European Journal of Electrical Engineering and Computer Science
Vol 7 | Issue 6 | December 2023
ISSN 2736-5751

RESEARCH ARTICLE

A Novel Mobile Vision Based Technique for 3D Human Pose Estimation

Sheldon McCall*, Liyun Gong, Afreen Naz, Syed Waqar Ahmed, Wing On Tam, and Miao Yu

ABSTRACT

In this work, we propose a novel technique for accurately constructing 3D human poses based on mobile phone camera recordings. From the originally recorded video frames by a mobile phone camera, firstly a mask R-CNN network is applied to detect the human body and extract 2D body skeletons. Based on the 2D skeletons, a temporal convolutional network (TCN) is then applied to lift 2D skeletons to 3D ones for the 3D human pose estimation. From the experimental evaluations, it is shown that 3D human poses can be accurately reconstructed by the proposed technique in this work based on mobile phone camera recordings while the reconstruction result is very close to the one by a specialized motion capture system.

Keywords: 3D human pose estimation, deep neural network, mask R-CNN, temporal convolutional network.

Submitted: August 31, 2023

Published: December 26, 2023

10.24018/ejece.2023.7.6.573

University of Lincoln, UK.

**Corresponding Author:*
e-mail: Sheldon.McCall@protonmail.com

1. Introduction

3D human pose estimation is the task of predicting the 3D positions of the articulated joint locations of a human body. It is an important research area as it is associated with a wide range of applications including but not limited to human-computer interaction, video surveillance, biomechanics and medication, sports, etc. Recently lots of research works have emerged concerning the prediction of 3D human pose based on certain sensor modalities (e.g., RGD/depth camera [1], wearable sensors [2], etc.). And as in [3], deep learning is the current most popular technique for 3D human pose prediction based on sensor recordings.

Despite the existence of a large variety of research works for 3D pose estimation, little research works have been done for accurately predicting 3D human body poses based on video recordings from a mobile phone, which is the most used electronic equipment nowadays in people's daily lives. To fill this gap, in this work we proposed a methodology of constructing the 3D human pose based on mobile vision, with the aid of modern deep learning techniques. The experimental studies have shown that the proposed approach can accurately construct the 3D human pose given video recordings from a normal mobile phone from both qualitative and quantitative studies.

2. Literature Review

Traditionally, a marker-based motion capture system is applied for 3D human pose estimation. Pose estimation is performed by tracking small reflective markers placed on the surface of subjects. Although they are now commonplace in laboratory environments while achieving high accuracy level [4], these motion capture systems are costly/expensive and limited to be applied to small and highly controlled laboratory environments. Besides, the requirement to wear markers is both inconvenient to the wearer and may also alter his/her natural movement patterns [5]. To ameliorate the limitations of the marker-based 3D pose estimation system, a variety of 3D markerless pose estimation methods have been developed nowadays. In [6], a methodology is developed to quickly and accurately predict 3D positions of body joints captured based on a Kinect sensor. Human body parts are segmented by a randomized decision forest from a single depth image. The inferred body parts are reprojected into the world space and spatial modes of inferred body part distributions and are computed using mean shift to obtain the 3D joint proposals. While the work in [7] estimates 3D human pose in real-world units from a single RGBD image, based on a VoxelPoseNet approach. The experimental results show that the performance outperforms monocular 3D pose estimation approaches from colour as well as pose estimation exclusively from depth. Multiple depths cameras are

exploited and information collected from different depth cameras is fused for human pose estimation [8]. Although no markers are required in the pose estimation systems in [6]–[8], special equipment such as Kinect or other depth cameras are needed in the aforementioned approaches for 3D human pose estimation.

Without relying on markers or special equipment such as depth cameras, estimation of the human pose from a monocular RGB camera has been an emerging research topic in the computer vision community. Reference [9] designs a simple neural network with two branches for simultaneously detecting the root location and regressing the relative locations of other joints, based on a single frame of a normal RGB camera. In [10], the human 3D pose is represented by a volume and a more complicated convolutional neural network (CNN) is applied to predict the voxel-wise likelihood for each joint in the volume. Both the 2D joint heatmap and image cues are fused for 3D pose estimation as in [11]. A kinematic object model consisting of several joints and bones is introduced in [12], to incorporate prior knowledge of the geometric structure of the human body to improve the pose estimation accuracy. Some works ([13], [14]) estimate the 3D pose from a 2D pose instead of directly from the image. In these works, 2D human poses are firstly extracted from an image based on a 2D pose extractor such as OpenPose [15], and a deep neural network (DNN) is applied to lift the estimated 2D poses to 3D ones.

The aforementioned methodologies estimate 3D human poses from a single image recorded by an RGB camera. Instead of relying on a single image, some research works have predicted 3D human poses based on a video sequence containing multiple video frames. Given an input video clip, the 2D human pose sequence is firstly extracted by a 2D pose extractor. Based on the extracted sequence of 2D poses, a variety of networks, such as recurrent neural network (RNN) [16], [17] and temporal convolutional network (TCN) [18], [19] are then exploited for estimating 3D human poses. The advantage of video sequence based pose estimation approaches is that temporal relationship/information between consecutive samples in the sequence can be effectively captured by LSTM or TCN models, to give more accurate 3D pose estimation results as reported in [16]–[19].

In this work, we propose a 3D human pose estimation approach based on video recordings based on a normal mobile phone, which is the most convenient electronic device for video recording nowadays, without relying on special equipment such as wearable markers, depth cameras or even a separate normal RGB camera. Specifically, we first exploit mask R-CNN [20] to extract 2D human poses based on video recordings from a normal mobile phone. A sequence of extracted 2D human poses is then fed into a TCN model for the final 3D pose estimation. By exploiting TCN, temporal dependency information between the 2D human pose sequence can be captured for accurately reconstructing 3D poses. Both qualitative and quantitative experimental results have shown the feasibility of estimating the 3D pose via a normal mobile phone.

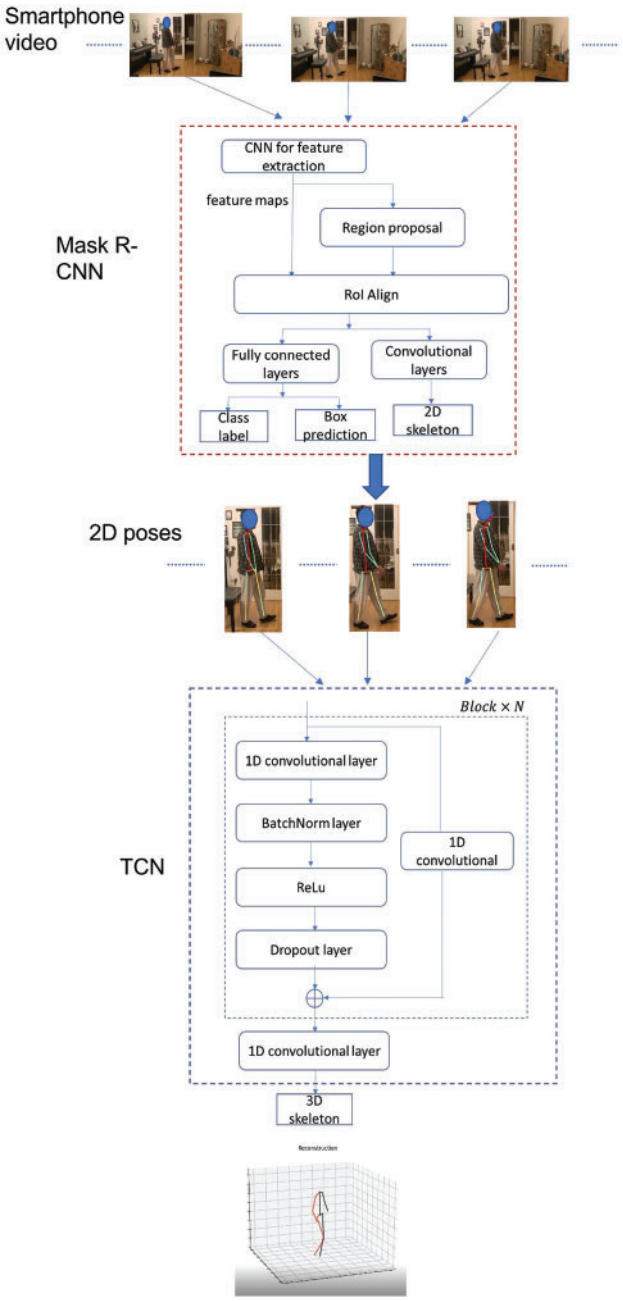

Fig. 1. The diagram illustrating the proposed methodology for 3D pose estimation.

3. Methodology

A diagram of the proposed methodology is shown in Fig. 1. The proposed methodology is composed of a hybrid of two deep neural networks (DNNs). Based on video recordings of a normal mobile phone, firstly a mask R-CNN [20] model is applied to obtain the 2D pose estimation from original video frames. The estimated 2D pose sequences are then fed into a TCN for obtaining the 3D poses.

3.1. *Mask R-CNN*

The mask R-CNN is applied for extracting human 2D skeletons from mobile phone recorded video frames. The classical mask R-CNN contains two stages. In the first stage, feature maps are extracted from an input video

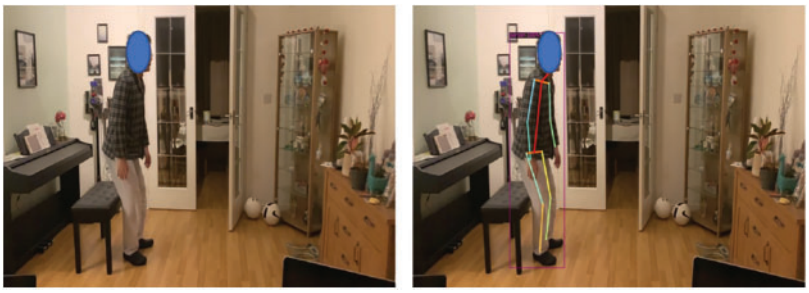

Fig. 2. Left: Original video frame Right: Detected human region (enclosed by the rectangle) and 2D pose by the trained mask R-CNN.

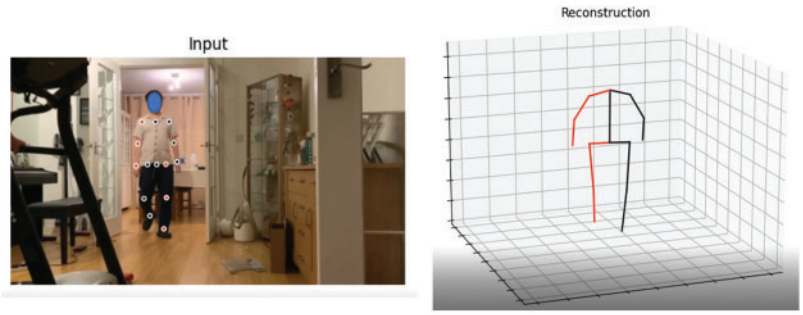

Fig. 3. Left: Input video frame and detected 2D joint positions (marked by circles) Right: Corresponding 3D reconstruction result.

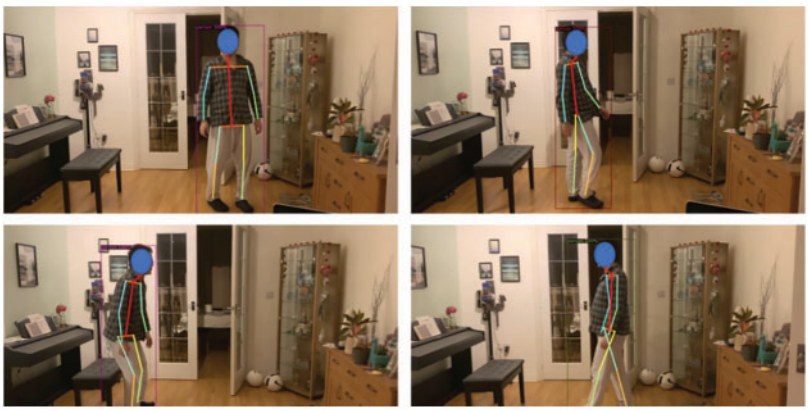

Fig. 4. 2D human skeleton detection results from the mask R-CNN based on video frames recorded by a smartphone.

frame based on certain CNN backbones (such as Resnet-50, Resnet-101, etc., as mentioned in [21]) while a Region Proposal Network Version March 4, 2023 submitted to Journal Not Specified 4 of 10 (RPN) [21] is adopted to propose candidate object regions based on extracted feature maps. In the second stage, the RoIAlign [21] is further exploited to extract features corresponding to the proposed candidate regions, which are fed into different network branches for object classification, bounding box prediction and other tasks as shown in Fig. 1. For the training of the mask R-CNN, a multitask loss is defined as below:

$$L = L_{cls} + L_{box} + L_{mask} \tag{1}$$

where L_{cls}, L_{box} and L_{mask} represent the classification loss, bounding box loss and object mask loss as detailed in [20]. Different optimization algorithms (e.g., SGD, Adagrad, Adadelta, etc. [22]) are applied to train the mask R-CNN by minimizing L. The traditional mask R-CNN is applied to detect the masks of a variety of objects (e.g., human, chair, TV, aeroplane, etc.). In this work, we only focus on the human object as well as the corresponding 2D skeleton key points (wrist, elbow, ankle, knee, etc.) instead of the whole silhouette region. Thus, the class output number associated with the classification branch of the original mask R-CNN is modified to 2 (human object and background) and the following loss function is adopted to train the modified mask-RCNN for detecting 2D skeleton:

$$L = L_{cls} + L_{box} + L_{skeleton} \tag{2}$$

where $L_{skeleton}$ represents the errors between the ground truth 2D skeleton key points and the network predicted ones.

An illustration of the trained mask R-CNN for human detection and 2D human pose detection is shown in Fig. 2.

3.2. TCN

A sequence of extracted 2D skeletons is then fed into a TCN for constructing the 3D human pose. As shown in Fig. 1, the TCN is composed of multiple blocks while each one is composed of a convolutional layer followed by batch norm, Relu activation and drop out, as well as a residual

TABLE I: Calibrated Smartphone Camera Parameters

Centre position	[993.1499, 551.1367]
Focal length	[1708.4, 1703.8]
Radial distortion	[0.2730, –1.1997, 1.5939]
Tangential distortion	[0.0013, 0.0078]

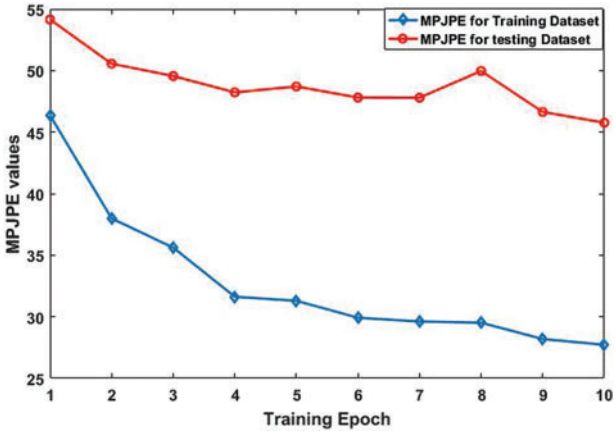

Fig. 5. The evolutions of MPJPE values for training/testing dataset with respect to network training epoch.

link containing one convolutional layer. Convolutional operations are performed in the layer as below:

$$O_t^k = f\left(\sum_{i,j} w_k^{i,j} I_{t+i}^j + b\right) \tag{3}$$

where O represents the k-th output corresponding to the time instance t represents the j-th input at t and b represents the convolutional kernel weight and bias respectively. f (·) represents an activation function (e.g., Relu, sigmoid, softmax, etc.).

With the aid of a series of 1D convolutional operations from multiple blocks, the TCN fully exploits temporal dependencies between 2D human pose samples in the sequence to achieve an accurate 3D pose estimation. The TCN can be trained by minimizing the Mean Per Joint Position Error (MPJPE) loss function, which is defined in (4).

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} ||P_p^i - P_{gt}^i|| \tag{4}$$

where P represents the i-th TCN predicted 3D human pose and P gt indicates the groundtruth ones. Fig. 3 shows an illustration of the 3D pose reconstruction result.

4. Experimental Studies

In this work, the mask R-CNN we use for 2D skeleton extraction is based on a backbone of feature pyramid network (FPN) [23] and trained on the COCO dataset [24]. The network training algorithm and corresponding parameters (e.g., learning rate, weight decay factor, etc.) are chosen the same as the ones in [20]. The trained mask R-CNN is applied for detecting human regions and 2D human poses from input mobile phone video frames. Fig. 4 shows the related results of human detection and 2D human pose detection based on video frames captured from a smartphone.

We obtain 3D human poses data of 7 subjects performing a variety of activities (e.g., walking, sitting, phoning, jogging, etc.) from the Human 3.6 M dataset [25]. Each 3D human pose is associated with 17 3D body joints (hip, right hip, right knee, right foot, left hip, left knee, left foot, spine, thorax, neck, head, left shoulder, left elbow, left wrist, right shoulder, right elbow and right wrist) whose positions are captured by a modern motion capture system. More details can be found in [25]; besides, we've calibrated a smartphone camera based on a chessboard and Matlab camera calibration toolbox [26], with the key calibrated parameters being provided in Table I.

Based on the calibrated smartphone camera parameters, we can map the 3D joint positions of 7 subjects in the Human 3.6 M dataset into corresponding 2D ones in the smartphone camera plane. The 2D/3D joint data

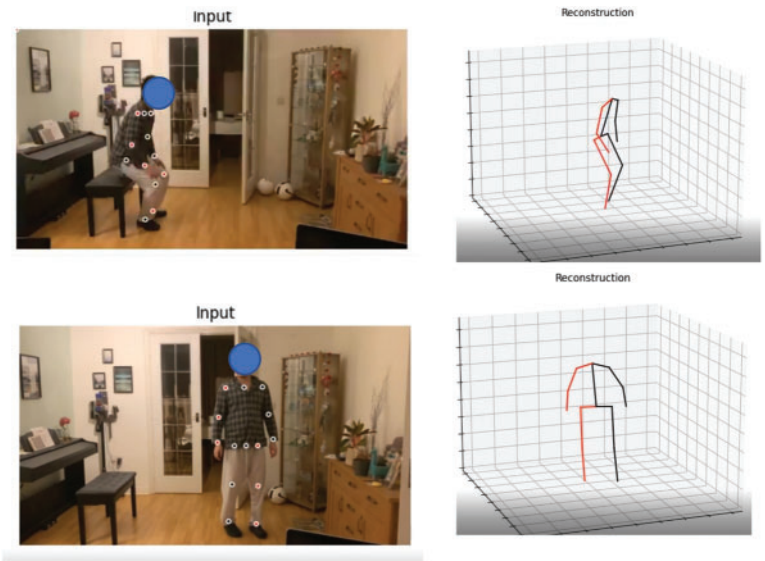

Fig. 6. Continued.

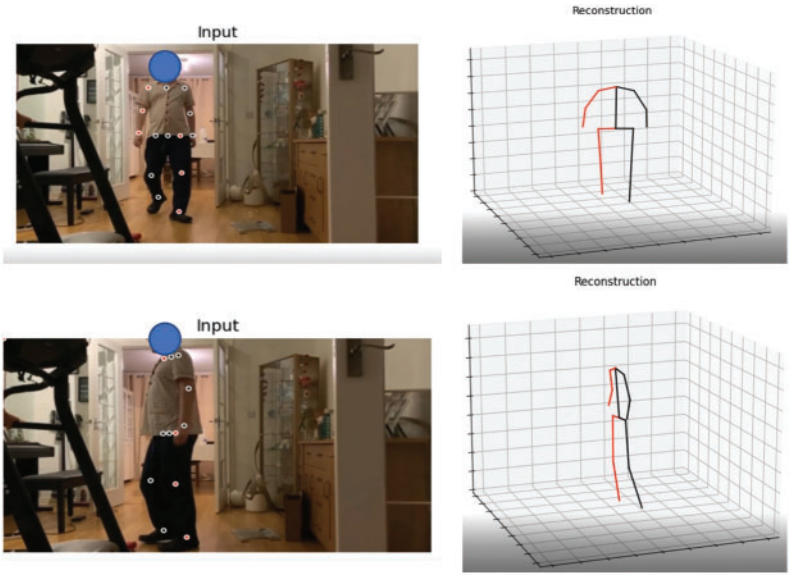

Fig. 6. Visualization of 3D pose estimation results based on mobile phone camera recordings.

corresponding to the five subjects is used for training the TCN model, to make the model able to map 2D poses in the smartphone image plane into 3D poses. And for the remaining two subjects are used for testing. The TCN is trained based on the Adam method with a learning rate of 0.001 and a learning rate decay of 0.95. Fig. 5 shows the evolutions of MPJPEs (in mm) for the training/testing datasets for different training epochs, from which we can observe the MPJPEs for both datasets tend to decrease as the epoch increases, which indicates that more accurate 3D pose estimation results can be obtained as the training epoch increases.

The trained TCN can then be used to map the 2D human poses in the smartphone image plane to 3D ones. Fig. 6 visualizes the obtained 3D construction results corresponding to different activities (sitting, standing, turning, walking) recorded by a smartphone camera. 2D body joints are first detected by the mask R-CNN from the video frames, which are then fed into the trained TCN to reconstruct 3D human poses.

Moreover, we've evaluated TCN models under different architectures, with different numbers of 1D convolutional filters and block numbers as shown in Tables II and III.

TABLE II: 3D Pose Reconstruction Performance with Different Number of convolutional Filters

Filter number	Filter number
256	48.3
512	44.3
1024	45.9
2048	45.7

TABLE III: 3D Pose Reconstruction Performance with Different Number of Blocks in the TCN Model

Block number	No. of network	MPJPE (mm)
3	2,171,443	44.3
4	3,222,067	44.5
5	4,272,691	44.0

We can find that the model with the convolutional filter number 512 can achieve the most accurate 3D pose reconstruction performance. And from Table III, we can find that there is no obvious advantage to introducing more blocks, with only 0.3 mm improvement by increasing the block number from 3 to 5 almost doubling the network parameters.

Finally, we've compared the TCN based method with the fully connected network (FCN) model approach proposed in [13], which estimates the 3D pose based on only a single 2D pose from one frame without exploiting any temporal information between 2D poses from multiple frames in a sequence. The MPJPEs (in mm) corresponding to different activities in our test dataset is calculated and summarized in Table IV, from which we can find that the TCN based method achieves better performance with smaller MPJPEs for the majority of activities as well as the mean MPJPE value, due to its advantage of exploiting

TABLE IV: Comparisons of the TCN model and the FCN in [13] for 3D pose estimation

	TCN model	FCN model in [13]
Directions	42.55	38.57
Discussion	48.93	44.80
Eating	41.54	44.13
Greeting	44.08	43.79
Phoning	48.31	48.38
Photo	49.82	56.67
Posing	47.15	44.69
Purchases	42.31	42.56
Sitting	52.78	54.32
Sitting down	54.55	60.32
Smoking	43.18	46.58
Waiting	45.02	46.62
WalkDog	45.24	49.79
Walking	35.17	38.36
Walking together	36.25	41.63
Average	44.30	46.75

temporal information within a 2D pose sequence for 3D pose construction.

5. Conclusions

In this work, we propose a deep learning-based technique for constructing 3D human body poses, based on video recordings from a mobile phone camera. The proposed approach is a hybrid of mask R-CNN and TCN. The mask R-CNN is applied to extract 2D human body poses from mobile camera recordings while the 2D poses are mapped to 3D ones based on the TCN. The experimental results show the effectiveness of the proposed approach for estimating 3D human poses based on mobile phone camera recordings. In future work, we will investigate more advanced deep neural network architectures for 3D pose estimation (e.g., graphical neural network, transformer, etc.); besides, we will apply the developed mobile phone-based 3D human body pose estimation technique for applications in a variety of domains (e.g., constructing 3D body poses for gait parameters extraction for medical applications, etc.).

Author Contributions

S. Mccall, L. Gong and A. Naz–writing; S. Ahmed and W. Tam–review and editing; M. Yu–supervision.

Funding

This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 778602 ULTRACEPT.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Informed consent was obtained from all participants involved in the study.

Conflict of Interest

Authors declare they do not have any conflict of interests.

References

[1] Wang J, Tan S, Zhen X, Xu S, Zheng F, He Z, *et al.* Deep 3D human pose estimation: a review. *Comput Vis Image Und*. 2021;210:1–21.

[2] Guzov V, Mir A, Sattler T, Pons-Moll G. Human POSEitioning System (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021.

[3] Zheng C, Wu W, Yang T, Zhu S, Chen C, Liu R, *et al.* Deep learning-based human pose estimation: a survey. *arXiv*. 2020, arXiv: 2012.13392.

[4] Topley M, Richards J. A comparison of currently available optoelectronic motion capture systems. *J Biomech*. 2020;106:1–5.

[5] Colyer S, Evans M, Cosker D, Salo A. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports Med*. 2018;4(24):1–15.

[6] Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, *et al.* Real-time human pose recognition in parts from single depth images. *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Colorado Springs, USA, 2011.

[7] Zimmermann C, Welschehold T, Dornhege C, Burgard W, Brox T. 3D human pose estimation in RGBD images for robotic task learning. *IEEE International Conference on Robotics and Automation*. Brisbane Convention Exhibition Centre, Brisbane, Australia, 2018.

[8] Hansen L, Siebert M, Diesel J, Heinrich M. Fusing information from multiple 2D depth cameras for 3D human pose estimation in the operating room. *Int J Comput Ass Rad*. 2019;14:1871–9.

[9] Li S, Chan A. 3D human pose estimation from monocular images with deep convolutional neural network. *Asian Conference on Computer Vision (ACCV)*, Singapore; 2014.

[10] Pavlakos G, Zhou X, Derpanis K, Daniilidis K. Coarse-to-fine volumetric prediction for single-image 3d human pose. *IEEE International Conference on Robotics and Automation, In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, 2017.

[11] Zhou K, Han X, Jiang N, Jia K, Lu J. HEMlets pose: learning part-centric heatmap triplets for accurate 3d human pose estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea,2019.

[12] Zhou X, Sun X, Zhang W, Liang S. Deep kinematic pose regression. *European Conference on Computer Vision (ECCV)*. Amsterdam, Netherlands,2016.

[13] Martinez J, Hossain R, Romero J, Little J. A simple yet effective baseline for 3d human pose estimation. *International Conference on Computer Vision (ICCV)*. Venice, Italy, 2017.

[14] Tome D, Russell C, Agapito L. Lifting from the deep: convolutional 3D pose estimation from a single image. *IEEE International Conference on Robotics and Automation, In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Hawaii, USA, 2017.

[15] Cao Z, Hidalgo G, Simon T, Wei S, Sheikh Y. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans Pattern Anal Mach Intelli*. 2021;43:172–86.

[16] Lee K, Lee I, Lee S. Propagating LSTM: 3D pose estimation based on joint interdependency. *European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018.

[17] Hossain M, Little J. Exploiting temporal information for 3D pose estimation. *European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018.

[18] Cheng Y, Yang B, Wang B, Yan W, Tan R. Occlusion-aware networks for 3D human pose estimation in video. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019.

[19] Pavllo D, Feichtenhofer C, Grangier D, Auli M. 3D human pose estimation in video with temporal convolutions and semi-supervised training. *IEEE International Conference on Robotics and Automation, In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[20] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. *International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[21] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *Twenty-Ninth Conference on Neural Information Processing Systems*, Montréal CANADA, 2015.

[22] Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MA, USA: MIT Press: Cambridge; 2016.

[23] Lin T, Dollr P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. *International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[24] Coco:common objects in context. 2021. Available from: http://cocodataset.org/home.

[25] Human3.6M dataset. 2021. Available from: http://vision.imar.ro/human3.6m/description.php.

[26] Camera calibration. 2021. Available from: https://uk.mathworks.com/help/vision/camera-calibration.html.