Supplementary - LInKs "Lifting Independent Keypoints" - Partial Pose Lifting for Occlusion Handling with Improved Accuracy in 2D-3D Human Pose Estimation

Anonymous WACV Algorithms Track submission

Paper ID 1272

1. Evaluating the Frequency of Complete 2D Pose Detection

In our paper, we discussed the two approaches employed in 3D human pose estimation (HPE): direct mapping from an RGB image and a two-stage process involving the acquisition of a 2D pose followed by its conversion to a 3D pose. Previous studies have indicated that the rationale behind adopting the two-stage approach is the exploitation of existing 2D pose estimation systems [2, 3, 5–7]. However, there is currently a lack of empirical evidence supporting the effectiveness of 2D detectors in detecting the full 2D pose for accurate 2D-3D pose conversion. Instead, it is more common for 2D detection papers to report the precision, recall, and object keypoint similarity (OKS) for all poses detected regardless of if all keypoints were detected for that particular pose. Furthermore, the human body exhibits a significant degree of malleability during motion, enabling a wide range of actions such as crouching, sitting, and crossing arms, all of which contribute to self-occlusion when viewed from a single camera's perspective. Consequently, when employing current 2D-3D lifting models with off-the-shelf 2D human pose detectors we are unable to provide an answer regarding the frequency at which we can successfully lift a complete 3D pose. To highlight the severity of self-occlusion, we conducted experiments using OpenPose (a popular 2D detection model) [1] on the Human3.6M dataset (H36M) [4]. It is important to note that H36M dataset captures videos in controlled conditions where subjects perform actions individually, in well-lit rooms, wearing fitted clothing, with self-occlusion being the primary factor affecting keypoint detection, along with a small stool for sitting and eating action. The full results of this experiment can be seen in Table 1. Our results showed that on average a complete 2D pose is recovered in 45.1% of all frames from both front and rear cameras, meaning in a best-case controlled scenario prior 2D-3D lifting approaches would function less than half of the time. In contrast, if we observe how often a partial 2D pose is detected this increase to 88.6%. Therefore utilising our approach would increase the number of frames where a full 3D pose could be retrieved by 96.5%. This is especially noticeable in difficult actions such as sitting down (SitD. in the table), where a 3D pose estimate could be retrieved in only 4.6% of frames using prior methods with an off-the-shelf 2D detector, which increases to 45.2% when using our approach. Moreover in scenarios such as greetings, our approach would be able to retrieve the full pose in 100% of frames from all viewpoints whereas a full pose lifting approach would only function in 51.3% of the frames. We hope that these findings help highlight the need and applicability of more robust 2D-3D lifting approaches which are able to handle occlusion.

References

[1] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(01):172–186, jan 2021. 1

[2] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5707–5717, 2019. 1

[3] Dylan Drover, Rohith M. V, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 78–94, Cham, 2019. Springer International Publishing. 1

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian 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. 1

[5] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose esti-

Action	Full Pose			Partial		
	Front Cams.	Rear Cams.	Avg.	Front Cams.	Rear Cams.	Avg.
Direct.	62.0%	27.1%	44.6%	99.8%	99.8%	99.8%
Discuss	69.1%	53.1%	61.1%	99.9%	99.9%	99.9%
Eat	69.1%	26.9%	48.0%	98.8%	84.2%	91.5%
Greet	64.0%	38.6%	51.3%	100%	100%	100%
Phone	57.9%	26.4%	42.2%	96.4%	73.7%	85.0%
Photo	67.2%	27.1%	47.1%	98.4%	98.7%	98.6%
Posing	75.9%	60.9%	68.4%	99.1%	99.4%	99.3%
Purchase	68.4%	24.5%	46.4%	90.0%	94.0%	92.0%
Sit	47.3%	1.2%	24.2%	91.4%	51.6%	71.5%
SitD.	5.6%	3.7%	4.6%	54.2%	36.2%	45.2%
Smoke	47.2%	43.6%	45.3%	85.9%	81.5%	83.6%
Wait	65.2%	55.3%	45.3%	99.6%	99.2%	99.4%
Walk	53.3%	57.6%	55.5%	98.9%	99.5%	99.2%
WalkD.	48.5%	45.0%	46.7%	93.4%	96.1%	94.7%
WalkT.	58.4%	60.0%	59.2%	98.9%	99.0%	99.0%
Avg.	54.8%	35.5%	45.1%	92.4%	84.8%	88.6%

Table 1. Showing the percentage of frames where either the full pose (denoted as "Full Pose") or partial body segments (denoted as "Partial"), were detected for each action in the Human3.6M dataset. The results were obtained by employing the MPI model of OpenPose on randomly chosen videos from the Human3.6M (H36M) dataset (two videos per action), using both front and rear cameras. The MPI model was chosen as it is consistent with the 2D keypoints provided in the H36M dataset, and thus representative of the model used in end-to-end 2D-3D lifting.

mation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017. 1

[6] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[7] Zhenbo Yu, Bingbing Ni, Jingwei Xu, Junjie Wang, Chenglong Zhao, and Wenjun Zhang. Towards alleviating the modeling ambiguity of unsupervised monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8651–8660, October 2021. 1