Evaluation of CNN-based human pose estimation for body segment lengths assessment

Saman Vafadar[1], Laurent Gajny[1], Matthieu Boëssé[1], Wafa Skalli[1]

[1]Arts & Métiers ParisTech

Abstract *Human pose estimation (HPE) methods based on convolutional neural networks (CNN) have demonstrated significant progress and achieved state-of-the-art results on human pose datasets. In this study, we aimed to assess the performance of CNN-based HPE methods for measuring anthropometric data. A Vicon motion analysis system as the reference system and a stereo vision system recorded ten asymptomatic subjects standing in front of the stereo vision system in a static posture. Eight HPE methods estimated the 2D poses which were transformed to the 3D poses by using the stereo vision system. Percentage of correct keypoints, 3D error, and absolute error of the body segment lengths are the evaluation measures which were used to assess the results. Percentage of correct keypoints – the standard metric for 2D pose estimation – showed that the HPE methods could estimate the 2D body joints with a minimum accuracy of 99%. Meanwhile, the average 3D error and absolute error for the body segment lengths are 5 cm.*

Keyword ergonomics, anthropometry, deep learning, stereo vision

Introduction

Work on Convolutional Neural Network (ConvNet, or CNN) as a neural network model to imitate the ability of human being for pattern recognition has already begun since the late seventies [1], [2]. However, the computational costs of ConvNets had restricted its extensive use. Nowadays, the GPU-accelerated computing techniques have made the training procedure more efficient [3], resulting in the wide applications of CNNs in handwriting recognition [4], behavior recognition [5], human pose estimation [6], and medical image analysis [7].

Human Pose Estimation (HPE) methods are computer vision techniques to localize the human body joints. HPE methods, depending on the interpretation of the body structure are categorized into generative, discriminative and hybrid methods [8]. Generative methods match the image observations – or image features which are the most representative information, e.g., edges, silhouettes – with the projection of the employed human body model to the image by adjusting the body model. On

the other hand, discriminative methods model the relations between the image observations and human poses [6]. For the moment, the most popular method for feature extraction is ConvNet. HPE methods based on ConvNets have demonstrated significant progress on challenging benchmarks (e.g. MPII [9]). The success of HPE methods based on ConvNets justifies investigation for specific applications such as anthropometry measurement.

The current applications of anthropometry measurements can be found in ergonomics. For instance, for designing fitting materials, such as the workspace and clothing to improve the safety and comfortability [10]. This study aimed to evaluate the CNN-based HPE methods for measuring anthropometric data. A stereo vision system combined with 2D HPE methods were used to recover the 3D body joint positions. Also, a marker-based motion capture system was used to assess the validity of the results.

Materials and Methods

Ten healthy subjects (5 males, 5 females) after informed consent participated in this study. The work has been approved by the relevant ethics committee (CPP 06036). Subjects were on average 24 years old (SD: 2, range: 21-27), mean height was 173 cm (SD: 9 cm, range: 160-187 cm), mean body mass was 64 kg (SD: 9 kg, range: 53-80 kg), and mean Body Mass Index (BMI) was 21 (SD: 2, range: 19-25).

Two calibrated and synchronized devices, a Vicon motion capture system (Vicon Motion Systems Ltd, UK) equipped with twelve Vicon Vero cameras as the reference system and a stereo vision system, were used to capture the data with the frequency of 100 Hz. The stereo vision system consisted of two GoPro Hero 7 Black cameras (GoPro, Inc., US) which recorded videos with the resolution of 1080p and linear field of view. The relative distance and angle between the two cameras were 75 cm and 15 deg, respectively, mounted on a tripod with a height of 120 cm.

Forty-eight reflective markers, according to a designed marker-set compatible with [11], [12], were attached to the subject's body segments. As shown in Figure 1**Erreur ! Source du renvoi introuvable.**, subjects were asked to stand in a static posture in front of the stereo vision system, approximately 3.5 m away from the device, for 3 seconds. Table 1**Erreur ! Source du renvoi introuvable.** shows the detail of the estimation of the morphological data using the reconstructed 3D positions of the reflective markers.

The 3 seconds of the videos recorded by the stereo vision system resulted in 300 frames. For the first frame of each video, a bounding box was manually defined to crop the frame in which the subject was at the center. Then, for each next frame, a dynamic bounding box was determined based on the estimated pose of the previous frame so that the height of the subject was equal to 75% of the bounding box height. Then, the cropped frames were resized to the resolution of 256 × 256 pixels and 368 × 368 pixels which are compatible with the size of the input image of the selected HPE methods. The cropped and subsequently resized images were saved to be used

Figure 1. One of the subjects standing in front of the stereo vision system (This image has been taken by the left camera of the stereo vision system).

as the input of HPE methods. Eight HPE methods [13]–[20] based on convolutional neural networks for which codes were publicly available, achieving the state-of-the-art results on challenging benchmarks (e.g., MPII [9]), have been selected to estimate the 2D poses. The 2D poses consist of 16 body keypoints including upper and lower head, neck, shoulders, elbows, wrists, torso, hips, knees and ankles (Wei et al., 2016 [17] estimate 14 keypoints; excluding lower head and torso). After 2D pose estimation for all the captured frames, since the subjects were standing in a static posture across the 300 frames, the mean values of the estimated body keypoints were computed. 2D to 3D pose lifting has been accomplished with the stereo vision system. After the intrinsic and extrinsic calibration of the stereo vision system, the 3D positions can be recovered using the perspective projection of the 3D point on the image planes. Herein, the retrieved 2D body keypoints were assumed to be the perspective projection of the corresponding 3D point on the left and right image plane; Thereby, the 3D positions of the body keypoints were obtained by using the linear triangulation method [21]. Hence, the morphological data were subsequently computed, as shown in Table 1**Erreur ! Source du renvoi introuvable.**.Three metrics evaluate the accuracy of the HPE methods, PCKh, 3D error and absolute error of the body segments lengths. **Percentage of Correct Keypoints** normalized with the **head** segment length (**PCKh**) defines an estimated keypoint to

be correct if the distance to the corresponding reference value is less than a threshold which is a function of the head segment length – for instance PCKh@0.5 considers the 50% of the head segment length as the threshold. **3D error** measures the Euclidean distance between a reconstructed 3D keypoint and its reference value. **Absolute error** which have been used to compare the segment lengths consists of the absolute difference between the measured values by the Vicon and stereo vision system.

Table 1. The acronyms stand for: HLE = Humeral Lateral Epicondyle, HME = Humeral Medial Epicondyle, USP = Ulnar Styloid Process, RSP = Radial Styloid Process, HJC = Hip Joint Center (femoral head based on the method of Bell et al [22]), FLE = Femoral Lateral Epicondyle, FME = Femoral Medial Epicondyle, LM = Lateral Malleolus, MM = Medial Malleolus.

Body segment	*Vicon system*	*stereo vision system*
Forearm	0.5×(HLE+HME) – 0.5×(USP+RSP)	Elbow – Wrist
Thigh	HJC – 0.5×(FLE+FME)	Hip – Knee
Leg	0.5× (FLE+FME) – 0.5×(LM+MM)	Knee – Ankle

Results

Figure 1**Erreur ! Source du renvoi introuvable.** shows the 2D pose estimation accuracy using the PCKh metric. The accuracy of all the selected HPE methods for 2D pose estimation was above 99% using the standard metric PCKh@0.5 [9]. Table 2 shows the 3D errors for body keypoints. The mean value of the 3D error is 5 cm. Table 3 shows the absolute error for the body segment lengths which were obtained based on the estimated 3D poses. The mean error for the lengths of the body segments, same as the 3D error, is 5 cm.

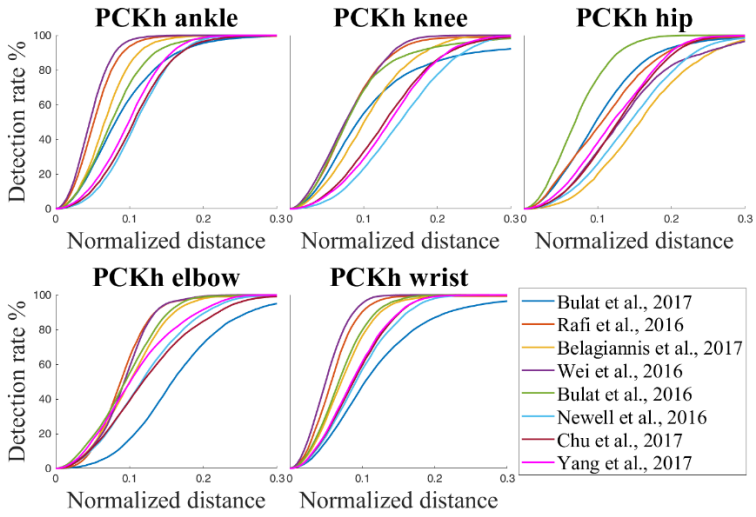

Figure 2. Quantitative results on 2D estimated keypoints using PCKh metric.

Table 2. Quantitative results on 3D reconstructed keypoints using 3D error. Mean (Min, Max) values are reported in millimeter.

	left ankle	*left knee*	*left hip*	*left elbow*	*left wrist*
Bulat et al., 2017 [13]	46 (16, 100)	43 (24, 78)	52 (25, 126)	68 (34, 140)	95 (31, 207)
Rafi et al., 2016 [14]	28 (8, 68)	48 (12, 182)	59 (30, 96)	88 (22, 598)	118 (25, 667)
Belagiannis et al., 2017 [15]	24 (8, 48)	44 (21, 81)	47 (28, 66)	34 (19, 62)	47 (10, 111)
Wei et al., 2016 [16]	23 (10, 42)	44 (19, 80)	51 (26, 101)	44 (21, 73)	60 (11, 106)
Bulat et al., 2016 [17]	32 (14, 73)	25 (7, 42)	36 (12, 64)	45 (30, 58)	36 (19, 56)
Newell et al., 2016 [18]	41 (32, 68)	64 (44, 99)	58 (44, 78)	61 (27, 122)	64 (23, 121)
Chu et al., 2017 [19]	37 (30, 61)	60 (34, 115)	50 (27, 69)	60 (37, 83)	46 (17, 127)
Yang et al., 2017 [20]	35 (24, 64)	54 (35, 83)	51 (32, 76)	52 (33, 91)	41 (14, 76)
	right ankle	*right knee*	*right hip*	*right elbow*	*right wrist*
Bulat et al., 2017 [13]	41 (18, 79)	41 (12, 74)	49 (20, 117)	95 (31, 207)	104 (35, 288)
Rafi et al., 2016 [14]	49 (22, 126)	53 (17, 167)	58 (21, 152)	118 (25, 667)	126 (21, 790)
Belagiannis et al., 2017 [15]	29 (4, 53)	32 (16, 55)	51 (28, 67)	47 (10, 111)	51 (19, 83)
Wei et al., 2016 [16]	24 (5, 45)	29 (7, 52)	51 (33, 85)	60 (11, 106)	42 (17, 77)
Bulat et al., 2016 [17]	40 (12, 70)	43 (9, 80)	35 (13, 62)	36 (19, 56)	41 (14, 78)
Newell et al., 2016 [18]	36 (22, 61)	41 (25, 85)	45 (21, 83)	64 (23, 121)	54 (26, 93)
Chu et al., 2017 [19]	33 (19, 52)	52 (21, 85)	63 (26, 148)	46 (17, 127)	53 (26, 101)
Yang et al., 2017 [20]	32 (14, 59)	41 (17, 73)	42 (18, 100)	41 (14, 76)	42 (18, 65)

Table 3. Quantitative results on the body segment lengths using absolute error. Mean (Min, Max) values are reported in millimeter.

	left forearm	*left thigh*	*left leg*
Bulat et al., 2017 [13]	83 (8, 195)	38 (1, 78)	34 (5, 87)
Rafi et al., 2016 [14]	63 (8, 173)	74 (14, 125)	47 (11, 142)
Belagiannis et al., 2017 [15]	32 (3, 64)	60 (11, 138)	42 (0, 140)
Wei et al., 2016 [16]	44 (3, 93)	49 (3, 130)	55 (21, 131)
Bulat et al., 2016 [17]	21 (3, 73)	30 (0, 65)	56 (24, 145)
Newell et al., 2016 [18]	58 (3, 215)	69 (15, 132)	71 (25, 123)
Chu et al., 2017 [19]	35 (3, 72)	33 (7, 87)	69 (1, 129)
Yang et al., 2017 [20]	33 (2, 78)	53 (0, 106)	54 (10, 118)
	right forearm	*right thigh*	*right leg*
Bulat et al., 2017 [13]	81 (11, 147)	43 (2, 114)	59 (4, 149)
Rafi et al., 2016 [14]	37 (0, 92)	40 (2, 124)	47 (0, 115)
Belagiannis et al., 2017 [15]	30 (8, 84)	54 (1, 126)	29 (4, 59)
Wei et al., 2016 [16]	33 (3, 89)	57 (13, 110)	25 (0, 59)
Bulat et al., 2016 [17]	46 (7, 97)	51 (11, 110)	43 (17, 98)
Newell et al., 2016 [18]	37 (21, 65)	29 (4, 82)	33 (0, 83)
Chu et al., 2017 [19]	53 (18, 104)	79 (5, 165)	49 (0, 112)
Yang et al., 2017 [20]	45 (5, 87)	44 (14, 78)	40 (0, 93)

Discussion

In this study, with the goal of evaluation of CNN-based HPE methods for measuring body segment lengths, eight HPE methods were employed to estimate the 2D poses which were subsequently transformed to 3D poses using a stereo vision system. The body segment lengths were computed based on the estimated 3D poses. The results show the errors of 2D pose estimation, 3D reconstruction, and anthropometric data – following the selected hierarchy to compute the anthropometric data.

Table 4 shows the accuracy of the HPE methods on the MPII dataset using the PCKh metric. These results showed that the most demanding body keypoints to estimate were the ankle, knee, and wrist. However, Figure 2, which shows the accuracy of the HPE methods on the dataset of this study, shows that the knee, hip, and elbow were the most difficult keypoints to locate accurately. Also, evaluation using PCKh@0.5 metric that measures the accuracy of all methods for all keypoints to be higher than 99% indicates that this value is comparably higher than the values reported in the literature. These two points highlight the role of the training and testing dataset. In the dataset of this study, there was no occlusion – neither self-occlusion nor occlusion by other entities – or challenging pose. Now, comparing all

the HPE methods may underline that one method cannot outperform all other methods for all the body keypoints. For instance, Wei et al. 2016 [16] achieve the best performance for estimating the positions of the ankle, knee, elbow, and wrist, but its accuracy for the hip is significantly less than the other HPE methods – i.e., there is no universal method outperforming all the other methods.

The results showed that the main parameters influencing the accuracy of an HPE method based on CNNs can be, the architecture of the convolutional neural network, the training and testing dataset, and the training strategy. Size of the input image of Wei et al., 2016 [17], which is the most accurate methods for the estimation of wrist and ankle based on the results shown in Figure 2, is 368 × 368 while for the others is 256 × 256. Thus, it strengthens the hypothesis that the resolution of the input image could also be a prominent factor in determining the accuracy of the HPE methods. Also, there are minor parameters which affect the output of the HPE methods and thereby their accuracies, such as the noise of the input image, the position, and scale of the subject inside the cropped frames.

Table 4. Quantitative results on the **MPII dataset** using the PCKh@0.5 metric.

	ankle	*Knee*	*hip*	*elbow*	*wrist*
Bulat et al., 2017 [13]	64.0	70.5	79.1	78.8	71.5
Rafi et al., 2016 [14]	73.4	80.6	86.8	86.4	81.3
Belagiannis et al., 2017 [15]	78.4	82.6	87.9	88.2	83.0
Wei et al., 2016 [16]	79.4	82.8	88.4	88.7	84.0
Bulat et al., 2016 [17]	81.9	85.7	89.4	89.9	85.3
Newell et al., 2016 [18]	83.6	87.4	90.1	91.2	87.1
Chu et al., 2017 [19]	85.0	88.0	90.6	91.9	88.1
Yang et al., 2017 [20]	85.3	88.6	91.1	91.9	88.2

3D pose recovery has been made using the linear triangulation method by assuming that the estimated 2D poses are the perspective projection of the 3D pose. However, the deviation of the 2D poses from their reference value may be exacerbated through triangulation. Figure 3 shows an explicit example that the estimated ankle joints on the left and right images are not stereo correspondent – i.e., the estimated ankles (either left or right ankle) do not refer exactly to the same anatomical point.

Table 2 shows the 3D error for the body keypoints. The mean 3D error for the estimation of body keypoints is 5 cm. In a similar study [23] which has used Microsoft Kinect™ for 3D pose estimation in a static posture, the average 3D error in standing posture has been reported to be 8 cm and 9 cm for the first- and second-generation Kinect sensor, respectively.

The 3D error, reported in Table 2, shows that the maximum errors, for the elbow and wrist using the HPE method of Rafi et al., 2016 [14], are 60 cm and 67 cm,

respectively. This occurrence is because of the false 2D detections of the HPE method for several frames of the static acquisition for one of the subjects.

Figure 3. 2D estimation of the ankle joints for a single frame. The red, green, and blue dots, represent the estimation of the right ankle, left ankle, and the reference values, respectively.

Table 3 shows the absolute error for the body segment lengths, while the mean error for the length of the body segments is 5 cm. A deeper look at this table shows that some methods cannot estimate the left and right body segment with the same accuracy. For instance, in the meanwhile that the 3D error for ankle and knee estimation, by Bulat et al., 2017 [13], is 4 cm, the absolute error for the left leg is 3 cm, and for the right leg is 6 cm. It also may highlight that post-processing could improve the uniformity of the results.

In conclusion, even though a stereo vision system combined with HPE methods can provide a cost-effective, easy to use, time efficient tool to measure the morphological data, the mean error is 5 cm that may not be adequate for applications in ergonomics. However, HPE methods may open new perspectives for measuring morphological data.

Acknowledgments

The authors thank the ParisTech BiomecAM chair program, on subject-specific musculoskeletal modelling and in particular Société Générale and COVEA.

References

[1] K. Fukushima, “Neocognition: a self,” *Biol. Cybern.*, vol. 202, pp. 193–202, 1980.

[2] K. Fukushima, “Cognitron: A self-organizing multilayered neural network,” *Biol. Cybern.*, vol. 20, no. 3–4, pp. 121–136, 1975.

[3] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” *Neurocomputing*, vol. 234, no. November 2016, pp. 11–26, 2017.

[4] Y. C. Wu, F. Yin, and C. L. Liu, “Improving handwritten Chinese text recognition using neural network language models and convolutional neural network shape models,” *Pattern Recognit.*,

vol. 65, no. February 2016, pp. 251–264, 2017.

[5] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, “Hybrid deep neural networks for face emotion recognition,” *Pattern Recognit. Lett.*, vol. 115, pp. 101–106, 2018.

[6] W. Gong *et al.*, “Human Pose Estimation from Monocular Images: A Comprehensive Survey,” *Sensors (Basel).*, vol. 16, no. 12, pp. 1–39, 2016.

[7] G. Litjens *et al.*, “A survey on deep learning in medical image analysis.,” *Med. Image Anal.*, vol. 42, no. December 2012, pp. 60–88, 2017.

[8] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, “3D Human pose estimation: A review of the literature and analysis of covariates,” *Comput. Vis. Image Underst.*, vol. 152, pp. 1–20, 2016.

[9] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D Human Pose Estimation: New Benchmark and State of the Art Analysis,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3686–3693, 2014.

[10] B. Bonnechère *et al.*, “Determination of the precision and accuracy of morphological measurements using the Kinect™ sensor: Comparison with standard stereophotogrammetry,” *Ergonomics*, vol. 57, no. 4, pp. 622–631, 2014.

[11] A. Leardini, Z. Sawacha, G. Paolini, S. Ingrosso, R. Nativo, and M. G. Benedetti, “A new anatomically based protocol for gait analysis in children,” *Gait Posture*, vol. 26, no. 4, pp. 560–571, 2007.

[12] VICON Motion System, “Nexus 2.6, Documentation, Full body modeling with Plug-in Gait,” *VICON Documentation*, 2017. [Online]. Available: https://docs.vicon.com/display/Nexus26/Full+body+modeling+with+Plug-in+Gait. [Accessed: 15-Jan-2019].

[13] A. Bulat and G. Tzimiropoulos, “Binarized Convolutional Landmark Localizers for Human Pose Estimation and Face Alignment with Limited Resources,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 3726–3734.

[14] A. Bulat and G. Tzimiropoulos, *Human pose estimation via convolutional part heatmap regression*, vol. 9911 LNCS. 2016.

[15] U. Rafi, I. Kostrikov, J. Gall, and B. Leibe, “An Efficient Convolutional Network for Human Pose Estimation,” *Br. Mach. Vis. Conf.*, pp. 1–11, 2016.

[16] V. Belagiannis and A. Zisserman, “Recurrent Human Pose Estimation,” in *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 468–475.

[17] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional Pose Machines,” 2016.

[18] A. Newell, K. Yang, and J. Deng, “Stacked Hourglass Networks for Human Pose Estimation,” 2016.

[19] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, “Multi-Context Attention for Human Pose Estimation,” *Cvpr2017*, pp. 1831–1840, 2017.

[20] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, “Learning Feature Pyramids for Human Pose Estimation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 1290–1299.

[21] J. Heikkilä and O. Silvén, “A Four-step Camera Calibration Procedure with Implicit Image Correction,” in *Cvpr*, 1997, vol. 97, p. 1106.

[22] A. L. Bell, D. R. Pedersen, and R. A. Brand, “A comparison of the accuracy of several hip

center location prediction methods," *J. Biomech.*, vol. 23, no. 6, pp. 617–621, 1990.

[23] X. Xu and R. W. McGorry, "The validity of the first and second generation Microsoft Kinect for identifying joint center locations during static postures," *Appl. Ergon.*, vol. 49, pp. 47–54, 2015.