Evaluation of Accuracy and Angle Dependency of 3D Pose Estimation through Stereo Camera Information Fusion with MediaPipe Pose

1st Sebastian Dill
*KIS*MED (AI Systems in Medice Lab)*
Technische Universität Darmstadt
Darmstadt, Germany
dill@kismed.tu-darmstadt.de

2nd Maurice Rohr
*KIS*MED*
TU Darmstadt
Darmstadt, Germany

3rd Gökhan Güney
*KIS*MED*
TU Darmstadt
Darmstadt, Germany

4th Christoph Hoog Antink
*KIS*MED*
TU Darmstadt
Darmstadt, Germany

***Abstract*—In recent years, significant research has been conducted on video-based human pose estimation (HPE). While monocular 2D HPE has been shown to achieve high performance, monocular 3D HPE is more challenging. Fusing the advantages of high accuracy in 2D HPE with the increased usability of 3D coordinates, we propose a method based on MediaPipe Pose 2D HPE on stereo cameras, epipolar geometry and direct triangulation to reconstruct 3D poses. We use the CMU Panoptic database, which provides recordings of humans from 31 different HD views and 3D ground truth data, to research which accuracy can be achieved from fusing only two cameras without prior stereo calibration. We also research which camera perspectives to employ, analyzing the angle dependency of our approach.**

***Index Terms*—computer vision, human pose estimation, fusion**

I. Introduction

Building on the advances in computer vision in recent years, significant research has been conducted on video-based human pose estimation (HPE) and motion capture. Two-dimensional HPE, i.e. the task of locating the human pose of a single person and their distinct joints in an image plane, has been shown to achieve high performance. Popular 2D approaches using deep learning techniques include OpenPose [1], DeepPose [2] and MediaPipe Pose [3]. The feasibility of popular 2D approaches has been shown for particular medical applications like gait analysis [4]. Conversely, three-dimensional HPE, i.e. the task of locating the human pose of a single person and their distinct joints in a 3D camera coordinate system, is more challenging than its 2D counterpart, because 3D pose estimation from monocular inputs presents an ill-posed problem, as multiple 3D predictions can correspond to the same 2D projection. Still, for obvious reasons 3D approaches offer a more accurate representation of the human's pose, and have been used for more complex medical applications such as joint load prediction [5]. One particularly popular 3D HPE library is, again, MediaPipe Pose, which is based on the BlazePose model [3] due to its computational efficiency, ease of use, and the fact that it is open-source. While MediaPipe Pose offers 3D pose estimations from a single camera view, the z-Axis, which is oriented perpendicularly to the image plane suffers from high noise. This deteriorates the overall estimation quality, as shown in one of our previous papers, where we have evaluated the accuracy of a MediaPipe-based pose estimation for physical therapy [6]. The idea, to combine the advantages of 3D pose estimation with the easier problem of 2D pose estimation, has led to research focusing on multi-view pose estimation [7] [8] [9] [10] [11]. In theory, two sets of 2D pose coordinates are enough to reconstruct a 3D pose utilizing intrinsic and extrinsic camera parameters as well as direct triangulation to identify matching epipolar lines between the two views. However, this 3D reconstruction is still highly dependent on the quality of the 2D estimation, which is hindered by problems such as (self-)occlusion [12]. Therefore, these approaches usually scale in accuracy with the number of cameras. However, with the recent increase of estimation accuracy in 2D estimators, we want to evaluate how well 3D stereo reconstruction, i.e. finding the position of a point in space given its position in just two images holds up to the multi-view version. Similar work, utilizing two distinct cameras that were approximately perpendicular to one another, and a machine learning model, has been done in [13]. Due to their focus on fixed positions for the cameras, the researchers concentrated on the upper body joints and could not evaluate the dependency of their approach on the camera angle.

In this work, we aim to do a 3D reconstruction based on stereo camera information fusion of 2D HPE performed with MediaPipe Pose. We use epipolar geometry, direct triangulation, and signal processing and evaluate our approach on the CMU Panoptic database [14]. We also aim to find the best camera views for this problem, evaluating the angle between the two cameras.

II. Methods

In order to give a qualitative evaluation of the 3D reconstruction based on stereo camera MediaPipe pose estimation,

The authors gratefully acknowledge financial support provided by the Hessian Ministry for Digital Strategy and Development [Hessisches Ministerium für Digitale Strategie und Entwicklung, Distr@l-Förderlinie 2, "SG4smartmedication", 21_0038_2A].

we utilize the CMU Panoptic dataset [14], which contains synchronized recordings of human movement from over 500 different views as well as ground truth 3D data. Out of the views, we repeatedly select sets of two cameras and feed their image streams into two distinct instances of the MediaPipe Pose framework [3] to receive two sets of 2D human pose joint estimations. With methods of epipolar geometry and triangulation, we give an estimate for the 3D world coordinates to evaluate which set of cameras is best suited for the reconstruction. To limit computation time, we perform a preliminary camera selection based on a projection of the 3D ground truth data onto the image planes.

A. CMU Panoptic Dataset

The CMU Panoptic dataset [14] features human subjects captured in video from 480 SD cameras with a resolution of 640×480 pixels and 25 frames per second (fps), as well as 31 HD cameras with 1920×1080 pixels resolution and 30fps. Due to the advantages of the higher resolution and fps, only the HD cameras were used. Among them, two cameras did not capture all of the recordings. Therefore, we limit ourselves to 29 HD cameras. In [14], Joo et al. present a method of 3D reconstruction utilizing all 480 SD camera views to generate 3D positions of 19 joints of all recorded subjects. This 3D data is also provided and will be considered the ground truth (GT) for our work. The data from all cameras and the GT has been synchronized. Lastly, the dataset provides camera calibration parameters in the form of intrinsic camera matrices, distortion parameters, and extrinsic camera projection matrices describing the projection of the GT onto each image plane for all cameras. These parameters will be explained in more detail in Section II-C. Overall, the dataset contains 146 recording sessions. Among these, 137 focus on either group activities, complex interactions, or the motion capture of specific body parts such as the hands. Due to our goal of performing single-person estimation of the whole body, we focus on the subset called *Range of Motion*, which consists of nine recordings featuring between one and eight subjects each. Specifically, we focus on the recording titled *171026_pose1* featuring 3 subjects over a recording time of 13 minutes and 28 seconds.

B. MediaPipe Pose

We present a brief overview of the BlazePose model, which is utilized by MediaPipe and therefore serves as the baseline for our research. MediaPipe's output consists of x-y-z-coordinates of 33 different joint positions as well as a visibility estimate for each joint, ranging from 0 to 1, with a higher value indicating more confidence on the estimated joint position. The coordinate system in which these positions are given depends on the operating mode. In *landmarks* mode, the output coordinates are given as image coordinates, where the x-y-plane is parallel to the image plane and the z-axis is oriented perpendicularly away from the camera. In *worldmarks* mode, these image coordinates are mapped onto an internal model of a human to create an estimate for real-world coordinates in meters. This coordinate system is centered between the hip joints and moves with the subject. In this work we want to focus on stereo camera information fusion through methods of epipolar geometry. Since the coordinates provided by MediaPipe's *worldmarks* mode are based in a world coordinate system independent of the camera, they are not suitable for these methods. Therefore, we only consider *landmarks* mode in the following. A visualization of the landmarks and the coordinate system is given in Fig. 1. Since MediaPipe outputs 33 joint positions and our GT data only consists of 19, which are in turn not all included in the 33, we select a common subset of 12 joints: shoulder, elbow, wrist, hip, knee, and ankle for both the left and right side.

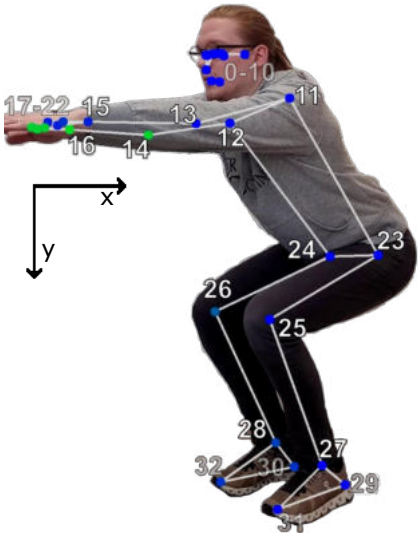


Fig. 1: General visualization of the MediaPipe Pose output of a person performing a squat. The output consists of x-y-z coordinates of 33 different landmarks. The coordinate system's origin is in the upper left corner of the image. MediaPipe also outputs a visibility estimate, which is color-coded from blue (1) to green (0). In this work, we only consider joints 11 to 16 for the upper body and 23 to 28 for the lower body.

While MediaPipe does output an estimate for the z-coordinate, it is far less reliable than estimations in the other two axes. This is because 3D pose estimation from monocular inputs presents an ill-posed problem, as multiple 3D predictions can correspond to the same 2D projection. To give more context to the relationship between the 2D and 3D coordinate systems, the following sections II-C and II-D introduce definitions and fundamentals of coordinate systems and 3D pose reconstruction.

C. Coordinate Systems

We define four different coordinate systems: 2D image coordinates $\mathbf{x} = [x, y]$, 3D image coordinates $\mathbf{x}_{3D} = [xz, yz, z]$, 3D camera coordinates $\mathbf{p}_{\text{cam}} = [x_{\text{cam}}, y_{\text{cam}}, z_{\text{cam}}]$ and 3D world coordinates $\mathbf{p}_{\text{world}} = [x_{\text{world}}, y_{\text{world}}, z_{\text{world}}]$. In the following, we will briefly describe how these four coordinate systems relate to each other.

In this work we will use homogeneous coordinates, i.e. coordinates with an extra scaling dimension, which helps displaying transformations between coordinate systems in a mathematically concise way. To reduce complexity in notation, homogeneous coordinates will be denoted by a Tilde symbol. A visualization of the relationships between the coordinate systems can be seen in Fig. 2.

The 3D camera coordinate system is defined such that the camera is positioned at the origin of the Euclidean coordinate system, with the camera's principal axis pointing directly along the z-axis. The image coordinate space is related to the 3D camera coordinate system through a projection, often approximated by the pinhole camera model. The pinhole camera model provides a fundamental explanation of how images are created through projection. It illustrates the geometric relationship between objects in space and their images on a flat image plane [15]. When applying the pinhole camera model, the projection from 3D camera coordinates $\mathbf{p}_{\text{cam}}$ into image space image coordinates $\mathbf{x}_{\text{3D}}$ is given by

$$\mathbf{x}_{\text{3D}} = \mathbf{K}\mathbf{p}_{\text{cam}}, \tag{1}$$

where $\mathbf{K}$ is referred to as the camera calibration matrix, which includes the intrinsic parameters of the camera. It is denoted by the following equation:

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

with f_x and f_y being the x and y components of the focal length f of the camera and c_x and c_y define the position of the principal point on the sensor. From the definition of $\mathbf{x}$ and $\mathbf{x}_{\text{3D}}$, it directly follows that $\tilde{\mathbf{x}} = \frac{1}{z}\mathbf{x}_{\text{3D}}$.

The assumptions of the pinhole camera model are idealized, and in real cameras, lenses and other optical elements are used to further focus and direct light. The curvature and material of the lens cause light rays not to be perfectly focused on a point, leading to both radial and tangential distortions in the image. Overall, we consider the following distortion coefficients $\mathbf{d}$:

$$\mathbf{d} = (k_1, k_2, p_1, p_2, k_3)$$

Since both the camera calibration matrix and the distortion coefficients can easily be estimated and are given in the CMU Panoptic dataset, we use them to undistort the 2D image coordinates $\mathbf{x}$ we receive from MediaPipe and then calculate the so-called normalized image coordinates:

$$\tilde{\mathbf{x}}_{\text{normalized}} = \mathbf{K}^{-1}\tilde{\mathbf{x}} = \frac{1}{z_{\text{cam}}}\mathbf{p}_{\text{cam}}, \tag{3}$$

which are independent of the camera-specific values for resolution and focal length. With this transformation out of the way, these camera-specific parameters can be ignored for the following sections. From now on, all mentions of the image coordinates refer to the normalized image coordinates.

The relationship between the 3D world coordinate system and the 3D camera coordinate system is described by the camera rotation and translation. Given a point $\mathbf{p}_{\text{world}}$ with coordinates $(x_{\text{world}}, y_{\text{world}}, z_{\text{world}})$ in the world coordinate system, to transform this point into the camera coordinate system, we must account for the camera's position and orientation in the world coordinate system. The camera is located at a specific point $\mathbf{c}$ in the world coordinate system, given by the vector $\mathbf{c} = (c_x, c_y, c_z)$. The orientation of the camera in space is described by a rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$, which transforms points from the world coordinate system into the camera coordinate system, taking into account the orientation of the camera. Alternatively, the rotation can also be represented by a 3-element rotation vector $\mathbf{r}$ whose direction represents the axis of rotation and whose norm represents the angle of rotation in radians. This representation can be used to concisely represent the rotation between two cameras by a single value, which is helpful when evaluating which angle between cameras is optimal without regard for the cameras' positions in relation to the subject. The overall transformation is then given by $\mathbf{p}_{\text{cam}} = \mathbf{R}(\mathbf{p}_{\text{world}} - \mathbf{c})$. In homogeneous coordinates the translation can be expressed by a matrix multiplication, resulting in the following equation:

$$\tilde{\mathbf{p}}_{\text{cam}} = \begin{bmatrix} \mathbf{R} & -\mathbf{R}\mathbf{c} \\ \mathbf{0} & 1 \end{bmatrix} \tilde{\mathbf{p}}_{\text{world}} = \mathbf{P}_{\text{world,cam}}\tilde{\mathbf{p}}_{\text{world}}, \tag{4}$$

where $-\mathbf{R}\mathbf{c}$ is usually referred to as the translation vector $\mathbf{t}$ and $\mathbf{P}_{\text{world,cam}} \in \mathbb{R}^{4\times4}$ is called the projection matrix from world to camera coordinates.

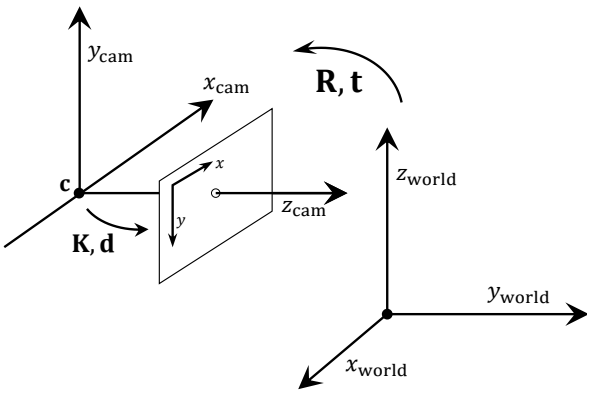


Fig. 2: A visualization of the relationships between the 2D image coordinates $\mathbf{x} = [x, y]$, 3D camera coordinates $\mathbf{p}_{\text{cam}} = [x_{\text{cam}}, y_{\text{cam}}, z_{\text{cam}}]$ and 3D world coordinates $\mathbf{p}_{\text{world}} = [x_{\text{world}}, y_{\text{world}}, z_{\text{world}}]$. The camera coordinates are obtained by transforming the world coordinates with the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$. The relationship between the camera coordinates and image coordinates is described through the intrinsic camera matrix $\mathbf{K}$ and the distortion coefficients $\mathbf{d}$.

D. Stereo-View 3D Pose Reconstruction

If we have not one but two cameras c1 and c2, the two 3D camera coordinate systems relate to one another through

a rotation and translation, which can be expressed by

$$\mathbf{p}_{c1} = \mathbf{R}_{21}\mathbf{p}_{c2} + \mathbf{t}_{21}, \tag{5}$$

where $\mathbf{R}_{21} \in \mathbb{R}^{3\times3}$ and $\mathbf{t}_{21} \in \mathbb{R}^{3\times1}$ denote the rotation and translation from camera coordinate system 2 to camera coordinate system 1, respectively. When applying equation 3, we get

$$\tilde{\mathbf{x}}_{c1} = \frac{1}{z_{c1}} \left(z_{c2}\mathbf{R}_{21}\tilde{\mathbf{x}}_{c2} + \mathbf{t}_{21} \right). \tag{6}$$

By first applying a cross product of $\mathbf{t}$ and then a scalar product of $\tilde{\mathbf{x}}_{c1}$ from the left, we obtain

$$\tilde{\mathbf{x}}_{c1} \cdot (\mathbf{t} \times \tilde{\mathbf{x}}_{c1}) = \frac{z_{c2}}{z_{c1}} \tilde{\mathbf{x}}_{c1}^T \mathbf{t}_{21} \times \mathbf{R}_{21}\tilde{\mathbf{x}}_{c2}, \tag{7}$$

where the left side equals zero. The relationship between the two coordinate systems can therefore be defined as

$$\tilde{\mathbf{x}}_{c1}\mathbf{E}_{21}\tilde{\mathbf{x}}_{c2} = 0, \tag{8}$$

where the so-called essential matrix $\mathbf{E}_{21} = \mathbf{t}_{21} \times \mathbf{R}_{21}$, $\mathbf{E}_{21} \in \mathbb{R}^{3\times3}$ describes the euclidean transformation from camera coordinate system 2 to 1. This equation enables us to calculate the relationship between the two cameras directly from a set of image points that form a linear system of equations. In the absence of noise, this problem is trivial and can be solved with a minimum of seven corresponding points. When Gaussian noise is present, the problem may be formulated as a least-squares minimization problem [16], which we solve with the LMS (least median of squares) algorithm [17] to accommodate for outliers. The retrieved essential matrix can be decomposed with the singular value decomposition to obtain an estimate for $\mathbf{R}_{21}$ and $\mathbf{t}_{21}$ [18].

The resulting problem of finding an optimal estimation for 3D camera coordinates from two corresponding sets of 2D cameras can again be formulated as a linear system of equations, which we solve with Direct Linear Transform (DLT) and SVD [19].

E. Camera Pre-Selection

Since the HPE estimation on all 29 camera perspectives would take a lot of time, we perform a pre-selection to find a smaller subset of well-suited cameras. For that, we select one of the videos featuring three subjects. We first project the GT data onto each camera perspective using the provided intrinsic camera matrices $\mathbf{K}_c$ and extrinsic projection matrices $\mathbf{P}_{GT,c}$:

$$\tilde{\mathbf{x}}_{GT,c} = \mathbf{K}_c \begin{bmatrix} \mathbf{R}_{GT,c} & \mathbf{t}_{GT,c} \\ \mathbf{0} & 1 \end{bmatrix} \tilde{\mathbf{p}}_{GT} \tag{9}$$

As has been explained in Section II-D, reconstruction from two camera images would be trivial without noise. Therefore, we add noise to the projected GT coordinates. The noise is generated in accordance with the original BlazePose paper [3], where it was stated that 84.1% of all joints were within 20% of the upper body size. Since it is not clear what measure they used for body size, we decided to base it on the hip width, which we have shown to be relatively stable in a previous study [6]. To calculate the hip width, we average the Euclidean distance of the left and right hip joints over all frames where both of these joints lie within the image size. With the so-calcualted hip width d_{HW}, we define the noise such that exactly 84.1% of all joints have less noise than $0.2\ d_{HW}$:

$$\mathbf{n} \sim \mathcal{N}(0, 0.142\ d_{HW}), \tag{10}$$

Then, we again exclude all frames where not all noisy joints are located within the image size, undistort these points with the respective intrinsic camera matrices, and calculate the essential matrix $\mathbf{E}_{21}$. Finally, we reconstruct the 3D coordinates in the 3D coordinate system corresponding to camera 1 through the methods described in Section II-D.

For evaluation, we transform the coordinates into the original GT coordinate system through a least-squares optimization. The residuals of this final transformation, which can be interpreted as a mean root-mean-square error over all joints, are used as an accuracy metric to evaluate how well this particular set of cameras works for the 3D reconstruction. The rate of frames where not all joints were visible is also used as a metric.

III. Results and Discussion

A. Camera Pre-Seletion

We first want to evaluate the preliminary selection of suitable cameras. The process explained in Section II-E is executed for every possible set of two cameras. With 29 cameras, this means we evaluate 870 possible reconstructions. The example presented in Fig. 6 shows results of two reconstructions. As can be seen, even though both reconstructions have a low ratio of frames where not all joints are visible (0.017 in Fig. 6a vs.0.028 in Fig. 6b), the better reconstruction has a much lower residual (1.173 averaged over all three axes in Fig. 6a vs. 5.542 in Fig. 6b).

Fig. 3 shows a histogram visualizing the distribution of the rate of frames where not all 12 joints were visible over all 870 combinations of two camera views. As can be seen, some rates are much more prevalent than others. This is due to the fact that we do not evaluate single cameras but instead jointly evaluate sets of two cameras. Therefore, the likelihood is usually determined by the worse camera. Since the movements in the videos are quite static, the occurrences of a joint not being visible, are likely to be biased, i.e. repeatedly happening for the same joint. Therefore, a high rate of "joint invisibility" translates to certain joints being biased in the reconstruction. Thus, motivated by the first steep drop visible in the histogram, we opt to set a clear cut-off for what is an acceptable rate at 7 %. Any camera combination with a worse rate is not considered any further.

The remaining 240 camera combinations are evaluated by the residual of the least-square algorithm estimating the optimal rotation of the reconstructed 3D coordinates onto the GT coordinates. Fig. 4 displays the residuals over the rotation angle between the cameras, which is calculated as the norm of the rotation vector as defined in Section II-C. These angles are calculated from the rotation matrices provided by the dataset. We can see that a small angle ($\leq 45°$) corresponds to an

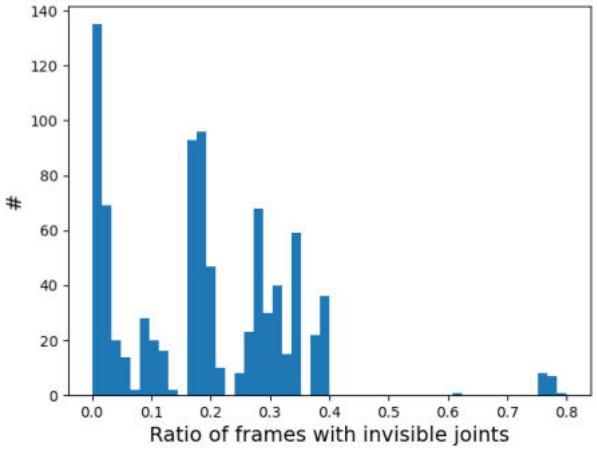


Fig. 3: A histogram visualizing the distribution of prevalence of frames where not all joints were visible over all 870 combinations of two camera views.

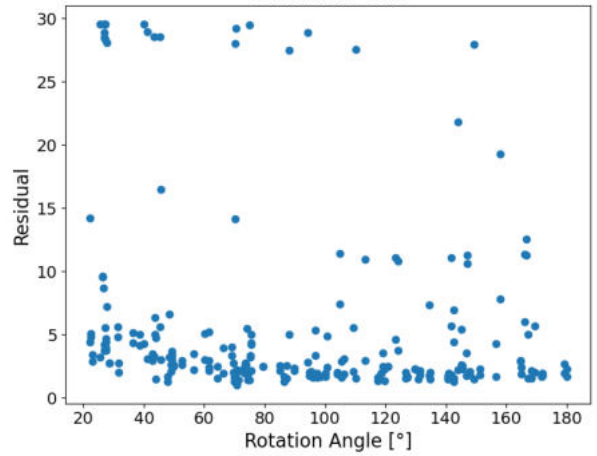


Fig. 4: A scatter plot for the camera pre-selection. Depicted is the residual from the final least-squares optimization over the angle calculated between the two cameras. The correlation is -0.23. The median value for the residuals is 2.68.

increased probability of high residuals. With increasing rotation angle, the amount of low residuals increases continuously with an overall correlation of -0.23. The median residual is given as 2.68, which is approximately 0.23 times the average hip width of 11.54. The minimum of 1.02 (0.0884 times the hip width) is achieved for the combination of the cameras *00_02* and *00_03*, which are rotated by an angle of $73.2°$. The corresponding reconstruction is shown in Fig. 6a. For greater angles, we can see a very slight upwards trend in the graph. We also notice that the residuals seem to be upper bound by a value of 30.

B. MediaPipe-Based Approach

To evaluate the MediaPipe-based 3D reconstruction, we select the 15 sets of cameras that heeded the best results in the pre-selection. In these 15 sets, there are 14 different

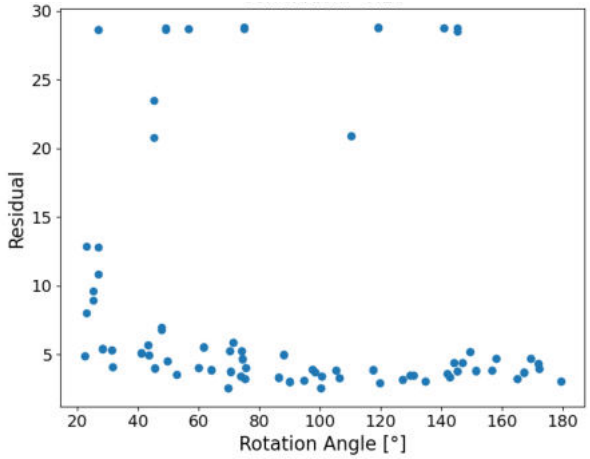


Fig. 5: A scatter plot for the MediaPipe-based 3D reconstruction. Depicted is the residual from the final least-squares optimization over the angle calculated between the two cameras. The correlation is -0.16, the median value for the residuals is 4.07.

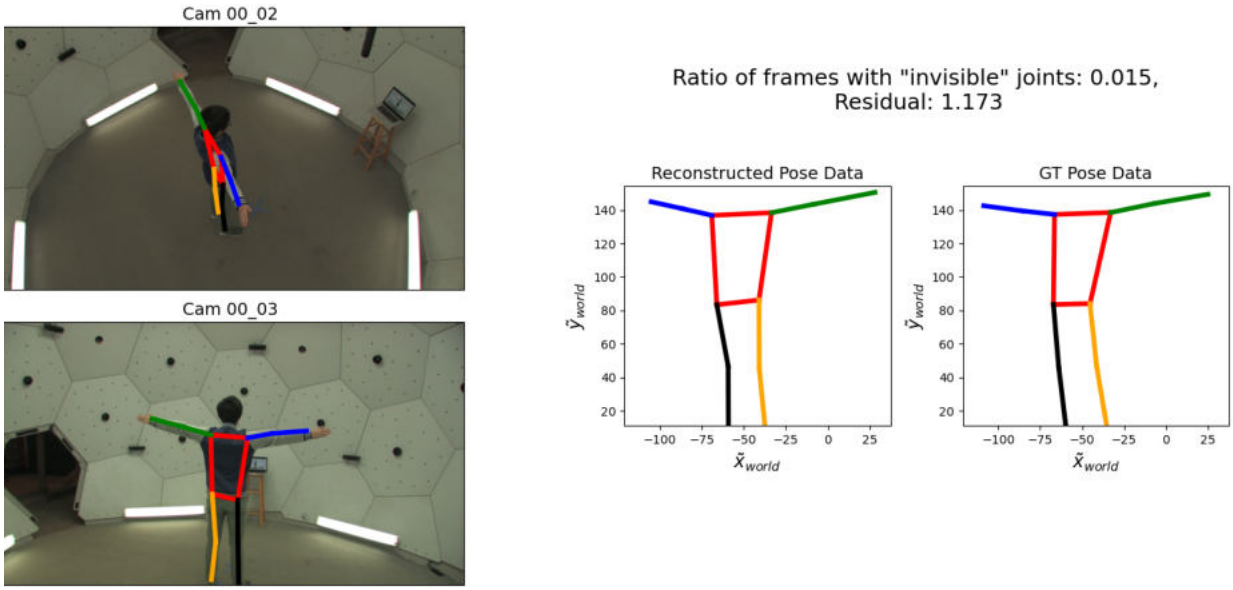
camera perspectives, from which we exclude two as they are not available for all recordings. We evaluate the remaining 12 cameras in every possible combination, resulting in 132 evaluations.

The first result is the far lower rate of frames with ”invisible“ joints. All combinations clear our previously set threshold of 7%. This is expected since we already filtered out cameras with a low visibility score in the pre-selection. Furthermore, MediaPipe has a tendency to always estimate the position of all joints when some part of the body is visible. This can be seen by lower visibility values MediaPipe attributes to the joints that are not actually in the frame.

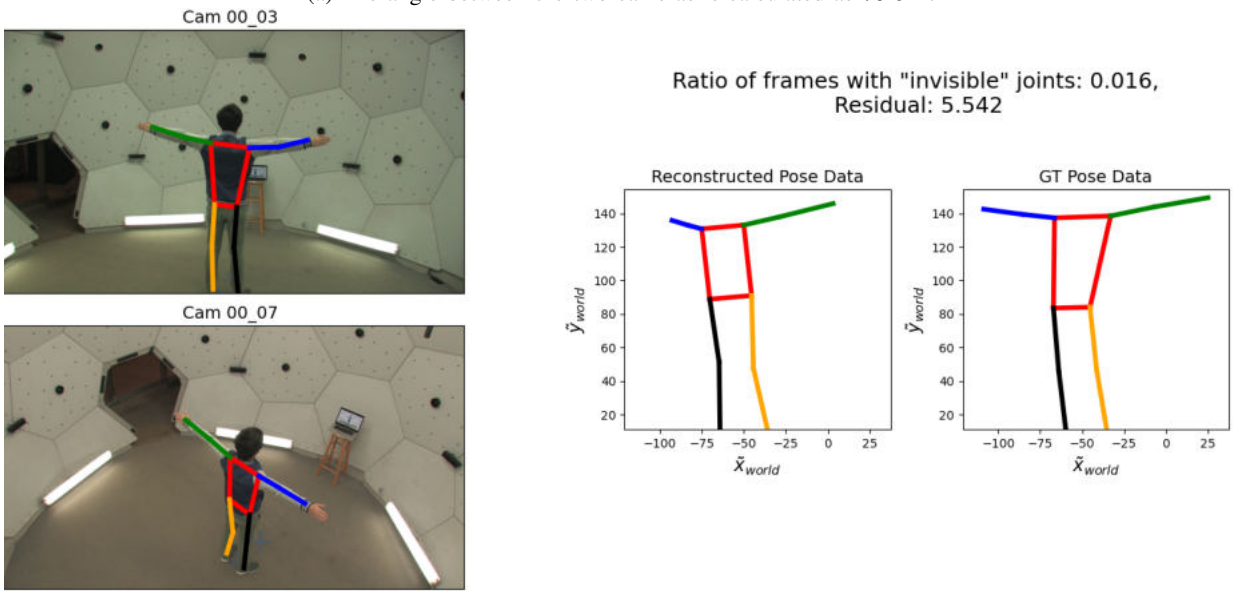
Fig. 5 shows the same scatter plot as before, depicting the residuals over the angle. We can see many similarities to Fig. 4, e.g. a higher probability for high residuals for small angles and an upper bound of 30 for the residuals. A local minimum of 2.551 can be seen for $69.6°$, but the global minimum of 2.538 (0.220 times the hip width) is achieved for the two cameras *00_23* and *00_22*, which are rotated by $100.1°$. The corresponding reconstruction is shown in Fig. 7. The camera set of *00_02* and *00_03*, which was found to be optimal in the pre-selection, achieves a residual of 3.754. With a correlation of -0.16, the statistical relationship between the angle and the residual is less pronounced, which might also be due to the fewer data points.

IV. Conclusion

In this work, we presented our research on how to employ MediaPipe Pose, epipolar geometry, and direct triangulation to reconstruct human 3D pose data from a stereo camera view. We used the 3D ground truth data provided in the CMU Panoptic dataset to generate artificial 2D pose estimations for 29 different camera views with a Gaussian noise model inspired by previous research on MediaPipe's accuracy. With



(a) The angle between the two cameras is calculated as $70.64°$.



(b) The angle between the two cameras is calculated as $44.04°$.

Fig. 6: Two examples of a set of two cameras and their corresponding reconstructions. On the left, the camera views are shown, with the projected GT with added noise drawn on top. On the right, the 2D projections onto the x-y-plane of the reconstructed 3D pose and the GT 3D pose data are shown. Even though both sets have a low ratio of frames with "invisible" joints, the one where the angle is higher, has a lower residual.

this artificial data, a pre-selection of the camera views was performed to find suiting pairs of cameras for further evaluation. To analyze the estimation accuracy we utilized the residuals of a least-squares optimization between the reconstructed 3D pose data and the ground truth. These residuals can be interpreted as the mean root-mean-square error over all joints and time. We could show a correlation of -0.23 between the so-calculated estimation accuracy and the rotation angle between the two cameras. The median values of the residuals is approximately 0.23 times the average hip width of the subject. The lowest residual of 1.02 is reached for the cameras *00_02* and *00_03*, which are rotated by an angle of $73.2°$.

The analysis of the MediaPipe-based approach shows similar results, but less pronounced. The median value of the residuals is significantly higher with 0.35 times the average hip width of the subject. The lowest residual is achieved for the cameras *00_06* and *00_18*, which are rotated by $100.1°$. The overperformance of the approach with artificially created 2D data might be caused by a too optimistic noise model. While some of MediaPipe's inaccuracies can be interpreted as

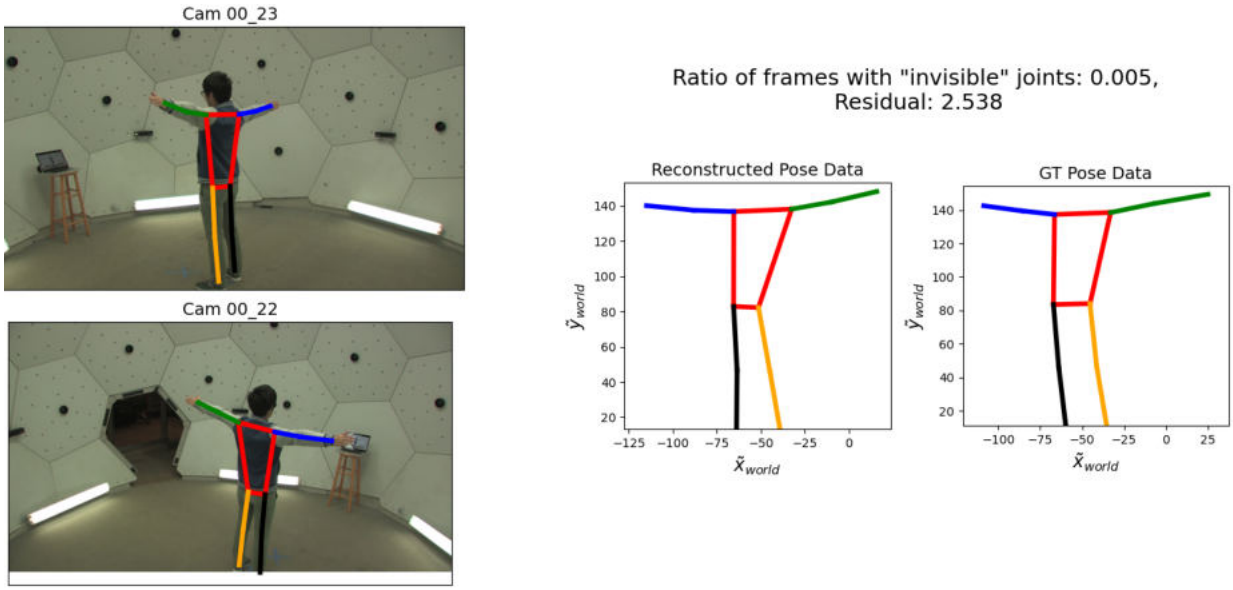


Fig. 7: The best reconstruction based on the MediaPipe approach. On the left, the camera views are shown, with the MediaPipe coordinates drawn on top. On the right, the 2D projections onto the x-y-plane of the reconstructed 3D pose and the GT 3D pose data are shown. The angle between the two cameras is calculated as $100.1°$.

Gaussian noise, other errors that have been observed but not modeled, include confusing left and right limbs and "seeing ghosts", i.e. predicting joints in regions without any person.

In future works, the dataset could be examined concerning the angle between the subjects' line of sight and the camera, which we have previously shown to play a big role in MediaPipe's estimation accuracy [6]. Furthermore, we did not utilize all the data available in the dataset. An expanded study utilizing the full dataset could yield more reliable results and research aspects like dependency on the subjects themselves. Lastly, the method's applicability to a real-world scenario should be evaluated. One possible application might be a physical therapy session, where it could enable patients to do parts of their therapy at home while getting detailed feedback by an automated evaluation system.

References

[1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[2] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," *CoRR*, vol. abs/1312.4659, 2013.

[3] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," 2020.

[4] P. Albuquerque, T. T. Verlekar, P. L. Correia, and L. D. Soares, "A spatiotemporal deep learning approach for automatic pathological gait classification," *Sensors*, vol. 21, no. 18, 2021.

[5] R. Mehrizi, X. Peng, D. N. Metaxas, X. Xu, S. Zhang, and K. Li, "Predicting 3-d lower back joint load in lifting: A deep pose estimation approach," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 1, pp. 85–94, 2019.

[6] S. Dill, A. Rösch, M. Rohr, G. Güney, L. D. Witte, E. Schwartz, and C. H. Antink, "Accuracy evaluation of 3d pose estimation with mediapipe pose for physical exercises," *Current Directions in Biomedical Engineering*, vol. 9, no. 1, pp. 563–566, 2023.

[7] M. Burenius, J. Sullivan, and S. Carlsson, "3d pictorial structures for multiple view articulated pose estimation," pp. 3618–3625, 06 2013.

[8] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele, "Multi-view Pictorial Structures for {3D} Human Pose Estimation," in *Electronic Proceedings of the British Machine Vision Conference 2013* (T. Burghardt, D. Damen, W. Mayol-Cuevas, and M. Mirmehdi, eds.), (Bristol, UK), pp. 1–12, BMVA Press, 2013. 24th British Machine Vision Conference.

[9] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Harvesting multiple views for marker-less 3d human pose annotations," *CoRR*, vol. abs/1704.04793, 2017.

[10] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, "Cross view fusion for 3d human pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4342–4351, 2019.

[11] X. Chen, K.-Y. Lin, W. Liu, C. Qian, and L. Lin, "Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10895–10904, 2019.

[12] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu, "Epipolar transformers," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 7779–7788, 2020.

[13] E. Gisleris and A. Serackis, "Enhancing 3d pose estimation accuracy from multiple camera perspectives through machine learning model integration," in *2023 IEEE 10th Jubilee Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, pp. 1–4, 2023.

[14] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[15] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 ed., 2004.

[16] R. I. Hartley and P. Sturm, "Triangulation," *Computer Vision and Image Understanding*, vol. 68, no. 2, pp. 146–157, 1997.

[17] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American statistical association*, vol. 79, no. 388, pp. 871–880, 1984.

[18] J. Dongarra, M. Gates, A. Haidar, J. Kurzak, P. Luszczek, S. Tomov, and I. Yamazaki, "The singular value decomposition: Anatomy of optimizing an algorithm for extreme scale," *SIAM Review*, vol. 60, no. 4, pp. 808–865, 2018.

[19] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.