Real-Time Omnidirectional 3D Multi-Person Human Pose Estimation with Occlusion Handling

Pawel Knap[1] Peter Hardy[1] Alberto Tamajo[1] Hwasup Lim[2] Hansung Kim[1]
pmk1g20* p.t.d.hardy* at2n19* hslim@kist.re.kr h.kim*
*@soton.ac.uk

[1] University of Southampton
[2] Korea Institute of Science and Technology

Abstract

We present a multi-person 3D human pose estimation system that addresses a major limitation in existing models, namely the focus on single-person pose estimation. By using an off-the-shelf 2D detectors and 2D-3D lifting model, we first obtain the 3D poses of detected individuals in their own local coordinate system from video. We then use a radar sensing data and people-matching approach to localise the 3D poses within a global coordinate system accurately reconstructing the scene in real-time.

Introduction

Existing 3D human pose estimation (HPE) models primarily target single-person HPE, while our approach introduces a method for multi-person HPE. By using a 360° panoramic camera and mmWave radar sensors, our system effectively resolves depth and scale ambiguities. It uses a real-time occlusion-handling 2D-3D pose lifting algorithm [2] allowing for accurate performance capture both indoors and outdoors, all while remaining affordable and scalable. As evidence, we find that our system maintains a consistent time complexity irrespective of the number of detected individuals, achieving a frame rate of around 7-8 fps on a commercial-grade GPU. Our method revolves around transforming 2D body keypoints, detected by OpenPose[1], into predicted 3D keypoints in a global coordinate space obtained via radar sensing data. The system proceeds through several stages, which can be seen in figure 1.

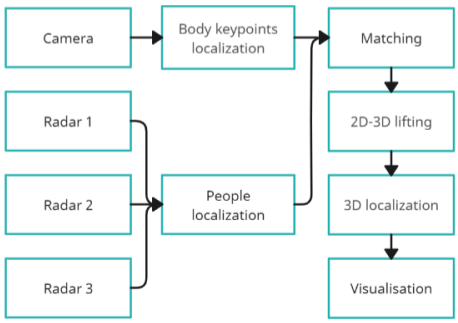

Figure 1: Overview of our approach, OpenPose extracts 2D keypoints from an image, while radars localize individuals. Subsequently, we match the individuals detected by both radar and camera systems. Next, the 2D keypoints are elevated into 3D, and their positions are adjusted within the global coordinate system using radar data.

System overview

Utilizing OpenPose [1], key body points' 2D Cartesian coordinates are extracted from the video frames, serving as inputs for the 2D-3D lifting model [2]. This algorithm employs a two-stage process, known as lift-then-fill, to address the problem of occlusion. Initially, it elevates the unobstructed 2D keypoints to form a partial 3D pose. Subsequently, an occlusion handling network completes missing joints caused by occlusions. The individuals detected by both radar and camera are matched using a binary search tree. This matching is based on the disparity between the average x-coordinate of 2D keypoints and the radar coordinates converted into the images x-coordinate space through a learned transformation. Finally, the detected pose is moved to its 3D coordinates by adding the radar-gathered positional data to 3D keypoints coordinates of identified individuals.

Results

Our system achieves a low average matching error of 4.63%, calculated as the disparity between a coordinate of correctly matched camera and radar individuals. The 2D-3D lifting algorithm [2] achieved competitive results with a PA-MPJPE of 37.2 and N-MPJPE of 61.7 on the GT 2D poses in the Human3.6M dataset, with qualitative results shown in Figure 2. Additionally, radar and camera calibrations reduced localisation errors by a few centimetres depending on the direction. We also conducted an ablation study which showed that the system performed consistently well with diverse poses, indoors and outdoors as shown in Figure 3. Furthermore, false positives were effectively filtered out. The system's primary constraint is the runtime of off-the-shelf 2D detectors, due to the large resolution present in 360° cameras. Additionally in our scenario, the radar detection range is approximately 3.5 meters.

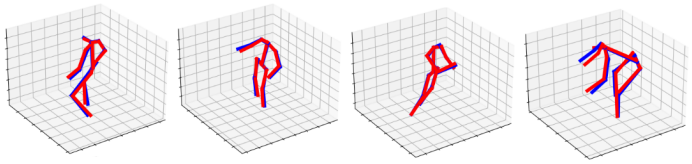

Figure 2: Qualitative pose reconstruction on the Human3.6M dataset. The GT 3D pose is in blue with our models predictions in red.

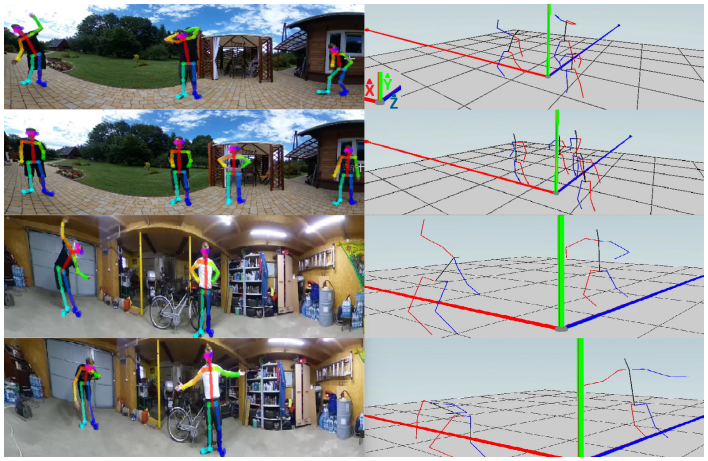

Figure 3: Actual poses captured by the camera, overlaid with OpenPose outputs, and reconstructed poses in the global 3D coordinate system.

Conclusion

Our real-time 3D multi-person detection system is robust, performs consistently regardless of the number of individuals, and theoretically can handle any number of detected people. The only limitations being, the speed of off-the-shelf 2D detectors and the range of the radar sensor. In our future work, we aim to develop a 2D detection approach that can process frames faster in high-resolution scenarios and extend the range of the radar which will enhance accuracy. Nonetheless, our contributions have paved the way for this system to be an affordable and dependable solution in the industry.

Acknowledgements

This work was partially supported by the EPSRC Programme Grant Immersive Audio-Visual 3D Scene Reproduction (EP/V03538X/1) and partially by the Korea Institute of Science and Technology (KIST) Institutional Program (Project No. 2E32303).

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y.A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[2] Peter Hardy and Hansung Kim. Links - lifting independent keypoints - partial pose lifting for occlusion handling with improved accuracy in 2d-3d human pose estimation, 2023.