Self-Occluded Human Pose Recovery in Monocular Video Motion Capture

Leila Malekian
School of Computing Sciences
University of East Anglia
Norwich, United Kingdom
l.malekian@uea.ac.uk

Rudy Lapeer
School of Computing Sciences
University of East Anglia
Norwich, United Kingdom
r.lapeer@uea.ac.uk

***Abstract*—Monocular video motion capture is a popular alternative to more expensive technologies such as marker-based optical motion capture. However, motions that are occluded from the single camera view, for example, due to self-occlusion, are difficult to recover. In this paper, we propose a machine learning-based method that is used in post-processing to reconstruct the incorrect motions that are caused by self-occlusion. The post-processing network is trained on a dataset acquired from three subjects doing 30 different basic exercise motions that include self-occlusion. The collected data comprise single video camera footage and optical motion capture data as the ground truth. To correctly reconstruct the occluded motion, action recognition information is used to select a machine learning model that is trained on the specific motion. The performance of predictive and non-predictive networks are compared to each other and also with the state of the art in human motion estimation. The results show a significant reduction of the overall pose error and the pose error for selected body parts with a large degree of self-occlusion.**

***Index Terms*—human pose estimation, self-occlusion, single view video, SMPL model, machine learning, deep learning**

I. Introduction

Human Pose and Mesh Estimation is an important problem that has attracted a lot of research interest in the last three decades. Vision-based estimation of the human pose and shape from monocular video whilst addressing the problem of self-occlusion is the focus of this research paper. There is a wide variety of methods introduced in the literature to solve the Human Pose Estimation (HPE) problem. Despite advancement, some unresolved challenges such as ambiguity, occlusion by self or an object, collision of body parts, unnatural poses, issues with low quantity and diversity of data, and use of different skeleton/human models, still exist.

Improving benchmarks, protocols and toolkits for 3D human mesh recovery methods are among the future directions of research in this area. There is a lack of large-scale online 3D mesh data sets and protocols for effective evaluation of 3D human mesh recovery methods. The SMPL human model [1], consisting of a human mesh and an underlying skeleton, is one of the most popular 3D virtual human models. Unlike the more traditional 3D key point skeleton representation of the human body, the joints of the SMPL model have both rotational and positional degrees of freedom. This makes SMPL a more complete representation of human body movement compared to the traditional 3D key point skeleton.

The majority of the traditional online benchmark datasets for 3D human pose estimation have 3D key points as ground truth. However, only a handful of datasets have been created with the SMPL human model ground truth. In this research, the focus is on improving the estimation of self-occluded motions using SMPL models. We have also created a dataset of self-occluded motions performed by three subjects with SMPL model ground truth. We also introduced a new error measurement method for evaluating the correct pose of SMPL models relating to the SMPL rotation parameters.

The input video in our research does a full-body capture of a single subject using a monocular video camera, whilst no external objects or other persons are in between the subject and the camera. The main source of occlusion is self-occlusion of the body parts due to having only one camera. The output is the parametric SMPL model. Current HPE methods are ranked based on the achievable pose error and not by handling extreme cases, so their ability to handle self-occlusion or external occlusion is not part of their performance measure. Most of the more recent work on occlusion is on external occlusion, for example being occluded by another person, external objects, being outside the frame, crowded scenes or having only a single image. In the area of video-based or 3D HPE with SMPL model estimation, the self-occlusion problem is largely ignored and remains an outstanding issue in state-of-the-art methods.

When a large part of human motion is occluded, the information about the performed action can be used to reconstruct the 3D motion in the output. It can be shown that the overall pose of the SMPL human model can be improved using machine learning post-processing, which can be reflected in a lower rotation error. To recover the occluded poses, the machine learning post-processing models should be trained on a specific action. Action recognition in the input video can help select the appropriate model that is trained on a specific action. Both frame-to-frame and predictive sequence-to-sequence machine learning are used. It is shown that both methods can improve the occluded limb error and reconstruct the occluded motion.

II. Background

HPE algorithms are divided into top-down and bottom-up approaches. Top-down methods first detect individual subjects and then estimate the pose for each subject. Bottom-up methods find all the key points in the image and then group them into individuals. They have lower accuracy than top-down methods but they are better at handling occlusion because of consideration of the joint relationships. One example of such a bottom-up method is ORPM [2] (occlusion robust pose-map) which uses the joint location redundancy that can be applied only to extremity joints to infer occlusion. Another bottom-up method is XNect [3] which encodes the joint's immediate local context in the kinematic tree to handle occlusion. Finally, there is also the depth-aware part association algorithm [4] which is robust to occlusion.

Various methods are designed to solve the occluded pose estimation problem specifically. One category of methods first finds the 3D skeleton that has some missing joints and then completes the missing joints with statistical and geometric models [3], [5]–[7]. The attention mechanism method is another approach that enforces the model to focus on non-occluded areas and results in more robustness in the final output [8]–[10]. If the input is video, it is possible to use temporal methods [11]–[19]. If the complexity of occlusion in the real world is higher than the available data, data augmentation methods [19]–[22] can solve this problem. In severe occlusion scenarios where there are little or no cues, some recent methods regress multiple plausible poses [23]–[26].

The problem of self-occlusion has also been specifically researched. For example, to reduce the ambiguity in inferring a 3D pose from a single image, both kinematic and orientation-related constraints are used [6]. This is done by projecting the 3D model onto the input image and other synthetic views, improving the ambiguity. In [27], the occlusion problem in a single image is solved by using the Euclidean Distance Matrix. In [28], a Markov random field is used to represent the occlusion relationship between human body parts in terms of occlusion state variables. The depth ordering of the body parts creates occlusion states that need to be estimated. The dataset is labelled according to how the depth ordering of body parts and self-occlusion is changing during the video. The inference is done in two separate stages: body pose inference and occlusion state inference. A new cue is used in [29] to address the problem of self-occlusion. The self-occlusion handling process uses the torso orientation as a cue. A new occlusion-aware graphical model is introduced in [30] that explicitly models both self-occlusion and other occlusions to improve robustness. The model learns the part-level occlusion relationship from data and infers the occlusion states of parts explicitly. In [31] pixel level hidden binary variables are used for self-occlusion reasoning. Some methods try to model self-occlusion holistically. In [32] self-occlusion of pedestrians is modelled in a joint shape and appearance tracking framework. In [6] self-occlusion reasoning is treated as a post-process with Twin-GP regression for 2D pose rectification.

Most of the previous research on self-occlusion is related to a single image and 2D human pose estimation [6], [27], [29]–[31] or 2D silhouette [32]. Some of the previous methods are demanding in terms of data, for example, [6] is taking advantage of multiple view training data and [27], [31] need a large training dataset. In [28] and [31] the depth ordering of body parts is required to be known beforehand. [28] is a video-based method but being part-based it is unable to estimate invisible body parts or limbs. The most widely used human body model in the previous work on self-occlusion is the body parts model [28]–[31]. Existing work on occlusion problems in SMPL-based methods [9] works for occlusion by other objects but does not perform well in self-occlusion scenarios where body parts are occluded by other body parts.

Our research focuses on the problem of self-occlusion in video-based 3D human model estimation, which is an under-researched area. We also believe this is the first work on the problem of self-occlusion using SMPL-based methods. Our approach does not require a complicated or large training dataset and only uses single-view video data. Furthermore, it is able to estimate invisible motions and body parts, unlike the previous video-based research.

III. Methods

The main aim of the research is to improve estimated human motion in challenging scenarios such as self-occlusion. We chose MotionBERT [33] as our baseline to evaluate the performance of our self-occlusion recovery methods. MotionBERT is a recent state-of-the-art 3D human pose estimation method that, given the input video, can predict several pieces of information such as 3D key points, the parametric SMPL model and the action performed by the subject. Then, a model that is trained on the specific action is chosen to predict the correct motion from the imperfect SMPL output of the human pose estimator. For better prediction of self-occluded and invisible motions, machine learning training data can be restricted to a specific action. Assuming we have an input video of arbitrary motions, the action label that is specified with action recognition is used to choose the model trained on the specific action in the video. Action segmentation is used to specify the start and end frame of the action from the video.

Figure 1 shows the process of self-occlusion correction. The video is given to the HPE block, where its output will be incorrect SMPL model prediction due to self-occlusion. A machine learning model that is trained on a dataset of incorrect and correct pose/motion pairs then predicts the correct SMPL pose/motion in which the self-occluded frames are improved.

A. Dataset Recording and Creation

The introduced post-processing method should be able to recover the 3D motion from the self-occluded videos when applied to any HPE program. For machine-learning-based post-processing, we are using pairs of incorrect SMPL (predicted by HPE) and correct SMPL (created using MoCap). This pair

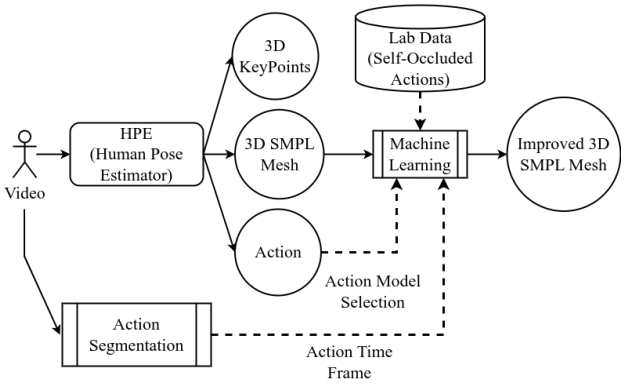


Figure 1. Human Pose Estimation (HPE) post-processing using machine learning for self-occlusion compensation. The action-specific machine learning model is chosen with the help of action recognition done by the HPE module. Action segmentation specifies the start and end frame of the action video segment within the input video.

of input and ground truth data is created by capturing synchronised monocular video and motion capture data from self-occluded motions. A total of 30 different motions with specific emphasis on self-occlusion from three different subjects were recorded. Performed actions in the data are repeated to add to the quantity and diversity of the data.

The motion capture was recorded at 120 fps and the video data frame rate was also 120 fps. The subjects were wearing motion capture suits or black outfits, which makes subject calibration and data recording easier. Since the motion capture lab walls and floor are black, we have added a green screen behind the subject to easily segment the subject from the background.

A (blurred) version of the video can be found in [34].

B. Data Preparation

The motion data that is processed in this work are the incorrectly estimated joint rotation values (affected by self-occlusion) of the SMPL human model. The SMPL model has 24 body joints and the rotational values are in axis-angle format with three rotation values. Therefore, each pose can be represented by a 1D array of length 72. A motion matrix is a 2D matrix of $n \times 72$ in which n is the number of frames recorded from MoCap and a single video camera. There are two corresponding motion matrices: one resulting from the MoCap and one from the synchronised video human pose estimation.

Suppose we have a motion sequence of length n which is down-sampled from 120 fps to 30 fps, then a 124-frame window from this sequence is chosen that is equal to around four seconds of video. This is sufficient time to complete one action. An overlapping window from the sequence with a one-second (30 frames) gap is chosen.

Each of the three subjects in the recorded dataset performed 30 different actions resulting in 90 video sequences. After dividing the data of three subjects into overlapping matrices we will obtain 10773 data matrices.

The SMPL-X model resulting from MoCap has extra information as compared to the SMPL model resulting from HPE, that needs to be removed. This includes the global position of the root joint and extra joints in the body e.g. additional finger joints.

C. Frame-based and Classic Machine Learning

The first proposed method uses a random forest (RF) model to learn the corresponding correct pose to each incorrectly predicted pose from HPE. As mentioned before, each pose is a set of joint orientations. The RF model is chosen due to being a multi-input and multi-output model suitable for this purpose. Each frame of the motion matrix is used as independent data for the RF model.

To properly compensate for the self-occluded motions, the uniqueness of correct to incorrect pose mapping can be guaranteed when having information about the action that is taking place. With the help of action recognition, the model that is trained on a specific action can be used to properly compensate for the occluded motion.

D. Predictive Sequence-based and Deep Learning

The second proposed method is a deep learning method and uses a motion sequence of poses instead of just one pose as the input data. It also creates a correspondence between a window of motion from the present to a window of motion from future frames. Two different networks with one and two LSTM layers - see Figure 2 - are used for this purpose.

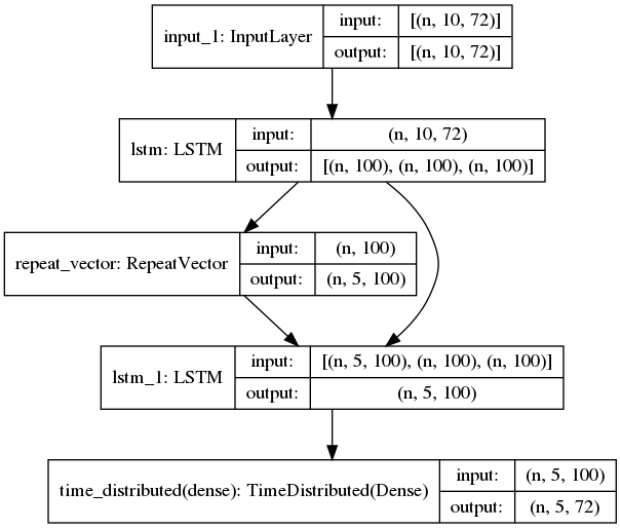


Figure 2. 1-layer LSTM AutoEncoder (AE1) for predictive pose correction with an input sequence of length 10 and output sequence of length 5. The brackets show the dimensions of the input and output data of each layer. The first dimension (n) is the batch size.

The data is first scaled between -1 and +1. To divide into overlapping windows, the length of the window from the past is 10 frames, which is mapped to a window of the future with a length of 5 frames.

As can be seen in Figure 2, the output of the network is the predicted window of future motion with a fixed length

of 5 frames. Since input/output windows are overlapping, for each frame there will be more than one predicted value from different predicted windows. The average of predicted values for each frame is computed to create the final output motion signal.

Similar to the frame-based method in section III-C, a model that is trained to recover the incorrect occluded motion can better predict future motion when it is trained on a specific action. Therefore, an additional step is needed to do action segmentation and action recognition of an arbitrary sequence of different actions and use the action-specific model to predict the output.

E. Error Measurement

This work is focused on solving the problem of self-occlusion and unnatural poses in SMPL human pose estimators. In HPE there is quantitative and qualitative comparison between the state-of-the-art and the implemented method. Since the errors are average measurements during the whole motion sequence, sometimes qualitative comparison is also needed to show how a method can improve the overall shape of the predicted pose in specific occurrences such as occlusion and unnatural poses.

The pose parameters of the SMPL model that are related to the output is a set of joint rotations in axis-angle format. Therefore, root-relative joint rotations are representative of the improvement of the overall pose in the SMPL model. The joints' 3D position error in the SMPL model can be affected by several factors such as root joint orientation or incorrect estimation of global position. The position is only representative of 3 degrees of freedom in joint movements. Therefore, in addition to the traditional mean joint position error, a new error measurement for SMPL-based human pose estimation based on quaternion differences is introduced.

IV. Results

The machine learning-based self-occlusion correction method is tested on the dataset of occluded motions collected in the lab. The orientation and position error of the output SMPL model were calculated. The data consists of 30 different actions with emphasis on self-occluded motions performed by three subjects.

The self-occluded motions from our dataset are tested on two different state-of-the-art (SOTA) methods, one of them is SMPL-based human pose estimation for occlusion [9] and the other is a recent human pose estimator [35]. When a large part of the motion is invisible or limbs are hidden from the single camera view, many of the recent HPE methods are unsuccessful. It should be noted that the proposed post-processing method can potentially find the self-occluded motion when applied to any HPE algorithm. In the demonstrated results, our method is added to the MotionBERT [33] HPE method which provides both action and 3D key points as well as the SMPL human model as the output.

There are three different post-processing methods: frame-to-frame machine learning using random forest (RF) and sequence-to-sequence machine learning using LSTM-based autoencoders with one and two layers respectively. There are three different subjects with 30 different actions, resulting in 90 different videos each doing specific actions.

The results from two different experiments are reported next. In the first experiment, the post-processing method is trained and tested on all different actions and the error of the pose is calculated - see Table I. To recover the occluded pose correctly, in the second experiment, the models are trained on specific action data. The result of self-occlusion human pose recovery is shown in the qualitative and quantitative results section (See Figure 4 and Table II).

A. Qualitative results

Figure 3 shows the results of a complex self-occluded pose: hands crossed over behind the back. The first sub-image shows the result for HPE whilst sub-image 2 shows the ground truth from the MoCap data of the test subject. The results of the post-processing methods, i.e. Random Forest and Predictive LSTM autoencoder with one and two layers, are in sub-images 3-5, respectively. The bottom sub-images 6 and 7 show the results of the HybrIK [35] and PARE [9] methods, respectively. Our proposed post-processing methods can resolve the self-occlusion problem as compared to the HPE baseline and the previous SOTA methods.

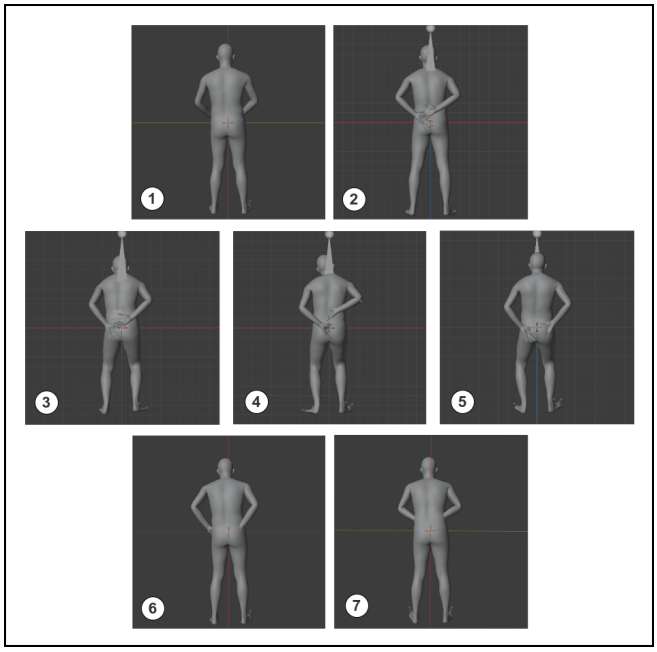


Figure 3. Self-Occlusion human pose recovery using post-processing methods, compared to the state of the art and the baseline HPE. 1: HPE, 2: Ground-Truth MoCap, 3: RF Post-Processing, 4: Predictive AutoEncoder Post-Processing with 1-Layered LSTM, 5: Predictive AutoEncoder Post-Processing with 2-layered LSTM, 6: PARE Method, 7: HybrIK Method.

B. Quantitative results

Table I shows the comparison of the average error of all joints using post-processing methods and the baseline when

using a model trained on all actions. It is shown that all post-processing methods show better rotational errors than HPE. For positional errors, the RF method is better than HPE, but the AE methods are not. The error calculations are across different subjects and actions in the test data (20% random selection of all subject-action data). The large positional errors are due to training across all actions and the predictive nature of the autoencoder model.

Figure 4 and Table II compare the errors of the HPE baseline, the HPE baseline with post-processing methods (trained on hand behind back action), and the state-of-the-art methods (PARE [9] and HybrIK [35]). It is observed that our proposed post-processing methods show better performance than the baseline and state-of-the-art methods. While the result of future motion prediction from predictive autoencoders might not be as good as current frame-to-frame RF motion prediction, they are still effective in reconstructing self-occluded motions when combined with action recognition.

Table I
AVERAGE ERRORS FOR TRAINING ACROSS ALL ACTIONS. ROTATIONAL ERRORS ARE IN 10^-3 RADIANS AND POSITIONAL ERRORS ARE IN MM. RF IS RANDOM FOREST AND AE1 AND AE2 ARE THE AUTOENCODER METHODS WITH ONE AND TWO LSTM LAYERS RESPECTIVELY.

	Joint Error	Baseline HPE	Post-Processing *RF*	*AE1*	*AE2*	SOTA *PARE*	*HBIK*
Rot.	All	467.1	186.33	262.8	258.1	480.9	535.5
	LArm	64.0	38.0	51.2	48.0	63.6	83.9
	RArm	68.0	40.0	53.2	50.3	76.5	83.2
	LLeg	52.4	27.7	48.6	47.3	59.6	62.6
	RLeg	53.0	28.0	46.3	47.4	57.8	69.1
Pos.	All	74.9	57.7	88.6	90.7	88.9	70.6
	LArm	15.6	2.9	20.2	21.1	19.3	15.2
	RArm	14.0	13.7	18.4	18.4	17.4	13.4
	LLeg	12.7	8.8	15.2	15.2	14.7	11.8
	RLeg	15.6	9.3	15.8	16.9	16.9	13.1

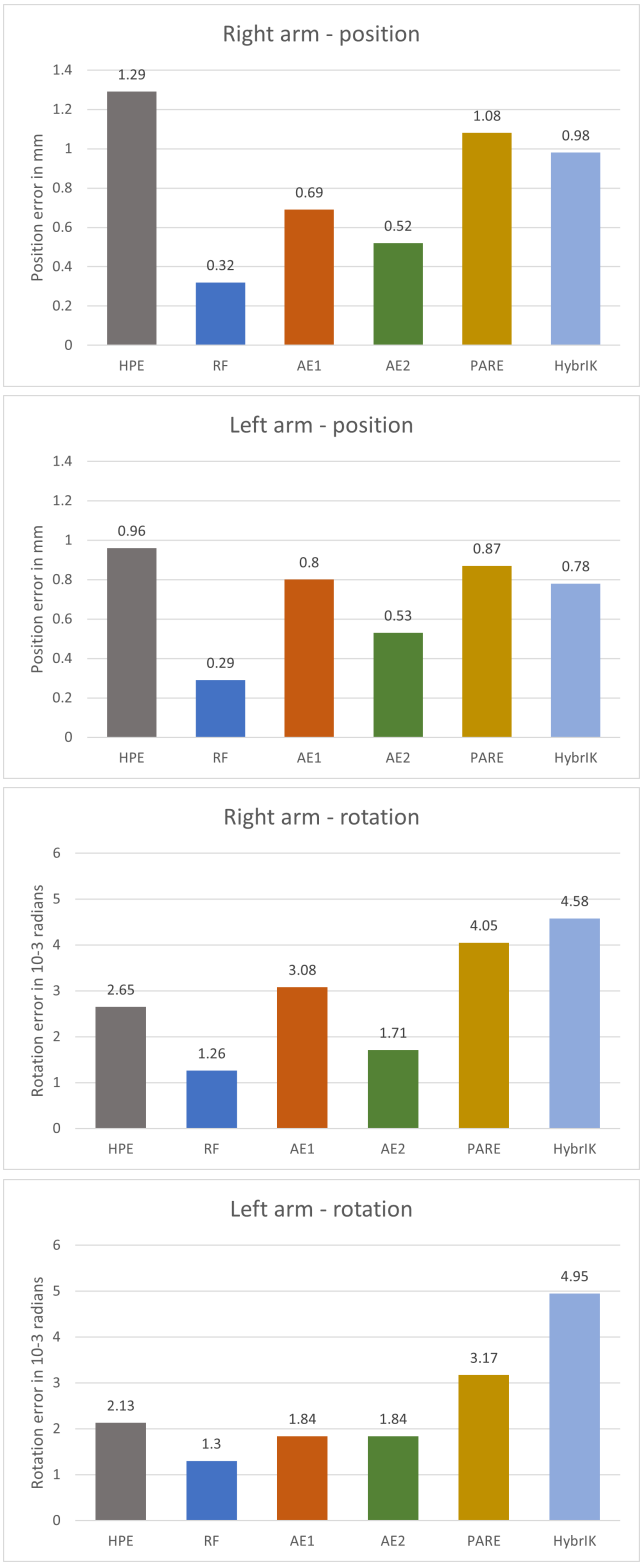
Table II
ERRORS IN MM OF POST-PROCESSING METHODS COMPARED TO THE STATE OF THE ART FOR ACTION-SPECIFIC LEARNING IN 10^-3 RADIANS FOR ROTATION AND MM FOR POSITION. ACTION: HAND BEHIND BACK - SEE ALSO FIGURE 3.

	Joint Error	Baseline HPE	Post-Processing *RF*	*AE1*	*AE2*	SOTA *PARE*	*HBIK*
Rot.	LArm	2.13	1.30	1.84	1.84	3.17	4.95
	RArm	2.65	1.26	3.08	1.71	4.05	4.58
Pos.	LArm	0.96	0.29	0.80	0.53	0.87	0.78
	RArm	1.29	0.32	0.69	0.52	1.08	0.98

V. CONCLUSION

This research focused on solving self-occlusion in SMPL-based single-video human pose estimation. Different post-processing methods were suggested to improve the result of the state-of-the-art human pose estimators and to recover the self-occluded poses. The action-specific models can be trained and selected using human action recognition depending on the input data action. Additionally, a new rotation-based error metric provides a more suitable evaluation of 3D human pose accuracy.

Figure 4. Arm joints error for action specific learning of hand behind back action - see also Figure 3 and Table II.



In future work, the quality and quantity (more subjects) of the ground truth data will be increased. Using more information from the original input video can also help improve the result. The predictive networks can also be compared with classic methods such as the Kalman filter which can do both position and orientation motion prediction, unlike our current method that does only the latter.

REFERENCES

[1] M. Loper, N. Mahmood, J. Romero, G.Pons-Moll and M.J. Black, 2023. "SMPL: A skinned multi-person linear model." In *Seminal Graphics Papers: Pushing the Boundaries*, Volume 2, pp. 851-866.

[2] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll and C. Theobalt, 2018, September. "Single-shot multi-person 3D pose estimation from monocular rgb." In *2018 International Conference on 3D Vision (3DV)*, IEEE, pp. 120-130.

[3] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.P. Seidel, H. Rhodin,G. Pons-Moll and C. Theobalt 2019. "XNect: Real-time multi-person 3D human pose estimation with a single rgb camera". *arXiv preprint arXiv:1907.00837*.

[4] J. Zhen, Q. Fang, J. Sun, W. Liu, W. Jiang, H. Bao and X. Zhou, 2020. "Smap: Single-shot multi-person absolute 3D pose estimation." In *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16, Springer International Publishing, pp. 550-566.

[5] R. De Bem, A. Arnab, S. Golodetz, M. Sapienza and P. Torr, 2018, November. "Deep fully-connected part-based models for human pose estimation." In *Asian conference on machine learning*, PMLR, pp. 327-342.

[6] I. Radwan, A. Dhall and R. Goecke, 2013. "Monocular image 3D human pose estimation under self-occlusion." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1888-1895.

[7] G. Rogez, P. Weinzaepfel and C. Schmid, 2017. "LCR-Net: Localization-classification-regression for human pose." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3433-3441.

[8] R. Gu, G. Wang and J.N. Hwang, 2021, January. "Exploring severe occlusion: Multi-person 3D pose estimation with gated convolution." In *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, pp. 8243-8250.

[9] M. Kocabas, C.H.P. Huang, O. Hilliges and M.J. Black, 2021. "PARE: Part attention regressor for 3D human body estimation." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11127-11137.

[10] L. Zhou, Y. Chen, Y. Gao, J. Wang and H. Lu, 2020. "Occlusion-aware siamese network for human pose estimation." In *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, Springer International Publishing., pp. 396-412.

[11] Y. Cheng, B. Yang, B. Wang, W. Yan and R.T. Tan, 2019. "Occlusion-aware networks for 3D human pose estimation in video." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 723-732.

[12] B. Artacho and A. Savakis, 2020. "UniPose: Unified human pose estimation in single images and videos." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7035-7044.

[13] Y. Cai, L. Ge, J. Liu, J. Cai, T.J. Cham, J. Yuan and N.M. Thalmann, 2019. "Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2272-2281.

[14] J. Liu, J. Rojas, Y. Li, Z. Liang, Y. Guan, N. Xi and H. Zhu, 2021, May. "A graph attention spatio-temporal convolutional network for 3D human pose estimation in video." In *2021 IEEE international conference on robotics and automation (ICRA)*, IEEE, pp. 3374-3380.

[15] M. Parger, C. Tang, Y. Xu, C.D. Twigg, L. Tao, Y. Li, R. Wang and M. Steinberger, 2021. "UNOC: Understanding occlusion for embodied presence in virtual reality." *IEEE Transactions on Visualization and Computer Graphics*, 28(12), pp. 4240-4251.

[16] M. Véges and A. Lőrincz, 2020. "Temporal smoothing for 3D human pose estimation and localization for occluded people." In *Neural Information Processing: 27th International Conference, ICONIP 2020*, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part I 27, Springer International Publishing, pp. 557-568.

[17] J. Wang, E. Xu, K. Xue, and L. Kidzinski, 2020. "3D pose detection in videos: Focusing on occlusion." *arXiv preprint arXiv:2006.13517*.

[18] J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang and W. Zhang, 2020. "Deep kinematics analysis for monocular 3D human pose estimation." In *Proceedings of the IEEE/CVF Conference on computer vision and Pattern recognition*, pp. 899-908.

[19] S. Park and J. Park, 2021. "Localizing human keypoints beyond the bounding box." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1602-1611.

[20] Y. Cheng, B. Yang, B. Wang and R.T. Tan, 2020, April. "3D human pose estimation using spatio-temporal networks with explicit occlusion training." In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 07, pp. 10631-10638.

[21] X. Peng, Z. Tang, F. Yang, R.S. Feris and D. Metaxas, 2018. "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2226-2234.

[22] Y. Xu, W. Wang, T. Liu, X. Liu, J. Xie and S.C. Zhu, 2021. "Monocular 3D pose estimation via pose grammar and data augmentation." *IEEE transactions on pattern analysis and machine intelligence*, 44(10), pp. 6327-6344.

[23] B. Biggs, D. Novotny, S. Ehrhardt, H. Joo, B. Graham and A. Vedaldi, 2020. "3D multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data." *Advances in neural information processing systems*, 33, pp. 20496-20507.

[24] E. Jahangiri and A.L. Yuille, 2017. "Generating multiple diverse hypotheses for human 3D pose consistent with 2d joint detections." In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 805-814.

[25] C. Li and G.H. Lee, 2019. "Generating multiple hypotheses for 3D human pose estimation with mixture density network." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9887-9895.

[26] T. Wehrbein, M. Rudolph, B. Rosenhahn and B. Wandt, 2021. "Probabilistic monocular 3D human pose estimation with normalizing flows." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11199-11208.

[27] X. Guo and Y. Dai, 2018, August. "Occluded joints recovery in 3D human pose estimation based on distance matrix." In *2018 24th international conference on pattern recognition (ICPR)*, IEEE, pp. 1325-1330.

[28] N.G. Cho, A.L. Yuille and S.W. Lee, 2013. "Adaptive occlusion state estimation for human pose tracking under self-occlusions." *Pattern Recognition*, 46(3), pp. 649-661.

[29] Y. Yu, B. Yang, and P.C. Yuen, 2016, August. "Torso orientation: A new clue for occlusion-aware human pose estimation." In *2016 24th European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 908-912.

[30] L. Fu, J. Zhang and K. Huang, 2015. "Beyond tree structure models: A new occlusion aware graphical model for human pose estimation." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1976-1984.

[31] L. Sigal and M.J. Black, 2006, June. "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation." In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2, IEEE, pp. 2041-2048.

[32] Y. Yang and G. Sundaramoorthi, 2013. "Modeling self-occlusions in dynamic shape and appearance tracking." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 201-208.

[33] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu and Y. Wang, 2023. "MotionBERT: A unified perspective on learning human motion representations." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15085-15099.

[34] https://youtu.be/LSPRKDgtcAQ

[35] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang and C. Lu, 2021. "HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3383-3393.