional multi-person pose estimation." Proceedings of the IEEE International Conference on Computer Vision. 2017, pp.2334-2343.

[20] Sapp, Ben, and Ben Taskar, "Modec: Multimodal decomposable models for human pose estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013, pp. 3674-3681.

[21] Johnson, Sam, and Mark Everingham, "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation." BMVC. Vol. 2. No. 4. 2010.

[22] Johnson, Sam, and Mark Everingham, "Learning effective human pose estimation from inaccurate annotation." CVPR 2011. IEEE, 2011, pp. 1465-1472.

[23] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis." Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014, pp. 3686-3693.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014, pp. 740–755.

[25] Yang, Yi, and Deva Ramanan, "Articulated human detection with flexible mixtures of parts." IEEE transactions on pattern analysis and machine intelligence 35.12 (2013): 2878-2890.

[26] Ferrari, Vittorio, Manuel Marin-Jimenez, and Andrew Zisserman, "Progressive search space reduction for human pose estimation." 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008, pp. 1-8.

[27] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis." Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014, pp. 3686-3693.