Real-time omnidirectional 3D multi-person human pose estimation with occlusion handling

Anonymous Author(s)
Submission Id: 12

ABSTRACT

This paper presents a real-time 3D multi-person human pose estimation system with a 360° panoramic camera and mmWave radar sensors. Our proposed system includes several contributions including, camera and radar calibrations, improved matching of camera and radar-detected people, and an advanced 3D pose estimation algorithm, designed to handle occlusion challenges. The system addresses both the depth and scale ambiguity problems by employing a lightweight 2D-3D pose lifting algorithm that is able to work in real-time while exhibiting accurate performance in both indoor and outdoor environments which offers both an affordable and scalable solution. Notably, our system's time complexity remains nearly constant irrespective of the number of detected individuals, achieving a frame rate of approximately 7-8 fps on a laptop with a commercial grade GPU.

CCS CONCEPTS

• **Computing methodologies** → **3D imaging**.

KEYWORDS

3D Human Pose Estimation, Occlusion Handling, Real Time System, Omnidirectional Camera, Radar Sensors

ACM Reference Format:

Anonymous Author(s). 2023. Real-time omnidirectional 3D multi-person human pose estimation with occlusion handling. In *Proceedings of the 20th ACM SIGGRAPH European Conference on Visual Media Production (CVMP 2023).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Monocular 3D human pose estimation (HPE) using RGB cameras is a complicated task with various applications such as security, 3D animation and physical therapy [???]. However, the retrieval of an accurate 3D pose from a single image in real-time is hampered due to the depth and scale ambiguity of people within the detected scene. Prior methods therefore, combined RGB cameras with depth sensors, like LIDAR (laser-based) or Kinect (IR-based) [??]. Despite being relatively effective, the high cost of laser-based sensors and the limited performance of IR-based sensors in outdoor environments due to sunlight interference hinders the large-scale deployment of these approaches. Moreover, solely relying on RGB cameras results in local coordinates without global localisation, limiting their effectiveness [??]. To address this, radar-based methods have emerged as a cost-effective solution for both indoor and outdoor scenarios. Combining omnidirectional cameras with mmWave radars offers a comprehensive 360-degree panoramic view of the surroundings, capturing full-body poses in a single frame, improving depth estimation and resolving the depth ambiguity problem in applying monocular depth estimation to practical applications [??]. In HPE applications, however, the occlusion of people remains a common challenge that impacts accuracy. Occlusions occur when body parts are hidden or obscured by objects such as other people, or self-occlusion due to body posture. Although studies have sought to tackle this problem in the 3D HPE field, proposing techniques to handle it effectively [???], occlusion occurrence varies based on the application and environment, being more prevalent in crowded or complex scenes. Therefore, in our system designed for multi-person HPE handling within 360° scenes, occlusion handling is a crucial task. For the accurate 3D pose estimation in the global coordinate, we use an unsupervised 2D-3D lifting algorithm for accurate local 3D pose estimation relative to the camera coordinate system and a localisation pipeline to accurately register the estimated 3D pose to the global coordinate system using the radar sensor data. The proposed system was built on two pieces of preliminary work. The first is [?], which is one of the first approaches to utilise an omnidirectional camera and multiple mmWave radars for real-time multi-person 3D HPE. The second is [?] which is the first unsupervised 2D-3D lifting algorithm to handle occlusion handling in 3D space which dramatically reduces the error. This paper presents significant enhancements to the system in [?], including radar and camera calibration and an improved matching algorithm. These updates have resulted in substantial improvements in precision, accuracy, and the ability to handle occlusions effectively.

The remainder of this paper is organised as follows: Section 2 provides an overview of related works, Section 3 introduces the proposed system, Section 4 presents the experimental analysis, and Section 5 concludes the paper. The full source code and supplementary material are made available to download: (URL hidden for double-blind review).

2 RELATED WORK

2.1 3D pose estimation from 2D keypoints

Existing methods of 3D HPE from 2D keypoints can be broadly categorised into supervised, weakly-supervised and unsupervised techniques. Among them, the following are supervised methods: Martinez [?] proposed a simple six-layer feed-forward neural network that achieved remarkable results in 3D pose estimation tasks, except for seated poses. Ludlow [?] improved upon Martinez's work by introducing three interventions that led to enhancements in 3D reconstructions for both general and seated poses. Whereas, Chen and Ramanan [?] and Yang et al. [?] utilised exemplar-based approaches, which involve large dictionaries and nearest-neighbour searches to determine the optimal 3D pose.

CVMP 2023, 30 Nov.–1 Dec., London, UK
2023. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Weakly-supervised algorithms fall between fully supervised and unsupervised models in terms of their required level of supervision. They utilize augmented 3D data or unpaired 2D-3D data to learn human body priors, without relying on explicit 2D-3D correspondences. Pavlakos et al [?] and Ronchi et al. [?] incorporated ordinal depth information such as the right wrist being behind the right elbow. Wandt and Rosenhahn [?] introduced a weakly-supervised adversarial method using kinematic chains, while Yang et al. [?] addressed 2D pose lifting without ground truth data, and Zhou et al. [?] employed transfer learning. Drover et al. [?] explored self-consistency but found the need for a 2D critic network to improve accuracy.

Unsupervised 2D-3D lifting algorithms do not use labelled 3D pose data for training. They utilise large-scale 2D annotated datasets and geometric constraints to infer 3D pose information. Pavlakos et al. [?] employed a coarse-to-fine architecture for 3D joint prediction from 2D keypoints. Tekin et al. [?] proposed a structured prediction framework that jointly estimates 2D keypoints and 3D pose. Tome et al. [?] introduced an end-to-end framework using a convolutional autoencoder to lift 2D keypoints to 3D space. Lastly, [?] approach involved a two-step process: first, transforming the occluded 2D pose into the 3D domain, and then reconstructing the missing occluded parts, arguing that it is more suitable to complete the occluded pose in 3D rather than 2D space.

In our preliminary work [?] Ludlow's approach with the Martinez [?] normalization technique was used. However, in this work, we adopted an unsupervised model proposed in our preliminary work [?]. This adaptation demonstrated the capability to handle occlusion and yielded similar results in our experiments. Furthermore, in comparison to supervised or weakly-supervised models, unsupervised models exhibit superior generalisation capabilities to new and unseen poses and environments - a crucial aspect for our application, considering the absence of specialised 360° lifting algorithms and datasets to train with.

2.2 mmWave Radar-based 3D HPE

The utilisation of mmWave radar technology for 3D HPE is challenging due to sparse point cloud data and the scarcity of labelled mmWave data. Nonetheless, several projects have explored its potential for estimating 3D human poses. For instance, Sengupta's mmPose [?] employed two radar devices and a forked-CNN architecture to process mmWave point cloud data and estimate human poses in 3D. Xue's mmMesh [?] constructed human meshes directly from mmWave point clouds using a human shape model to improve the prediction accuracy with sparse data. An's [?] proposed mmWave-based rehabilitation system estimates 3D human poses by processing sorted mmWave point clouds through matrix transformations prior to feeding them into a CNN architecture. However, these approaches do not address the scarcity of labelled mmWave data and struggle to achieve accurate multi-person 3D HPE.

3 PROPOSED SYSTEM

3.1 System overview

The core of our method revolves around transforming 2D body keypoints detected by OpenPose [?] in the image space into 3D keypoints in a global coordinate space through radar sensing data. As depicted in Figure 1, the system proceeds through several stages, including data fetching, body keypoints localisation, people localisation, matching, 2D-3D lifting, and 3D coordinate estimation.

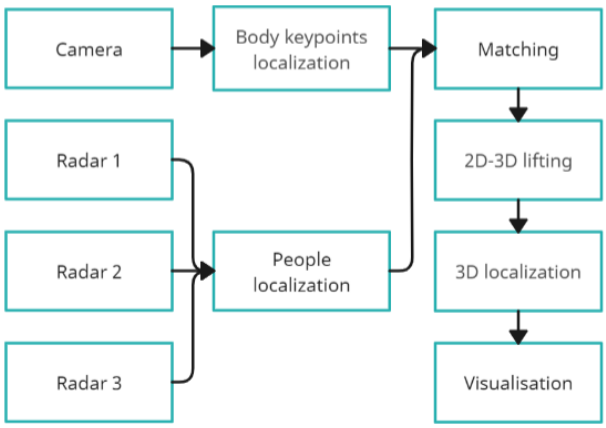


Figure 1: System pipeline.

3.2 Sensor calibration

To achieve accurate people localisation and 3D human pose estimation (HPE), the camera underwent calibration using the Zhang method [?]. Similarly, the radars were calibrated using the affine transform. For each radar's $(\dot{x}, \dot{z})$ direction, the affine transform was obtained using the LM algorithm. To perform the calibration, multiple radar readings were collected by placing an individual at multiple known radar coordinates spaced 0.5m apart. These readings were recorded for several seconds, and the average value was then compared to the correct location, allowing for precise estimation of the affine transform.

3.3 Data fetching, body keypoints and people localisation in the local coordinate

The data fetching process follows the methodology established in previous work [?]. It involved acquiring sensor data, including video frames and radar data, while ensuring proper synchronisation. The camera data is retrieved in the main thread, while radar data is concurrently fetched through separate threads. Synchronisation is guaranteed by the camera thread, which signalled the radar threads to add their data to a shared queue, ensuring synchronisation. For body keypoints localisation, the 2D Cartesian coordinates of human-body key points are estimated within the video frame, leveraging 15 keypoints $(\mathbf{x}, \mathbf{y}, \mathbf{k}) \in R^{15}$ derived from OpenPose's BODY_25 model [?]. In the simultaneous stage of people localisation, the radar sensor data is subjected to processing using Texas Instruments' (TI) people counting algorithm [?]. This enables the extraction of $(\dot{x}, \dot{z})$ coordinates corresponding to the detected individuals in each

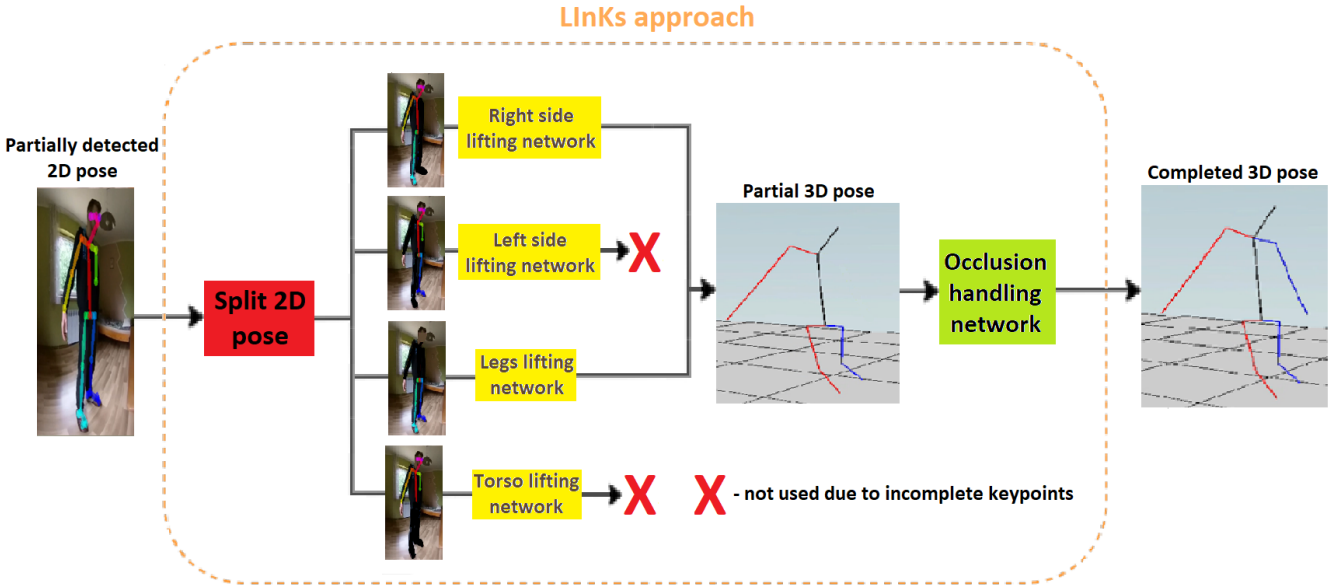


Figure 2: Overview of the lifting and occlusion handling of the LInKs [?] approach. When an image is occluded, a partially detected 2D skeleton is generated. This skeleton is divided into its distinct components: torso, legs, and left and right-hand keypoints. These parts are then individually processed by specific lifting networks, and their outputs are aggregated to produce a partial 3D pose. This partial 3D pose is subsequently input to an occlusion handling network, which predicts the missing keypoints to achieve a complete 3D pose reconstruction. In the given scenario, the occlusion affects the left arm, leading to the exclusion of the left-hand side and torso lifting networks due to incomplete 2D keypoint data. Diagram adapted from [?].

radar coordinate system. The coordinates are then subsequently transformed into the common coordinate system, and employed to calculate the distance of a person away from the camera. The distance information plays a crucial role in normalising the inputs of the lifting algorithm, thereby ensuring a consistent size for the estimated poses, irrespective of the person's distance from the camera. This improvement is significant, as the previous approach [?] suffered from a problem where the size of the reconstructed poses decreased as individuals moved farther away from the camera.

3.4 Matching Camera and Radar-Detected Individuals

Matching is a critical step in associating camera-detected individuals with their corresponding radar-detected counterparts. To achieve efficient processing time, we employed the binary search tree method with a threshold value. The matching technique relied on the disparity between the average image x-coordinate of OpenPose keypoints, denoted as $x_{mean} = \mathbf{x}/15$, and the radar coordinates transformed to image x-coordinate through a learned transform, as described in the pseudo-inverse section of [?]. The number of matched individuals is limited by the minimum count of people detected by either the camera or radars.

3.5 2D-3D lifting

Once matched, the estimated 2D human-body keypoints $(\mathbf{x}, \mathbf{y})$ are given as input to a 2D-3D lifting algorithm known as LInKs [?], leading to the estimation of 3D coordinates $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \in R^{15}$. The LInKs algorithm adopts a two-stage approach, called lift-then-fill, to handle occlusions in 3D HPE. In the first stage, it elevates the 2D keypoints that are not occluded to obtain a partial 3D pose. Subsequently, this partial pose is passed to an occlusion handling network, which fills in the missing body joints caused by occlusions. When the pose is not occluded, any two out of the four keypoints groups shown in figure 2 can be chosen for the first stage, with the left and right sides selected in our scenario as it achieved the best results. As it is impossible to retrieve absolute depth from a single view, LInKs predicts each pose to be at a fixed depth offset c from the camera. In our study, c is set to be 10 to be consistent with prior research [??]. Additionally, we normalise the 2D poses by the mean root head distance in the Human3.6M dataset (which was used to train LInKs) ensuring an average distance of 1/c from the head to the pelvis in our detected 2D poses. The final X, Y and Z coordinates are obtained via perspective projection:

$$(X, Y, Z)_i = (x_i a_i, y_i a_i, a_i) \quad (1)$$

$$a_i = max(1, d_i + c), \quad (2)$$

where d_i is the estimated depth-offset for keypoint i, and c is the predicted depth offset of the human away from the camera. The LInKs algorithm independently lifts the legs, torso, and left and right side keypoints, enabling effective occlusion handling for any or all points of a single limb, left or right side and torso or legs. This

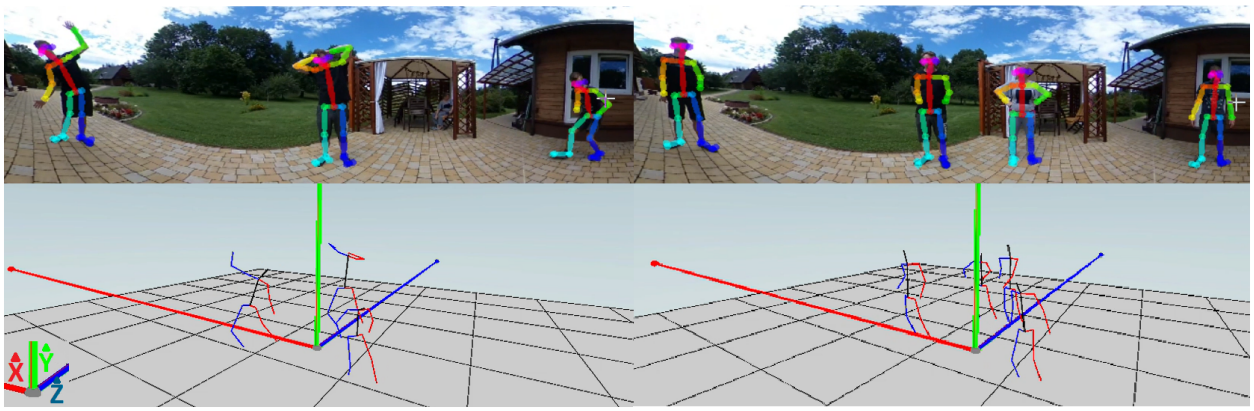


Figure 3: On the top, we have real poses captured by the camera with OpenPose outputs superimposed on them, while on the bottom, we can see the corresponding reconstructed poses in global 3D coordinate system. All pictures are partially cropped.

lifting approach also reduces long-range dependencies between keypoints e.g. the 2D keypoint of the right elbow affects the 3D keypoint of the left knee.

During the training phase, additional data is generated from the estimated distribution of a pre-trained normalising flow, accompanied by the implementation of a novel sampling loss mechanism to guarantee consistent representations of human poses. This involved drawing similar samples closer in the embedding space while pushing dissimilar ones apart. Furthermore, the method embraces self-consistency and rotational consistency during training via a virtual second view, culminating in enhanced accuracy. The occlusion handling network plays a pivotal role by transferring insights from independent lifting networks to predict absent 3D coordinates for occluded keypoints. To enhance its robustness, the model introduces two supplementary losses that consider relative bone lengths and temporal deformation between poses. The training process encompassed multiple lifting networks, each paired with its corresponding normalising flow, followed by the training of occlusion networks.

The lift-then-fill approach offered several advantages over the fill-then-lift method, where the occluded joints are filled in 2D and then lifted. Firstly, it resulted in more natural and anatomically valid 3D poses by better capturing the specific ranges of motion and dependencies between neighbouring joints. Additionally, it improved the accuracy of occluded joint estimation by effectively disentangling multiple 3D configurations that may correspond to the same 2D projection, leading to a more reliable and precise solution. Moreover, the lift-then-fill approach reduced error propagation, preventing inaccuracies introduced during occlusion handling from affecting subsequent lifting processes.

3.6 3D localisation

In line with previous work and the standard protocol of Human3.6M, [?] normalised the 2D poses by subtracting the middle of the hip (pelvis). This means that the root (x, y) coordinate is $(0, 0)$ and can be used as the root keypoint. This choice allowed for smooth transitions of poses from the local $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ to the global $(\hat{\mathbf{X}}, \hat{\mathbf{Y}}, \hat{\mathbf{Z}})$ coordinate system, ensuring alignment with the camera location. The transformation is achieved through radar data, which involved adding the radar-detected $\dot{x}$ and $\dot{z}$ coordinates to the estimated 3D keypoints' $\mathbf{X}$ and $\mathbf{Z}$ coordinates, after subtracting the depth offset c from $\mathbf{Z}$. Additionally, to maintain constant contact with the ground, the $\mathbf{Y}$ coordinates are updated by subtracting the $\mathbf{Y}$ coordinate of the lowest ankle. Notably, our system is compatible with any other off the shelf 2D-3D lifting algorithm, as long as it can derive $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ body joint local coordinates from pixel keypoints.

4 RESULTS

4.1 System setup

The proposed system utilised a commercial NVIDIA GeForce RTX 3060 graphic card, Ricoh Theta V omnidirectional camera and three TI AWR1843BOOST mmWave radars. However, the system works with any GPU, mmWave radar or omnidirectional camera as long as the latter outputs video frames in the equirectangular format. The x and z axes of the camera have to be aligned with the x and z axes of the first radar. The remaining two radars follow counterclockwise to maintain the necessary spatial consistency. The camera and radars were mounted on a tripod, as depicted in figure 4. Whereas figure 3 illustrates the camera image with the OpenPose skeleton overlaid on camera-detected individuals, along with their corresponding reconstructed poses.

4.2 Matching Camera and Radar-Detected Individuals

As part of our improvements, we made changes to the matching algorithm. The baseline approach [?], which relied on the angle between radar-detected people and camera-detected people relative to the camera x-coordinate, did not yield precise results, as demonstrated in Table 1. To address this, we implemented an alternative matching technique that solely focused on the x-coordinate in camera space, as it exhibits high distinctiveness among the located individuals when compared to the y-coordinate. For the matching

Figure 4: System experimental setup

technique to work, we transformed the radar coordinates of detected individuals into the camera space using the pseudo-inverse technique outlined in the study by Oh et al. [?]. The transformation was computed using the Levenberg-Marquardt (LM) algorithm. Data points for this calculation were collected as follows: ensuring the detection of only one person by both the radar and the camera, we gathered the radar coordinates of the detected person, along with the averaged coordinates of 15 keypoint locations obtained from the camera. For each radar, data acquisition and transform calculations were repeated. However, we observed relatively high error rates for our method in radar 2, which can be attributed to its orientation aligning with the direction where the camera x-coordinates are the smallest.

Radar	Radar 1	Radar 2	Radar 3
Baseline sys.	23.89 ± 6.57	33.57 ± 50.55	66.89 ± 263.89
Proposed sys.	2.52 ± 2.51	9.44 ± 13.27	1.94 ± 1.52

Table 1: The matching error for baseline [?] and proposed matching algorithm is computed as the absolute difference between the corresponding radar and camera values, divided by the camera value, and expressed as a percentage. For a baseline, the values used are the angles between radar-detected people and camera-detected people relative to the x-coordinate. In our algorithm, we utilize the transformed radar x-coordinate and camera x-coordinate, both in camera space.

4.3 2D-3D lifting algorithm

In its original publication, the LInKs algorithm is trained unsupervised on the Human3.6M dataset [?]. This allows for a better adaptation to different domains, which is essential for our application,

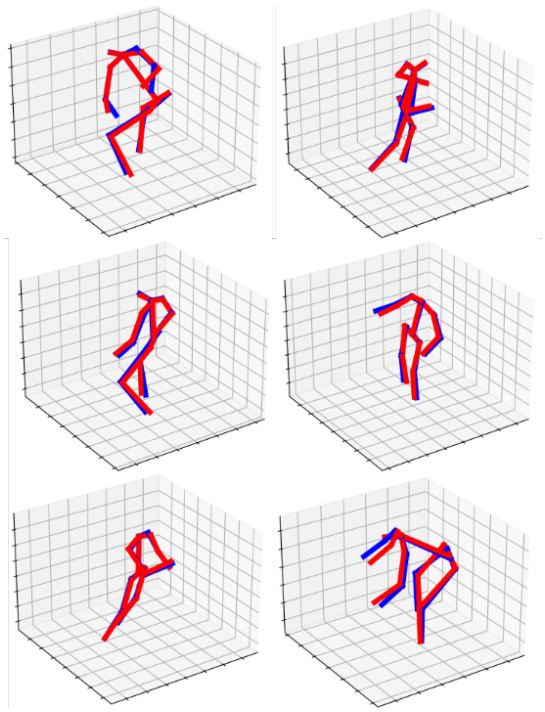
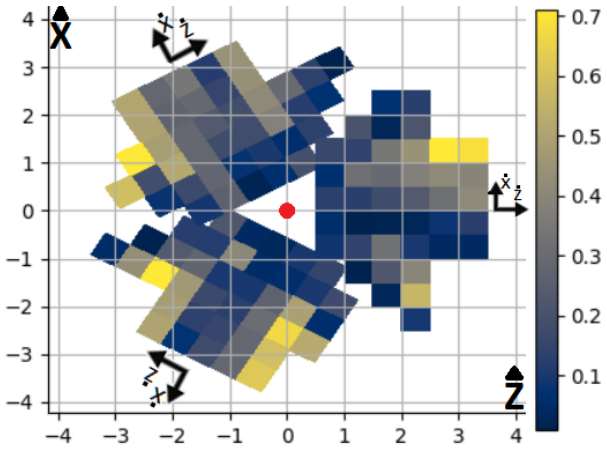


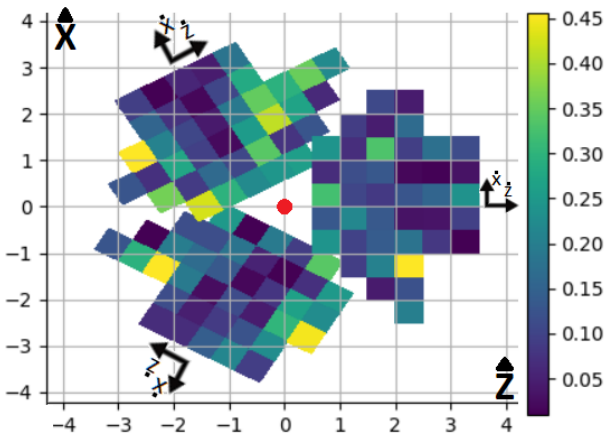
Figure 5: Qualitative pose reconstruction on the Human3.6M dataset. The GT 3D pose is in blue with our models predictions in red.

as there are no bespoke 360° lifting algorithms. Additionally, due to its independent lift-then-fill approach, the LInKs algorithm is able to deal with common forms of occlusion which occur frequently in a 360° scene. During our study, we modified this algorithm in order to lift the 15 2D keypoints obtained from OpenPose by removing the spine and head-top keypoint present in the Human3.6M training data. The limitation of this approach is that it is currently only able to successfully complete specific forms of occlusion. These are for the left leg, right leg, left arm, right arm, left leg and arm, right leg and arm, both legs and full torso. If we have any other form of occlusion present for example left leg and right arm then the occlusion handling step will not function as intended. In addition, the Pelvic coordinate is used to centre and normalise the 2D pose prior to being seen by the lifting. If this is not correctly detected then the system may detect a person, but a 3D pose cannot be accurately reconstructed, resulting in no visualisation. The examples of reconstructed poses against ground truth pose from Human3.6M dataset can be found in figure 5.

For evaluation, we employed two standard metrics: the Procrustes-aligned Mean Per Joint Position Error (PA-MPJPE) and the Normalised Mean Per Joint Position Error (N-MPJPE). The PA-MPJPE computes the average Euclidean distance between the estimated joint positions and their corresponding ground-truth positions, utilising Procrustes analysis to align the estimates accurately through scaling, rotation, and translation. Conversely, the N-MPJPE is the



(a) Localization errors in each radar's $\dot{x}$ directions



(b) Localization errors in each radar's $\dot{z}$ directions

Figure 6: Localization error in meters at various points around the system. Errors are evaluated in each radar's $\dot{x}$ (figure 6a) and $\dot{z}$ (figure 6b) directions, but the figures display them in the $(\hat{X}, \hat{Z})$ 2D global coordinate system. The red dot marks the system location.

Supervision	Method	PA-MPJPE	N-MPJPE
Full	?	37.1	45.5
	?	40.2	48.8
Weak	?	38.2	50.9
	?	38.2	-
Unsupervised	?	42.0	85.3
	?	36.7	64.0
	LInKs [?]	33.6	60.8
	Ours	**37.2**	**61.7**

Table 2: Evaluation results of our unsupervised 2D-3D lifting algorithm on the GT 2D poses in the Human3.6M dataset. Numbers are taken from their respective papers. Note that we lift 15 keypoints whereas other unsupervised works and the original LInKs algorithm lift 17.

Euclidean distance between the joints of our predicted and GT 3D poses after scaling alone. Lower values of both PA-MPJPE and N-MPJPE indicate improved accuracy in pose estimation.

The evaluation results of our unsupervised model are shown in table 2, where comparisons are made with other state-of-the-art models. It is worth noting that the compared approaches utilised 17 keypoints for lifting, while our model used 15 keypoints, excluding the head-top and spine. The omission of these keypoints, which are relatively straightforward to estimate depth for, causes results to be inflated. Still, the original LInKs algorithm surpassed all other unsupervised methods. Furthermore, we still attain the second lowest N-MPJPE error among the algorithms considered with only the original implementation of LInKs surpassing ours. Table 3 showcases the evaluation scores of LInKs in occlusion scenarios. Note that these results came from [?] and therefore use the full 17 keypoints of the Human3.6M dataset during evaluation and lifting.

Model	Occlusion	PA-MPJPE	N-MPJPE
LInKS [?]	Left Arm	52.1	78.1
	Left Leg	46.0	73.2
	Right Arm	49.8	75.7
	Right Leg	44.5	71.6
	Left Arm & Leg	62.0	86.0
	Right Arm & Leg	60.2	83.7
	Both Legs	69.3	99.8
	Full Torso	88.4	122.0

Table 3: Evaluation results for the Human3.6M dataset in occlusion scenarios. When a single limb is occluded in the occlusion scenarios of LInKs all keypoints belonging to that limb are occluded e.g. for the arm it will be the shoulder, elbow and wrist. Results were taken from [?]

4.4 3D localisation - radar calibration

The application of the affine transform for radars' calibration resulted in a reduction of the mean localisation error, as demonstrated in Table 4. However, despite these improvements, some imperfections still remain. Figure 6 visually illustrates the localisation errors at various points around the system. These errors are evaluated separately in each radar's $\dot{x}$ and $\dot{z}$ directions, just like the data presented in Table 4. However, in the figures, we present these errors in the global coordinate system, offering a collective view of the data from all radars. Due to our system's operational characteristics,

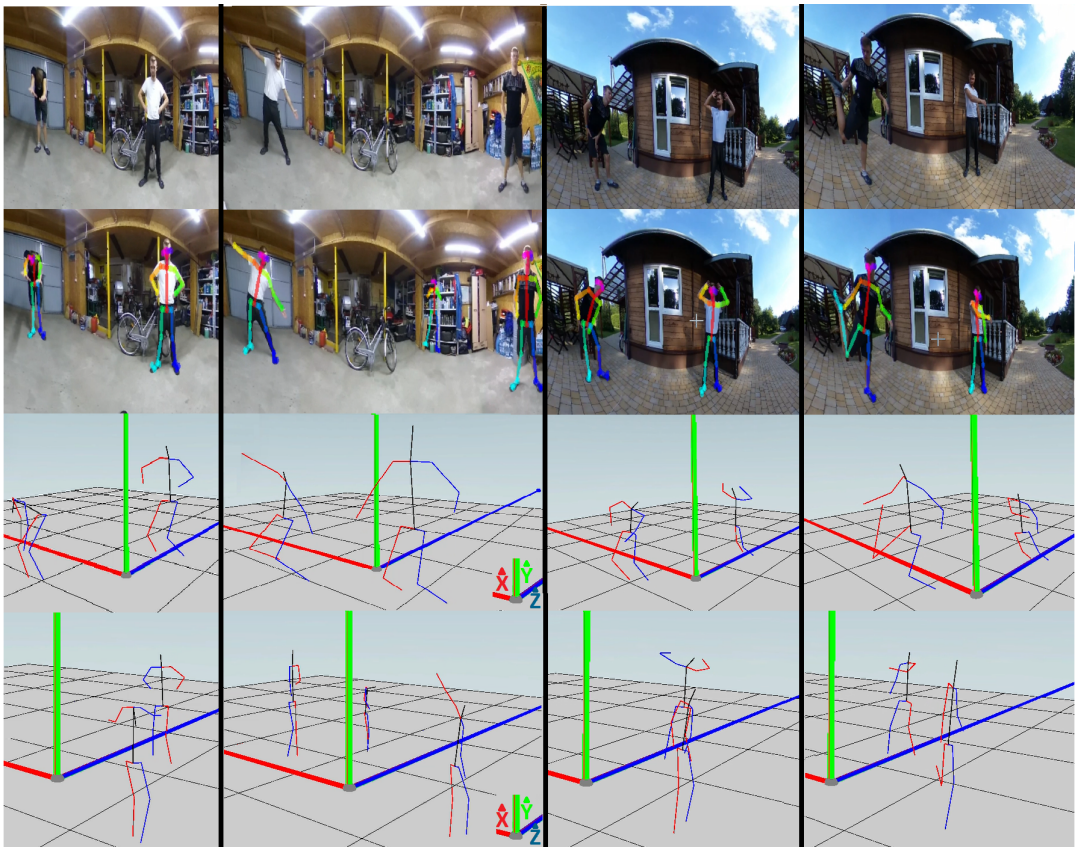


Figure 7: In the uppermost section, we exhibit the inputs to the system sourced from the camera. Positioned in the upper-middle area are the OpenPose outputs, which are overlaid on the corresponding images taken by the camera. The bottommost sections present the reconstructed poses in the global 3D coordinate system, with our system's results showcased above and those of the baseline [?] displayed below. As clearly seen in the second OpenPose output, our visualization does not reconstruct the camera-detected false-positive, whereas the baseline [?] does. Please note that all pictures are partially cropped.

these radar errors directly translate into 3D localisation errors.

Furthermore, figure 6 highlights two significant issues. Firstly, there is a lower accuracy of radars at approximately 60° degrees away from the radar centre, which results in decreased reliability in those regions despite the radar's specified 120° coverage. Secondly, void spaces exist in these areas, where no radar can detect a person. To address these limitations, a straightforward solution is to incorporate an additional radar to ensure complete coverage of the entire space.

Radar	Dir.	Baseline	Proposed with calibration
1	x	20.65 ± 16.86	16.45 ± 10.57
	z	11.41 ± 8.86	11.45 ± 8.58
2	x	26.19 ± 17.2	24.86 ± 15.98
	z	15.39 ± 9.81	10.77 ± 7.66
3	x	16.88 ± 12.61	15.94 ± 12.91
	z	13.83 ± 11.48	13.46 ± 10.26

Table 4: The average absolute error in centimetres in the x and z direction for radar 1,2 and 3 before and after the calibration. Where the z direction is perpendicular, and the x direction is parallel to the front surface of the radar.

4.5 Subjective evaluation

Conducting a subjective evaluation, we analysed the system's performance as individuals executed diverse poses in both indoor and outdoor settings. Figure 7 presents a subset of these instances. Moreover, we evaluated the system's occlusion handling capability, as portrayed in Figure 8. Remarkably, the system showcased consistent and stable operation across both environments, maintaining a frame rate of 7-8 fps, regardless of the number of individuals detected. The primary speed constraint is attributed to OpenPose, which operates relatively slowly due to the high resolution (1960x980) of the omnidirectional camera. Furthermore, the system's operational range is defined by the limited capability of the radars to detect people at distances beyond approximately 3.5 meters. Nevertheless, provided that the pose occlusion remains within the system's capacity and

individuals are not completely stationary but within the system's range, they are consistently detected in the scene.

False positives may occasionally be detected by OpenPose, as depicted in figure 7, or by the radars. However, their appearance in the visualisation is virtually impossible since both sensors would have to detect false positives at the exact same location. Consequently, false positive poses are never detected by our system, in comparison to the baseline, where a simple matching technique doesn't prevent false positive poses. Our solution maintains a consistent pose size irrespective of the person's location, whereas, in the baseline, the human pose size varies based on the camera's distance. Notably, we addressed the baseline issues that could wrongly depict a left hand being up instead of the right in visualisation, or misrepresent a person's movement direction, which our solution rectifies. As our system performs a multi-person 3D human pose estimation, in theory, it has no limitations on the maximum number of detectable people. However, the number of detectable people may be limited depending on the type of mmWave radars used.

As can be seen in figure 7, individuals standing straight appear to have bent knees. We posit this issue arises due to the compression of the width of 2D poses in a 360° view when compared to that in the Human3.6M training data. As the 2D-3D lifting model is trained via a pin-hole camera system, the detected 2D poses, have different scales in our 360° scenario. As an example, the omnidirectional camera provides approximately 5.3 pixels per degree FOV, whereas the Human3.6M dataset, used for training, offers approximately 12.5 pixels per degree FOV. To tackle this issue, a straightforward solution would be to utilise a different 360° camera with higher resolution, ideally having 4500 pixels width to achieve 12.5 pixels per degree.

5 CONCLUSIONS

In conclusion, we introduce a novel 3D multi-person detection system that is able to work in real time. Our approach has multiple advantages and improvements to the baseline system [?], resulting in a simpler, more robust, and higher-performing solution. Our system maintains stable performance regardless of the number of detected individuals, and theoretically, there is no limitation on the number of detectable people. Nonetheless, the existing limitations lie in the system's speed, range, and occlusion handling capabilities. Future research aims to enhance occlusion handling, optimise the algorithm execution time for improved speed, extend the system's range, and enhance the accuracy of pose estimation in challenging real-world environments. Moreover, different hardware, including additional radar and cameras with higher resolution would remove the current limitation of blank detection spaces and bent knees of visualised individuals. Overall our findings and improvements are substantial, making the technology more accessible and robust for computer vision systems. As a result, our contributions have paved the way for this system to be an affordable and dependable solution in the industry.

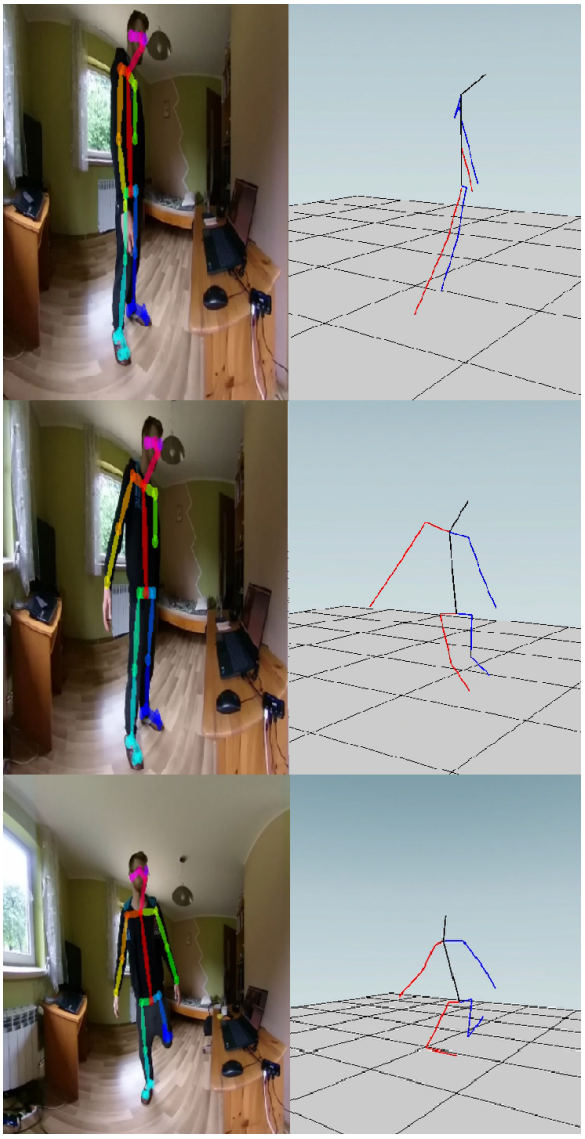


Figure 8: System's ability to handle occlusion. On the left, OpenPose outputs are overlaid on the respective camera-captured pictures. On the right, the reconstructed poses in the global 3D coordinate system are displayed. The occluded joints from top to bottom are the left arm, left arm, and left leg. Please note that all pictures are partially cropped.

REFERENCES

M. Alawadh, Y. Wu, Y. Heng, L. Remaggi, M. Niranjan, and H. Kim. 2022. Room Acoustic Properties Estimation from a Single 360 Photo. In *European conference on signal processing 2022*. 857.

S. An and U. Y. Ogras. 2021. MARS: mmWave-based Assistive Rehabilitation System for Smart Healthcare. *ACM Transactions on Embedded Computing Systems (TECS)* 20, 5s (2021), 1–22.

Anonymous. 2023a. LInKs - Lifting Independent Keypoints - Exploring Partial Pose Lifting for Occlusion Handling and Improved Accuracy within 2D-3D Human Pose Estimation. In *Submitted to WACV 2023 (submission version attached)*.

Anonymous. 2023b. Real time 3D multi-person human pose estimation using an omnidirectional camera and mmWave radars. In *Submitted to ICEET (submission version attached)*.

Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. 2019. Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks. In *2019 IEEE/CVF International*

Conference on Computer Vision (ICCV). 2272–2281.

Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y.A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1.

C.-H. Chen and D. Ramanan. 2017. 3D human pose estimation = 2D pose estimation + matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7035–7043.

Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 2020. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10631–10638.

Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. 2019. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF international conference on computer vision*. 723–732.

Dylan Drover, Rohith M. V, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. 2019. Can 3D Pose Be Learned from 2D Projections Alone?. In *Computer Vision – ECCV 2018 Workshops*, Laura Leal-Taixé and Stefan Roth (Eds.). Springer International Publishing, Cham, 78–94.

M. Furst, S. Gupta, R. Schuster, O. Wasenmuller, and D. Stricker. 2020. HPERL: 3D Human Pose Estimation from RGB and LiDAR.

Keegan Garcia. 2019. Bringing intelligent autonomy to fine motion detection and people counting with TI mmWave sensors. https://www.electronicspecifier.com/products/sensors/people-counting-and-tracking-using-mmwave-radar-sensors

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. 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 (2014), 1325–1339.

C. Keskin and et al. 2013. Real time hand pose estimation using depth sensors. In *Consumer Depth Cameras for Computer Vision*. 119–137.

L. Kumarapu and P. Mukherjee. 2021. AnimePose: Multi-person 3D pose estimation and animation. *Pattern Recognition Letters* 147 (2021), 16–24.

R. Ludlow. 2018. 3D Human Pose Estimation from 2D Keypoints. *github.com/rludlow/3d-pose-2d-keypoints* (June 2018).

K. Ludwig, S. Scherer, M. Einfalt, and R. Lienhart. 2021. Self-Supervised Learning for Human Pose Estimation in Sports. In *2021 IEEE International Conference On Multimedia & Expo Workshops (ICMEW)*. 1–6.

M. Martin, S. Stuehmer, M. Voit, and R. Stiefelhagen. 2017. Real time driver body pose estimation for novel assistance systems. In *2017 IEEE 20th International Conference On Intelligent Transportation Systems (ITSC)*. 1–7.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. 2017. A simple yet effective baseline for 3D human pose estimation. In *Proceedings of the IEEE international conference on computer vision*. 2640–2649.

Jiyong Oh, Ki-Seok Kim, Miryong Park, and Sungho Kim. 2018. A Comparative Study on Camera-Radar Calibration Methods. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. 1057–1062.

Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. 2018. Ordinal Depth Supervision for 3D Human Pose Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7307–7316.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marco Roberto Ronchi, Oisin Mac Aodha, Robert Eng, and Pietro Perona. 2018. It's all Relative: Monocular 3D Human Pose Estimation from Weakly Supervised Data. In *British Machine Vision Conference 2018, BMVC 2018*. Northumbria University, Newcastle, UK, 300.

A. Sengupta, F. Jin, R. Zhang, and S. Cao. 2020. MM-pose: Real-time human skeletal posture estimation using mmwave radars and CNNS. *IEEE Sensors Journal* 20, 17 (2020), 10032–10044.

Bugra Tekin, Sudipta N Sinha, and Pascal Fua. 2016. Structured prediction of 3D human pose with deep neural networks. In *BMVC*.

Denis Tome, Chris Russell, and Lourdes Agapito. 2017. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

B. Wandt, J.J. Little, and H. Rhodin. 2022. Elepose: Unsupervised 3D human pose estimation by predicting camera elevation and learning normalizing flows on 2D poses. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bastian Wandt and Bodo Rosenhahn. 2019. RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 7774–7783.

Jiaxin Wang, Zhenbo Yu, Zekun Tong, Huazhong Wang, Jun Liu, Wenjun Zhang, and Xiaogang Wu. 2022. OCR-Pose: Occlusion-Aware Contrastive Representation for Unsupervised 3D Human Pose Estimation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. Association for Computing Machinery (ACM), New York, NY, USA, 5477–5485.

Y. Wu, Y. Heng, M. Niranjan, and H. Kim. 2021. Depth Estimation from a Single Omnidirectional Image using Domain Adaptation. In *18th ACM SIGGRAPH European Conference on Visual Media Production*.

H. Xue and et al. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 269–282.

Jia Yang, Li Wan, Wei Xu, and Song Wang. 2019. 3D human pose estimation from a single image via exemplar augmentation. *Journal of Visual Communication and Image Representation* 59 (2019), 371–379.

Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy SJ Ren, Hongsheng Li, and Xiaogang Wang. 2018. 3D Human Pose Estimation in the Wild by Adversarial Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5255–5264.

Zhenbo Yu, Bingbing Ni, Jingwei Xu, Junjie Wang, Chenglong Zhao, and Wenjun Zhang. 2021. Towards Alleviating the Modeling Ambiguity of Unsupervised Monocular 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8651–8660.

Zhengyou Zhang. 2000. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence* 22, 11 (2000), 1330–1334.

Xingyi Zhou, Qixing Huang, Xiangxu Sun, Xiangyang Xue, and Yichen Wei. 2017. Towards 3D Human Pose Estimation in the Wild: A Weakly Supervised Approach. In *The IEEE International Conference on Computer Vision (ICCV)*.