IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of current version December 20,2019.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

3D Human Pose Estimation

MD JAMIL, UR RAHMAN[1], HARSHAL, SONI[2], ARSHAM, SARSHOGH[3], AND SHUYU, ZHAO.[4]
[1]University of Alberta, Edmonton,Alberta, Canada (e-mail: mdjamilu@ualberta.ca)
[2]University of Alberta, Edmonton,Alberta Canada (e-mail: hsoni@ualberta.ca)
[3]University of Alberta, Edmonton,Alberta Canada (e-mail: sarshogh@ualberta.ca)
[4]University of Alberta, Edmonton,Alberta Canada (e-mail: szhao5@ualberta.ca)

"This work was supported in part by the Department of Computing Science- Multimedia Research Lab at University of Alberta"

ABSTRACT The proliferation of computer vision techniques has kindled multiple applications that facilitate computers to detect humans as well as provide smart services. Estimating 3D spatial coordinates of skeleton in real time with reasonable latency using a single RGB image is the motivation behind this work. We are considering image processing as well as multi-layer neural network based approach as the modus operandi. The image processing method titled "Multi- Person human pose estimation over inaccurate bounding boxing" has four major components in its architecture [1]: Symmetric Spatial Transformer Network (SSTN), Parametric Pose Non-Maximum-Suppression (NMS), Pose-Guided Proposals Generator (PGPG), and Non-parametric Pose Lifter. This method is able to handle inaccurate bounding boxes and redundant detection, allowing it to achieve 80.7 mAP on the MPII (multi person) data-set [2]. The multi-layer neural network approach is implemented in an unique way to avoid the extensive training time by taking 2D points as an input rather than raw images and perform end-to-end learning to 3D pose points. Neural network based 2D pose estimators have an exceptionally well accuracy and 2D points are easily derived for our network from it. This facilitates us to train entire Human 3.6M data-set [3] on a single GPU because of the low- dimentionality.

INDEX TERMS HPE-Human pose estimation, SSTN -Symmetric Spatial Transformer Network, NMS-Parametric Pose Non-Maximum-Suppression , PGPG - Pose-Guided Proposals Generator, hourglass model.

I. INTRODUCTION

HUMAN pose estimation- or the problem of localizing anatomical key-points or "parts" has been one of the most popular and widely engaging computer vision topics, acclaimed for its multiplicity of applications ranging from action recognition, surveillance, human-robot interaction and motion detection [4]. Subscribing to the potentials, there exists different approaches that often use video streams, multi-view cameras and depth images. In practise, usage of specialised equipment, such as wearable motion sensors, depth cameras (e.g. MicrosoftKinect), or marker-based motion capture systems are standardized. Since these equipment are not economical enough to be used in real-life scenarios, estimation of human poses using images or videos taken from standardize cameras has caught up the heat.

In practise, image processing and machine learning are the two complementary techniques used by researchers around the globe to solve the the problem of single human pose estimation and multi-human pose estimation. Additionally, predicting the poses in the wild using monocular RGB images is even more challenging than controlled lab environment poses.

The devoid of depth data in the RBG image is an hurdle to infer the three-dimensional pose proving out to be an under-constrained problem. However, humans have been successful to manually estimate such 3D poses from the single image using the prior knowledge of depth cues, spatial relations between body parts, length of the limbs et cetera, transmuting the same concept explicitly into a hand-crafted algorithm is a great impediment.

The modern techniques approaches this hurdles by using one of the two methods [1] : a part based architecture or a two-step architecture. The part-based frame-work [5] [6] first detects all possible body parts independently and then assembles the detected body parts to form multiple human poses. The two-step framework [6], [7] first estimates all the possible human bounding boxes and then estimates the pose within each box independently. In the second approach, assembled human poses can be inappropriate if humans are organized in a tight crowd and the quality of bounding boxes can be the bottleneck of first approach.

On the contrary, deep learning approaches are able to

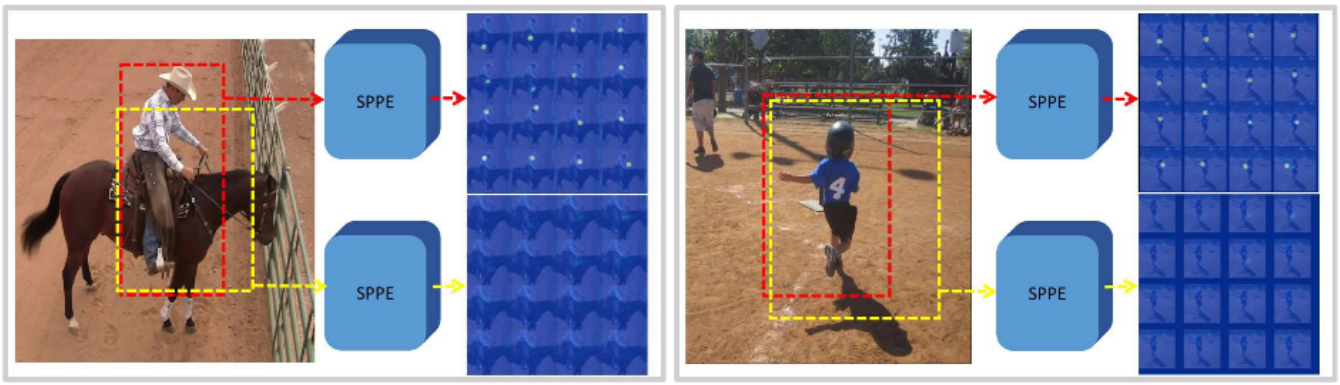


FIGURE 1. The localization errors and redundant bounding box error can be generated by 2D pose detectors that can deteriorate 3D poses in later stages

implicitly learn the spatial mapping of bones in the skeleton and develop a relationship between input and output [4]. The robustness of deep neural networks(CNNs) to extrapolate the patterns and relationship between the training data have proved to be highly tangible to solve the undermining problem of human pose estimation. The corroboration by the results for popular benchmark data-sets such as MPII Human Pose [2] and Human3.6M [3] can justify the success of deep learning techniques.

The recent Deep learning based methods either uses volumetric heat-maps or fully connected end-to-end regressors to form joint location predictions. However, heat-map based methods used for 3D pose estimation are too computationally expensive and requires quite a memory to perform efficiently and the fully connected layers based approach can fail to learn the spatial correlation which can fail miserably in the wild data values.

In this work, we have addressed both the field of research with two techniques. Our approaches encapsulate the best of what the field has to deliver and tries to address the challenges mentioned above. The first approach is based on Image processing paradigm to extract key-points of 2D human skeleton joints from RGB image and build 3D poses using trainable ResNet backbone. Our second approach is based on deep learning paradigms and it attempts to predict 3D human poses based on multi-layer neural network.

II. RELATED WORK

A. 2D HUMAN POSE ESTIMATION

It has been proven over the course of years that deep neural networks are extremely successful at 2D pose estimations.With DeepPose [8] being the oldest one to use a convolutional network to predict the coordinates of the joints of human body.It also uses a refinement algorithm to improve the predicted candidates.The current state-of-the art technique Stacked Hourglass [9] also utilizes this concept of cascaded refinement along with residual connections [10] with semi- supervising windows. The high robustness and accuracy of Stacked Hourglass model as justified by testings done on the MPII Human Pose dataset —has motivated us to build one of our 3D pose estimator (Approach 2) on top of it as done by others as well.[11], [12], [12], [13]

However, the extensive popularity of neural network based techniques has taken the attention on more accurate heat-map matching approach [14]. Heat-map based approach functions by creating high intensity pixel locations on the image called heat-maps and training the model to learn these spatial locations. To infer the 2D joint locations, cor-ordinates are calculated using a nondifferentiable argmax operation. Along with this, some techniques also uses the neighboring pixel values to make fine adjustments.[9]. Recently, soft-argmax based technique [13] has proven to be a viable alternative to argmax based techniques because it is easier to backpropogate and make refinements to predicted poses.

B. 3D HUMAN POSE ESTIMATION

3D human pose estimations are more complex in comparison to 2D human pose estimation because not only the model needs to learn to estimate pose in xy plane but also in yz and zx planes. Hence, it is difficult to train a network to regress candidate poses in three planes as well as it is memory extensive in case of heat-maps.A single plane heat-map cannot encompass enough depth related information that predictions for other planes can be made. Roughly speaking, following the few techniques used by researchers to estimate human poses in 3D:

1) Fully connected output: This is the most simple approach that can be formulated for 3D pose estimations [15], [16], [17]. The network is made end-to-end fully connected and it is trained. The biggest downsize of this approach is that it is expensive to train and it cannot be generalized to unseen data-sets. It is therefore necessary not to rely on this approach and explore further.
2) Volumetric heatmaps: To overcome the downsizes of the first class of techniques, the concept of heat-maps can be applied and extended to other two dimensions. However, memory requirement to store and process these heat-maps can be problematic especially

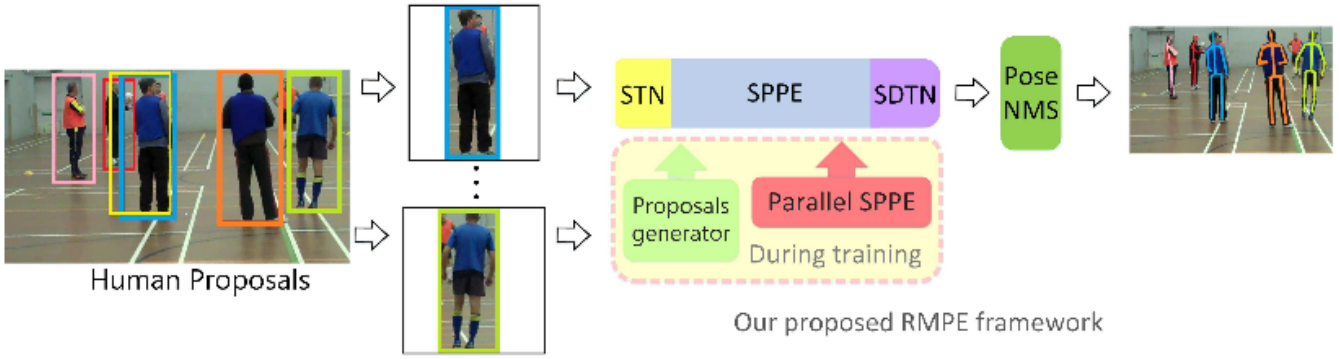


FIGURE 2. Architecture of 2D pose estimator based on [1]

with argmax operations to calculate the final poses. Pavlakos et al. [18] used coarse-to-fine build up technique to include the depth into network in an attempt to reduce memory usage. Luvizon et al. [19] use the soft-argmax operation to calculate coordinates from volumetric heatmaps.

3) Location maps: Mehta et al. [20] has been the precursor of"location maps" where every value in the map is a pixel estimate of spatial co-ordinate value of the skeleton. The location-maps are 2D representation and for the z dimension, it is similar to depth maps. This makes it memory economic to be processed. Tragically, not much of the research is being carried out here and performance on Human 3.6M data-set is not up to mark and the field is open to research.

Our multi-layer neural network approach is inspired from [16] works in real-time because of low-dimensional input of 2D skeleton points and triangulate them into 3D space. Furthermore, it uses global coordinate frame for 3D pose estimation rather than any arbitrary frame of reference as it makes the pose invariant to different camera poses and orientation. Since the network has to only learn the correlations of 2D to 3D points,our network is light-weight. Our image processing based "Multi- Person human pose estimation over inaccurate bounding boxing" is inspired from [1], [21] follows a two level framework that is able to detect poses even in case of inaccurate bounding boxes.

III. OUR APPROACHES

Since, both of our methods are based on "lifting" the 2D points to 3D space, they suffer from the localization error problem and the redundant detection problem if the 2D points are obtained from 2D human pose detectors rather than 2D ground truth from data-sets. Figure 1. explains these two errors even when the bounding boxes are considered as correct with IoU >0.5, the detected human poses can still be wrong. Both the framework address the aforementioned challenge using solution described by [1][22][16].

The Multi-person human pose estimation over inaccurate bounding box approach has two tiers in its architecture. The first tier comprises of three levels [1]. Firstly, a symmetric

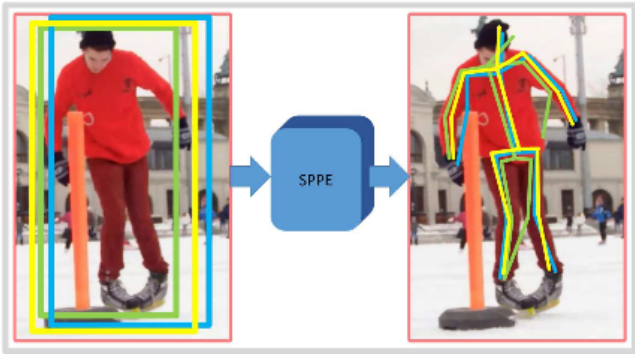


FIGURE 3. Detail visualization of localization error [1]

spatial transformer network (SSTN) is fused to the SPPE to extract a high-quality single person region. Secondly, NMS is introduced that eliminates the redundant poses by using a novel pose distance metric to compare pose similarity. Lastly, pose-guided human proposal generator (PGPG) is used to augment training samples and learn the poses. In the second tier, predicted 2D poses are "lifted" to 3D through simple ResNet based network produces absolute 3D human root localization, and root-relative 3D pose estimation modules [22].

The multi-layer neural network approach demonstrates a simple yet effective baseline for 3D human pose estimation. The principle idea here is to minimize the training overhead in order to make the system work in real-time. To facilitate this, rather than taking input as an RGB image, it relies on 2D feature points of human skeleton as its input. The network is end-to-end trained with several fully connected single layers to produce 3D human points as its output. [16]

IV. IMPLEMENTATION- MULTI- PERSON HUMAN POSE ESTIMATION OVER INACCURATE BOUNDING BOXING

This approach is based on the fact as proven by [1] that 3D Pose is conditionally independent to the 2D pose facilitating a "lift" of 2D pose vectors into its 3D poses.

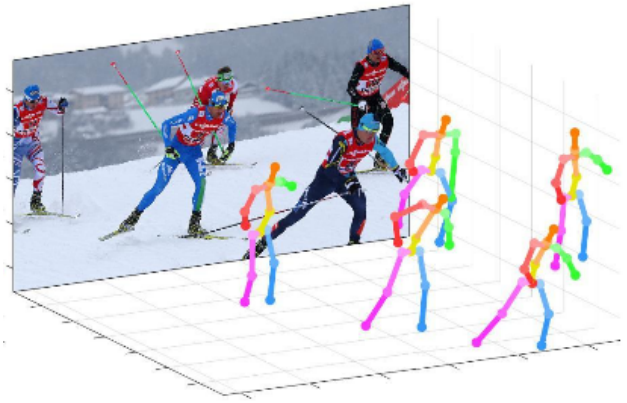


FIGURE 4. Figure explains how 2D points are lifted in 3D using Pose library

A. IMAGE BASED 2D POSE ESTIMATION

This is the first phase of our architecture. It directly takes an image as an input and compounds the 2D pose estimates.The pipeline of our proposed RMPE is illustrated in Figure 3 and described in the following units.

1) Spatial transformer network: It extracts high-resolution feature pyramids and regions of single person even with inaccurate bounding boxes. Used 3D Affine transformation for extracting pose regions. Furthermore, Spatial transformer network is used to mask poses back to original frame.
2) Eliminates redundant poses by using pose distance metric by comparing relative pose similarity. (arXiv:1908.10357v2) . Firstly, the most confident pose is selected as reference, and some poses close to it are subject to elimination by applying elimination criterion.
3) Pose guided proposal generator: It augments the training data and learns the output distribution of a human detector for different poses(each with 5 degree step increase over the space)

Our RMPE produces N dimensional 2D pose vectors(or marginal distributions over the location of individual joints)). Our RMPE model is currently trained on MPII dataset but since its size is limited, we project to train it on Human3.6M dataset whose annotations are adjusted to fit to modern MOCAP systems.

B. NONPARAMETRIC 3D MODEL

Our nonparametric 3D network estimates the root-relative 3D pose P=(x,y,Z)from the proposed 2D annotations from the first phase. We are using the state-of-the-art model proposed by Sunet al. [44]. This model has two parts in it. The first part is ResNet backbone that extracts global feature. The second part uses the features maps extracted from the first part and upsamples it using three consecutive deconvolutional layers with batch normalization layers [23] and ReLU activation function.

Furthermore, 3D heat-maps for each joint is produced by a 1-by-1convolution applied to the upsampled feature map. The soft-argmax operation is used to coalesce the extracted the 2D image coordinates(xj,yj),and the root-relative depth values Z(relative) to be used as final pose. The weights are updated by the Adam optimizer with a mini-batch size of 128. The initial learning rate is set to 0.001 and reduced by a factor of 10 at the 17th epoch.

V. IMPLEMENTATION - RESIDUAL MULTI-LAYER NEURAL NETWORK APPROACH

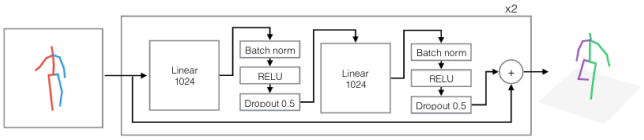


FIGURE 5. Figure describes the architecture of our deep-learning method.

Input to our network is 2D human skeleton points that are directly obtained from state-of-the-art hourglass network of Newellet al. [9],pre-trained on the MPII dataset . We experimented between OpenPose, CPM, and Hourglass networks for the optimal selection and came to conclusion that hourglass network is 10 times faster than the the peers which can serve as a benefit in the case of Human3.6M dataset.

Figure 5. elucidates the architecture of our network. It comprises of a simple batch normalization [23], dropout [24] and Rectified Linear Units (RELUs) [25], as well as residual connections [10] as its backbone. We have 4 to 5 million trainable parameters and 6 linear layers that produces the output of 3n.

1) Linear ReLu layers: Many deep learning approaches aforementioned utilises translation-invariant filters that are applied to entire images [26], [27], [28], [18], [17]. However, low- dimentionality of our input enables us to use lightweight and computationally feasible ReLu to add non-linearity in the network.
2) Residual connections: We have used two residual blocks into our network as it has been proven that it helps to stem down the errors into very deep neural network [10].
3) Batch normalization and dropouts: Since our network takes the 2D point as its input from another network rather than using the 2D ground truth, some marginal errors creeps into the network. In order to mitigate it, we have used Batch normalization [23]and dropout [24].
4) Max-norm constraint: To stabilize the network, we use set up maximum constraints on the wights at every layer to be less than or equal to 1.0

Lastly, entire network was trained for 200 epochs on Nvidia 1080 GPU with decay rate of 0.001 and batch size of 64.

VI. RESULTS AND METRICS

Figure illustrated the multifarious poses network can detect and their 3D representation in global coordinates. Since, we have our first method trained on MPII dataset and second on HUman 3.6M dataset, we have noted that majority of the errors stems into final output in the representation of 3D points rather than "lifting" process.

Table 1 gives the accuracy of our multi-layer deep learning based approach categorized in different pose-actions using protocol 2 (rigid alignment in post-processing).

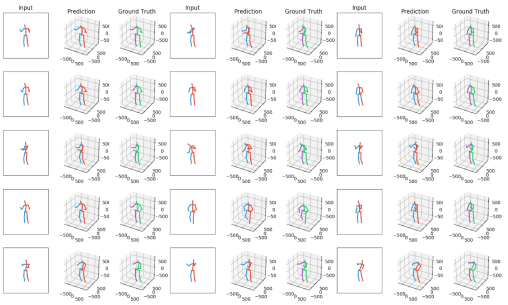


FIGURE 6. Figure describes the output of our deep learning method. It is a chronological representation of how the 2D pose is processed into 3D

Activity	Result
Direct	37.7
Discuss	44.4
Eating	40.3
Greet	42.1
Phone	48.2
Photo	54.9
Pose	44.4
Purch	42.1
Sitting	54.6
SittingD	58.0
Smoke	45.1
Wait	46.4
WaitD	47.6
Walk	36.4
WalkT	40.4

Table 2 and 3 explains the Results on Multi person MPII dataset in comparison to other methods.

Method	Head	Shoulder	Elbow
NIPS17[27]	92.1	89.3	78.9
Ours	88.4	86.5	78.6

TABLE 1. Results of first network over MPII dataset

As it is clear from Figure 7 and 8 , our methods work efficiently in real time on "unseen" samples validating our approaches in real-time.

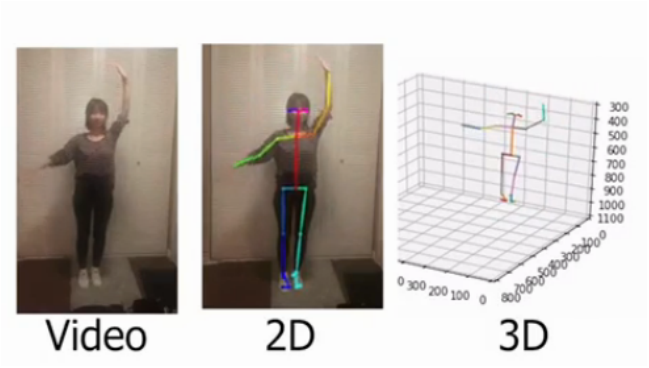


FIGURE 7. Illustration of our deep learning method on unseen sample

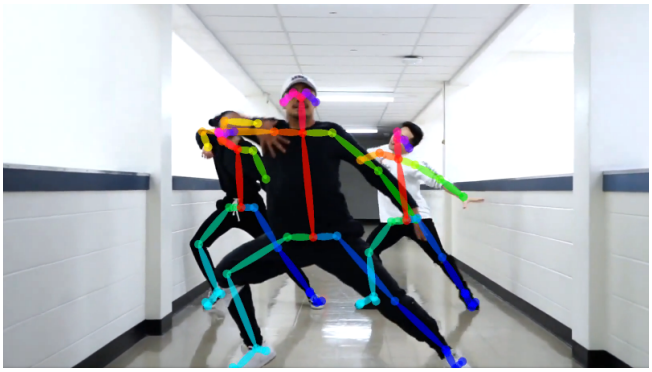
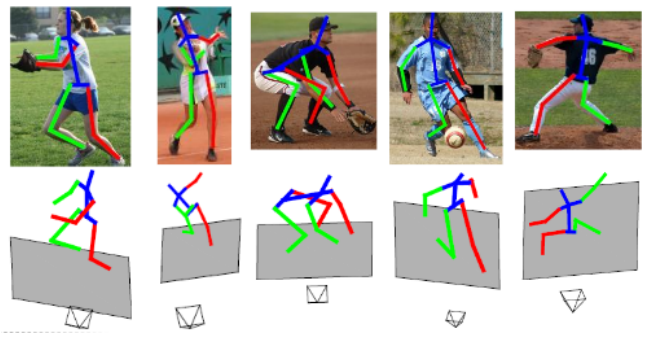


FIGURE 8. Illustration of our image processing method on unseen sample for multi-person pose estimation



Method	Wrist	Hip	Knee	Ankle
NIPS17[27]	69.8	76.2	71.6	64.7
Ours	70.4	74.4	73.0	65.8

TABLE 2. Results of first network over MPII dataset

VII. CONCLUSION

In this work, we have approached the problem of 3D human pose estimation from RGB using two complementary approaches. Whilst the first approach relies on image processing based multi-person framework for intermediate 2D points and the second relies on deep learning based hourglass network for 2D points, the concept of "lifting" the poses from 2D to 3D remains the same. We observed that it is much faster

and efficient in comparison to traditional end-to-end image to 3D pose estimation techniques. Additionally, we were also able to apply the concept to multi-person pose estimation in the wild which is crucial for many real-world applications.

REFERENCES

[1] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2334–2343.
[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, 2014, pp. 3686–3693.
[3] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," IEEE transactions on pattern analysis and machine intelligence, vol. 36, no. 7, pp. 1325–1339, 2013.
[4] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "3d human pose estimation with 2d marginal heatmaps," in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019, pp. 1477–1485.
[5] X. Chen and A. L. Yuille, "Parsing occluded people by flexible compositions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3945–3954.
[6] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 3178–3185.
[7] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "Using k-poselets for detecting people and localizing their keypoints," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3582–3589.
[8] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1653–1660.
[9] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in European conference on computer vision. Springer, 2016, pp. 483–499.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[11] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1212–1221.
[12] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1831–1840.
[13] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in European Conference on Computer Vision. Springer, 2016, pp. 467–483.
[14] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in Advances in neural information processing systems, 2014, pp. 1799–1807.
[15] M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng, "Recurrent 3d pose sequence machines," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 810–819.
[16] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2640–2649.
[17] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua, "Learning to fuse 2d and 3d image cues for monocular body pose estimation," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3941–3950.
[18] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7025–7034.
[19] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," Computers & Graphics, vol. 85, pp. 15–22, 2019.
[20] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," ACM Transactions on Graphics (TOG), vol. 36, no. 4, p. 44, 2017.
[21] C.-H. Chen and D. Ramanan, "3d human pose estimation= 2d pose estimation+ matching," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7035–7043.
[22] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10 133–10 142.
[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.
[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.
[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.
[26] M. F. Ghezelghieh, R. Kasturi, and S. Sarkar, "Learning camera viewpoint using cnn to improve 3d body pose estimation," in 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016, pp. 685–693.
[27] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3d human pose estimation," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2848–2856.
[28] S. Park, J. Hwang, and N. Kwak, "3d human pose estimation using convolutional neural networks with 2d pose information," in European Conference on Computer Vision. Springer, 2016, pp. 156–169.