

Post-Data Augmentation to Improve Deep Pose Estimation of Extreme and Wild Motions

Kohei Toyoda*
The University of Tokyo

Michinari Kono†
The University of Tokyo

Jun Rekimoto‡
The University of Tokyo
Sony Computer Science
Laboratories, Inc.

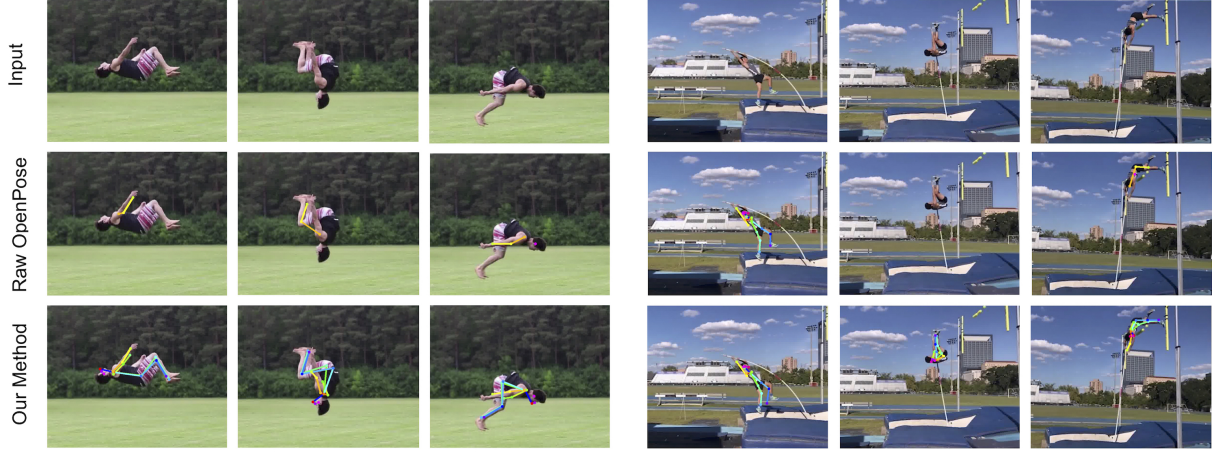


Figure 1: Applying data augmentation approach for the estimation/generation phase in deep pose estimation can improve the quality of extreme/wild motion pose estimation. Our approach can be used with pre-trained models, and without new training with self-collected dataset. This figure shows results compared with using raw OpenPose [1].

ABSTRACT

Contributions of recent deep-neural-network (DNN) based techniques have been playing a significant role in human-computer interaction (HCI) and user interface (UI) domains. One of the commonly used DNNs is human pose estimation. This kind of technique is widely used for motion capturing of humans, and to generate or modify virtual avatars. However, in order to gain accuracy and to use such systems, large and precise datasets are required for the machine learning (ML) procedure. This can be especially difficult for extreme/wild motions such as acrobatic movements or motions in specific sports, which are difficult to estimate in typically provided training models. In addition, training may take a long duration, and will require a high-grade GPU for sufficient speed. To address these issues, we propose a method to improve the pose estimation accuracy for extreme/wild motions by using pre-trained models, i.e., without performing the training procedure by yourselves. We assume our method to encourage usage of these DNN techniques for users in application areas that are out of the ML field, and to help users without high-end computers to apply them for personal and end use cases.

Index Terms: Computing methodologies—Motion capture; Computing methodologies—Neural networks; Human-centered computing—Human computer interaction (HCI)

*e-mail: toyoda-kohei648@g.ecc.u-tokyo.ac.jp

†e-mail: mchono@acm.org

‡e-mail: rekimoto@acm.org

1 INTRODUCTION

Measuring human pose is an important topic for analyzing the motion of the human body, which has been studied and applied to various research fields, such as sports science and human-computer interaction (HCI). Typical approaches for motion capturing were to use optical or mechanical sensors fixed or worn by the user. However, the recent growth of the deep-neural-network (DNN) techniques have presented remarkable results for human pose estimation by using simple monocular cameras that can retrieve 2D poses [1, 12, 14] or even 3D poses [7, 8]. These approaches have several benefits, for example, they do not require camera calibration and can be obtained by single RGB images.

In DNN techniques, they usually require to have large datasets that can be used for training models, and online available data is used (e.g., [6]). Unfortunately, using such data sometimes find difficulty to be applied for rare or specific postures, such as extreme or wild motions. In order to solve this problem, a typical solution is to gather data by yourselves by using other motion capturing tools for reference data, however, there are still difficulties to gather data for some motions. For example, gathering data for swimming may require an expensive or special environment for motion capturing. Instead of gathering such data, Peng et al. [11] suggested that rotation data augmentation can improve the performance of the pose estimation for *in the wild* images. This solution may improve the quality of pose estimation for extreme and wild motions.

However, this may result to another problem, where the training data becomes very large. Training with large dataset may require much more time for training and require high-grade GPUs. Therefore, it may be difficult for some *users* to use the DNN pose estimation by themselves. For such *users*, they may find benefit from just using the pre-trained model that is often provided by the developers. What we mean by saying *users* here, are people working out of the

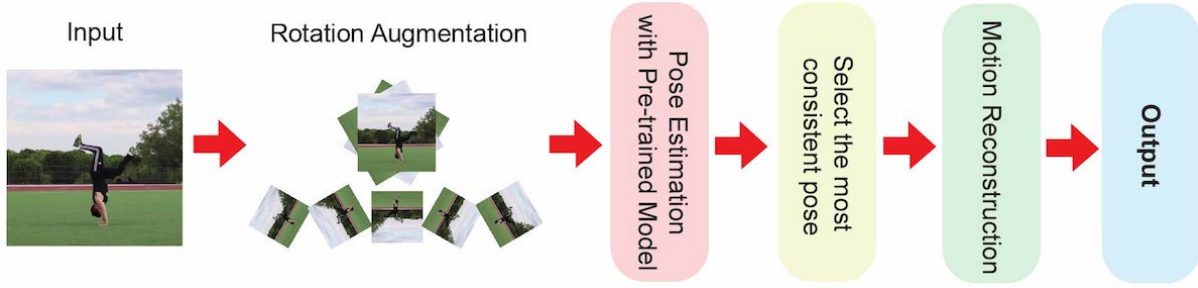


Figure 2: The overview of our approach. Instead of augmenting data for training, we apply augmentation for the estimation process. Pose estimation is applied multiple times for each frame, and then we select the best consistent pose from the set. A simple reconstruction is applied to get the final output.

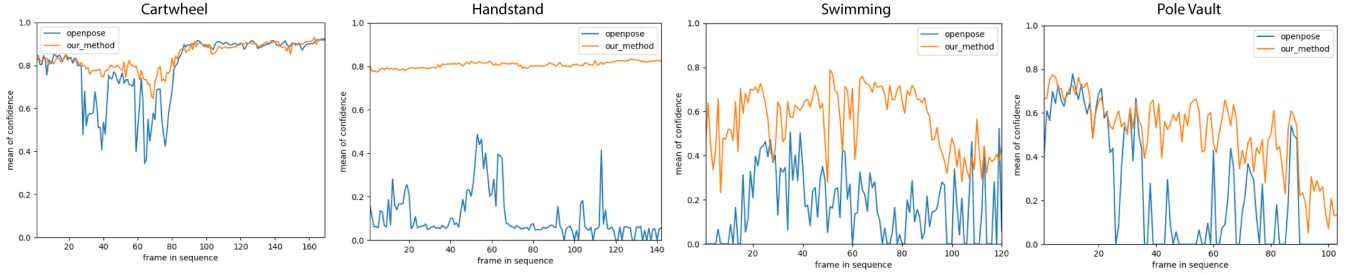


Figure 3: Confidence levels of some example video sequences. Note that the confidence do not exactly present the accuracy, but is a simple reference for the quality of prediction. The confidence level presented here is the output retrieved from OpenPose, with and without augmentation with our method.

ML field, who do not develop DNN by themselves, but use them for developing applications.

DNN pose estimation techniques are used for various applications, for both realtime and non-realtime use cases. Move Mirror¹ allows you to find images with the similar pose with the input. Other applications include use cases for entertainment purposes [5, 13] and sport analysis [2, 3, 9]. Some even focus on analyzing hands or facial expressions for other purposes [4, 10]. We can see that pose estimation is important for the users, and do not always have to be available for realtime use.

In this paper, we propose a simple but effective method to improve the pose estimation for extreme/wild motion videos, without gathering new reference data, and by just using the provided pre-trained model. Referring to the approach of Peng et al. [11], we apply pose estimation for videos rotated multiple times for each frame to find a confident pose, i.e., we apply a kind of data augmentation technique at the estimation process. As a proof-of-concept, we present example results using OpenPose [1] with *in the wild* videos and some videos recorded in specific activities (Figure 1).

2 METHOD

The approach of our method is shown in Figure 2. First, we apply OpenPose at various rotations to get predictions obtained from various angles, which is expressed by equation 1 and 2.

$${}^k j_t^\theta = {}^k j_t' R^{-1} \quad (1)$$

$$J_t^\theta = \{{}^k j_t^\theta\} \quad (2)$$

${}^k j_t'$ is the predicted k th joint at frame t when the image is rotated θ , and ${}^k j_t^\theta$ is the joint fixed to the original coordinates. R is the

¹<https://experiments.withgoogle.com/collection/ai/move-mirror/view>

rotation matrix for a 2D rotation at θ , and θ takes values at $[0^\circ, 360^\circ]$ sampled every d degrees. d is 10° in our current case. J_t^θ is sets of estimated joints when the image is rotated θ . Therefore, J_t^θ can be understood as a pose that is predicted at a certain rotation but is fixed to the original image coordinates and consists of a set of poses as $\{{}^k j_t^\theta\}$. When $t = 1$, J_1^θ takes the pose that satisfies objective function 3, where \bar{x}_t^θ is the mean of confidences of J_t^θ . We then select the ideal J_t^θ from the set, by using objective function 4. We find the top 5 poses for objective 4, and then use objective 3 to select the best matching pose. In addition, we defined a threshold for objective 4 as 500, and when all the calculation exceeds the threshold, we used objective 3 to find the best pose instead. This approach was taken because we experimentally found that simply selecting the pose with the best confidence level did not always present the correct pose.

$$\max \bar{x}_t^\theta \quad (3)$$

$$\min \sum_k \|{}^k j_t^\theta - {}^k j_{t-1}^p\|_2 \quad (4)$$

Here, p is the final reconstructed pose that is obtained by equation 5 and 6.

$$J_t^p = w J_t^\theta + (1 - w) J_{t-1}^p \quad (5)$$

$$J_t^p = \{{}^k j_t^p\} \quad (6)$$

w represents the weight for the current frame and is set to $w = 0.8$. J_t^p is used for representing the pose at every t . Note that for the procedures that are taken when selecting the best joints, we do not consider the joints around the head.

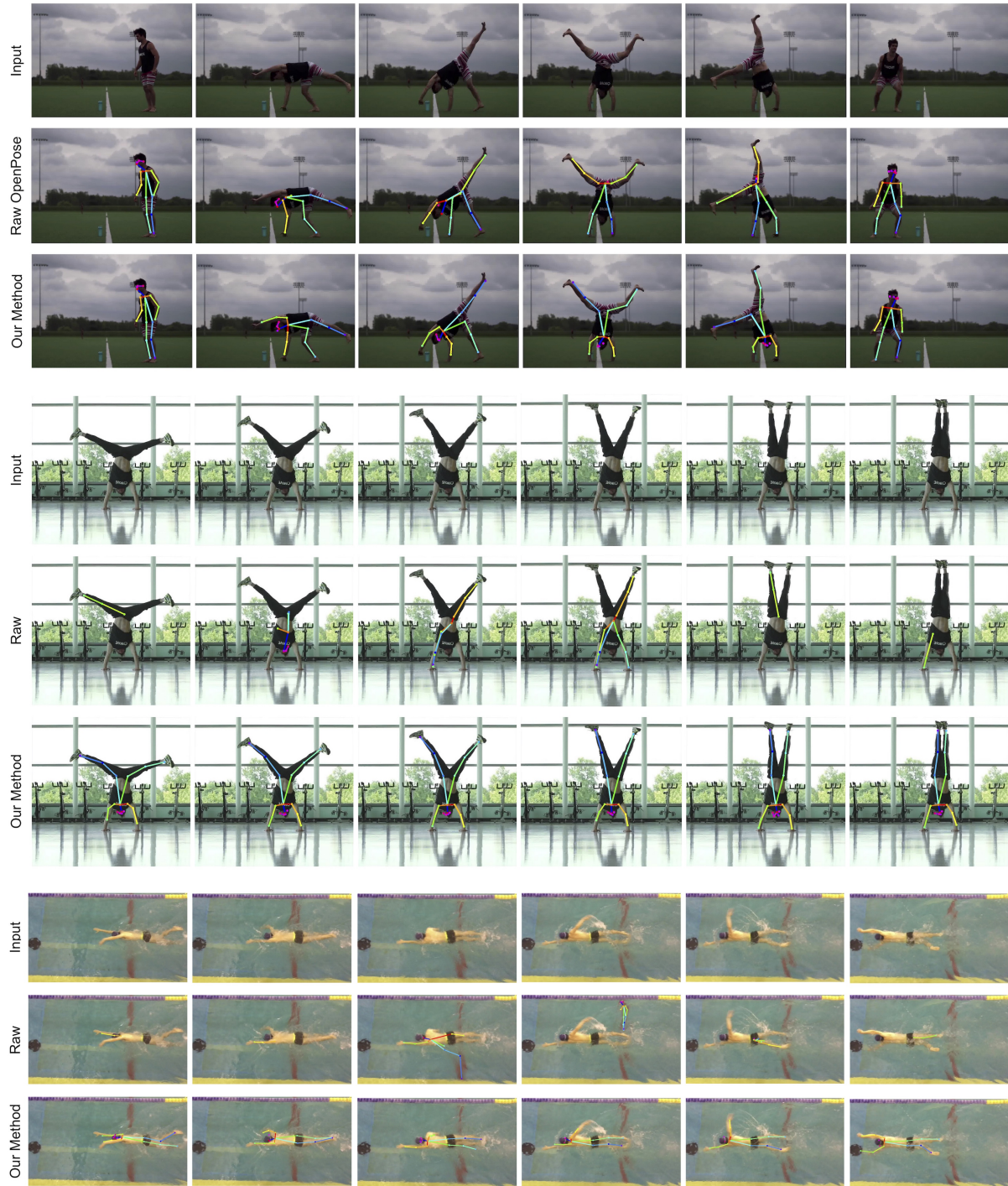


Figure 4: Example result of pose estimation applied to a video sequence. The top row is the input, the middle is the result obtained by using raw OpenPose, and the bottom is the result obtained by using our method.

3 RESULTS

We applied our method for various video sequences, and the followings are some examples.

- Cartwheel: A person performs a cartwheel, and jumps vertically at the end.
- Handstand: A person performs a handstand, and moves his legs.
- Swimming: A person swims across the camera settled above the swimming pool.
- Pole Vault: A person performs a pole vault.

Figure 3 displays confidence levels of some video sequences that were tested. We can observe that the overall confidence level tends to be higher by using our method. However, since we work on typical body models, it is still difficult to estimate poses that exceed the constraints of such models. We can also observe that the confidence level sometimes becomes lower than the results of using raw OpenPose. This happens when the predicted pose has fine confidence, however, do not match the pose from the prior frame, and therefore exceeds the threshold of objective 4.

Figure 1 and 4 presents some visual results of pose estimation using our method for videos with extreme/wild motions. We can see that using OpenPose with its default usage can suffer to estimate sufficient results especially when the human body is sideways or upside down. Our method successfully allows estimating a better pose throughout the motion. However, we still observed some difficulties for estimating poses of pole vault sequences. This happens when the body curls up, or when the pole disturbs the pose estimation.

Since our interest was to work on extreme/wild motions that do not have open source dataset provided (*in the wild scenes*), we currently have not conducted a quantitative evaluation for the accuracy. In the future, we plan to collect data for these extreme/wild motions with motion captures, and to evaluate our method against other methods, including training models created with the data.

4 DISCUSSION AND FUTURE WORK

Our results suggest that our method can improve the quality of deep pose estimation, however, still has some limitations.

Since our work applies pose estimation multiple times for each frames, it requires more time for the whole pose estimation procedure. Therefore, we can say that there is a trade-off between time and accuracy. We must note that our method cannot be applied for use cases that require realtime analysis. In addition, while the strength of OpenPose is where it can be applied for multi-person simultaneously, our current method focuses on a single person.

For future work, we will apply our approach to 3D pose estimation as well. We will also investigate other data augmentation approaches to be used for our method. Currently, we apply the pose estimation for all considerable rotation of the frame. However, we assume that the computation cost can be minimizing by referring to the selected rotation angle of adjunct frames, and to limit the rotation angle of the frame so that we can decrease the number of the pose estimation times applied for each frame. Figure 5 presents the θ value of a cartwheel sequence. In this sequence, a person starts standing, and then performs a cartwheel, and finally when the cartwheel is completed, he performs a small jump vertically. During the cartwheel, the θ value gradually decreases following the progress of the cartwheel. When the person is standing or jumping, θ takes values near the original angle (typically around 0–30° and 330–360°). We may observe that pose estimation can perform better when the head is in the upper area of the image and the legs are at the bottom. Therefore, we may possibly estimate a range of θ , which can make a good estimation.

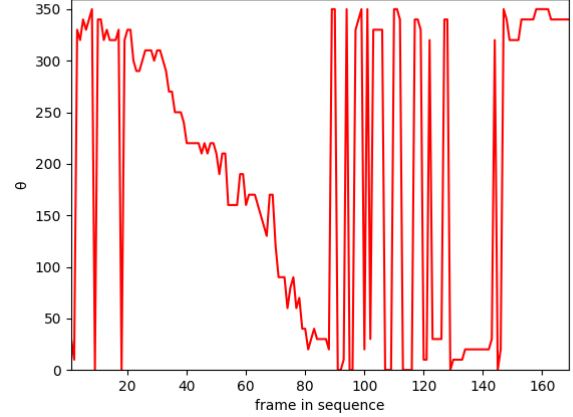


Figure 5: Plot of the θ value of a video sequence (cartwheel).

We also expect our approach of *post-data augmentation* can be applied for other application domains, not limited to pose estimation. The approach where we adapt the input data and to help the training model to perform the prediction more easily can improve the overall performance of estimation, and can be effective for various models. We assume that the approach can be expanded with other data augmentation methods, not limited to rotation augmentation.

Our motivation for this work is to encourage researchers in human augmentation and HCI to use DNN based pose estimation methods for various sports/activities with irregular motions. Introducing a method that does not require data collection and training on your own, will hopefully improve the acceptability and reduce the obstacles for applying such methods.

5 CONCLUSION

This paper introduces a method to improve pose estimation with DNN, but with a post-data augmentation approach that can be applied for pre-trained models. When using traditional methods, it can be difficult to obtain accurate poses from extreme/wild motions. Our method augments the input data with rotation augmentation, and use pose estimation method multiple times for every frame. We then select the most consistent pose, followed by a motion reconstruction for smoothing. The results show that our method can improve the overall quality of pose estimation for videos where people take irregular poses, such as when performing acrobatic motions. We expect our method to be used for applications where collecting datasets for the activity are difficult, and do not require real-time estimation. This approach can be considered as a *post-data augmentation* method, which can be used after the training process.

ACKNOWLEDGMENTS

The work is supported by ISID Technosolutions, Ltd.

REFERENCES

- [1] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1611.08050, 2016.
- [2] H. Fani, A. Mirlahi, H. Hosseini, and R. Herperst. Swim stroke analytic: Front crawl pulling pose classification. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 4068–4072, Oct 2018. doi: 10.1109/ICIP.2018.8451756
- [3] P. Felsen and P. Lucey. Body shots: Analyzing shooting styles in the nba using body pose. In *MIT Sloan Sports Analytics Conference*, 2017.
- [4] K. Fujii, P. Marian, D. Clark, Y. Okamoto, and J. Rekimoto. Sync class: Visualization system for in-class student synchronization. In

- Proceedings of the 9th Augmented Human International Conference, AH '18*, pp. 12:1–12:8. ACM, New York, NY, USA, 2018. doi: 10.1145/3174910.3174927
- [5] D.-H. Hwang and H. Koike. Parapara: Synthesizing pseudo-2.5d content from monocular videos for mixed reality. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, CHI EA '18*, pp. LBW608:1–LBW608:6. ACM, New York, NY, USA, 2018. doi: 10.1145/3170427.3188596
 - [6] 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*, 36(7):1325–1339, jul 2014.
 - [7] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. *CoRR*, abs/1712.06584, 2017.
 - [8] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. *CoRR*, abs/1705.03098, 2017.
 - [9] M. Nakai, Y. Tsunoda, H. Hayashi, and H. Murakoshi. Prediction of basketball free throw shooting by openpose. In *Proceedings of Fifth International Workshop on Skill Science, SKL '18*, 2018.
 - [10] T. Okumura, S. Urabe, K. Inoue, and M. Yoshioka. Cooking activities recognition in egocentric videos using hand shape feature with openpose. In *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management, CEA/MADiMa '18*, pp. 42–45. ACM, New York, NY, USA, 2018. doi: 10.1145/3230519.3230591
 - [11] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine. Sfv: Reinforcement learning of physical skills from videos. In *SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia '18*, pp. 178:1–178:14. ACM, New York, NY, USA, 2018. doi: 10.1145/3272127.3275014
 - [12] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. *CoRR*, abs/1312.4659, 2013.
 - [13] S. Tsuchida, S. Fukayama, and M. Goto. Automatic system for editing dance videos recorded using multiple cameras. In A. D. Cheok, M. Inami, and T. Romão, eds., *Advances in Computer Entertainment Technology*, pp. 671–688. Springer International Publishing, Cham, 2018.
 - [14] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *CoRR*, abs/1602.00134, 2016.