



EÖTVÖS LORÁND UNIVERSITY

FACULTY OF INFORMATICS

DEPT. OF ARTIFICIAL INTELLIGENCE

Deep NRSFM for Multi-view Multi-body Pose Estimation

Supervisor:

Fóthi Áron

PhD Student

Author:

Lu Fengjiao

Computer Science Msc

Budapest, 2023

Abstract

This paper addresses the challenging task of unsupervised relative human pose estimation. We leverage the potential offered by multiple (2-3) uncalibrated cameras. It is assumed that spatial human pose and camera parameter estimation can be solved as a block sparse dictionary learning problem with zero supervision. The resulting structures and camera parameters can be used to fit individual skeletons into a common space. To do so, we leverage the fact that all individuals in the image are viewed from the same camera viewpoint, thus taking advantages of the information provided by multiple camera views and overcoming the lack of information on camera parameters. To the best of our knowledge, this is the first solution that does not necessitate either 3D ground truth or knowledge of the camera parameters, whether intrinsic or extrinsic. Our approach demonstrates the potential of using multiple viewpoints to solve challenging computer vision problems.

Contents

1	Introduction	3
2	Related Work	6
2.1	Multi-View Single-Person Pose Estimation	6
2.2	Multi-View Multi-Person Pose Estimation	7
2.3	Non-rigid Structure from Motion	8
3	Proposed Method	9
3.1	Motivation to the Proposed Method	9
3.2	Problem Description	10
3.3	Non-rigid Structure from Motion	10
3.4	Single Pose Estimation	11
3.5	Procrustes Transformation	12
3.6	Orthogonal Projection	13
3.7	Multi Pose Estimation	13
4	Ablation Studies	15
4.1	Introduction	15
4.2	Methodology	15
4.2.1	Eight-point Estimation Method of Fundamental Matrix	15
4.2.2	Triangulation Reconstruction and Bundle Adjustment	17
4.3	Results and Conclusion	18
5	Experiment	19
5.1	Datasets and Parameter Settings	19
5.2	Evaluation Metrics	20
5.3	Main Results	21
5.3.1	Results on CMU Panoptic Dataset	22
5.3.2	Results on In the wild Dataset	23

CONTENTS

5.3.3 Discussion	23
6 Conclusion	25
Acknowledgements	26
Bibliography	26
List of Figures	30
List of Tables	31

Chapter 1

Introduction

In certain scenarios, such as sports broadcasts and CCTV recordings in public spaces, multiple camera perspectives may be accessible, yet the associated camera parameters remain unknown. Despite the absence of these parameters, it is still essential to interpret what happened in the recording, especially regarding the physical interaction between two or more individuals. However, due to optical illusions and occlusion of the human skeleton, this can be a challenging task even for the human eyes, let alone for machine vision-based solutions.

In this context, the use of machine vision-based solutions offers a potential solution to the problem, but it requires an accurate estimation of camera parameters. Remarkably, if several camera views are available, it is possible to take advantage of them to give a more accurate estimate of the problem at hand. Despite the advantages of using multiple camera views, accurately estimating camera parameters remains a critical challenge. In cases where only single camera view is available, a previous study [1] demonstrated that the task can be solved without knowing the parameters of the camera, by applying the non-rigid structure from motion procedure and learning the statistics with a directory learning approach. By utilizing neural shape prior proposed by [1], a hybrid multi-view 2D-3D lifting architecture that combines multi-view geometry and monocular deep 3D lifting approaches was proposed by [2] .

However, there is still a gap in methods using multiple views for multi-body pose estimation. Our proposed method fills this gap by taking advantage of multiple views and non-rigid structure from motion to perform multi-person body pose estimation.

Simultaneously, it is worth noting that the majority of current multi-person pose

estimation techniques necessitate 3D ground truth for model training. However, obtaining such ground truth in real-world scenarios can be resource-intensive and pose significant challenges.

To address these existing challenges, we have proposed a novel method Multi View Multi Body Non-rigid Structure from Motion(MVMB NRSfM). Our approach builds upon the established deep non-rigid structure from motion (NRSfM) model [1], which facilitates the estimation of 3D human pose without requiring camera parameters while simultaneously providing predicted camera parameters. However, the global coordinate systems for each individual skeleton differ, resulting in an undefined relative pose between any two skeletons. To address this issue, we utilize these predicted camera parameters to compute the rotation matrix between two distinct skeletons, taking into account the constant relative pose of the camera and allowing us to use the information provided by multiple camera views. Through incorporating the 2D ground truth as our loss function, we can effectively determine the translation matrix required for the alignment of the skeletons. By applying both the rotation and translation matrices, we can successfully obtain the relative poses between two skeletons. (Seeing in Figure 1.1) This enables us to estimate the relative pose between two human skeletons with enhanced accuracy and precision.

We evaluate our approach on a large scale datasets CMU Panoptic [3] as well as on custom video recordings captured in-house. Our evaluation metric, Procrustes aligned group mean per-joint position error(PAG-MPJPE), enables a thorough examination of the effectiveness of our approach. Remarkably, our method demonstrates high performance even under constrained conditions, such as the utilization of merely three uncalibrated cameras and the absence of 3D ground truth. This is particularly evident in the accurate estimation of the relative pose between two distinct skeletons, highlighting the robustness and adaptability of our technique in various scenarios.

Our proposed approach inherits many favorable properties from Multiple View Non-rigid Structure from Motion(MV NRSfM) [2]. It assumes they are known 2D projected measurements, so the pixel intensity is not used directly. Therefore, our approach is integrable with any available 2D keypoint detector such as HRNet [4], Stacked Hourglass Networks [5], ViTPose [6] and others. Finally, the camera is not static, so the proposed approach is agnostic to camera movement. Therefore, all we require is a synchronized multi-camera setup. Yin et al. [7] proposed a self-supervised

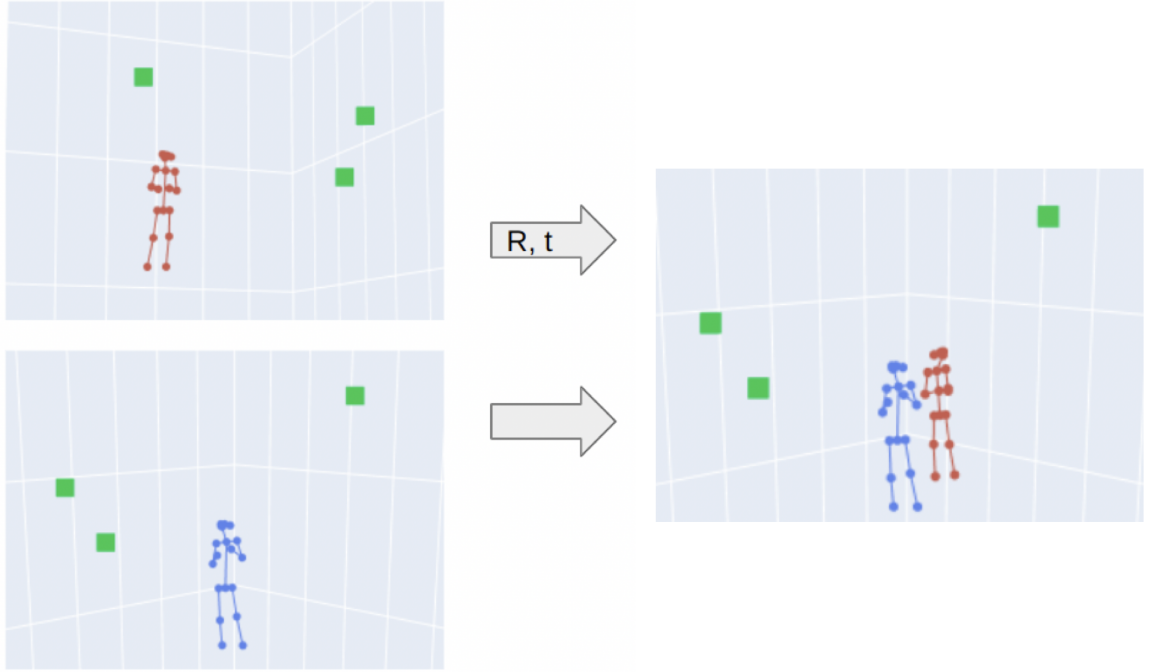


Figure 1.1: The method we propose simultaneously estimates each pose (reference in blue and other in red) and the corresponding camera parameters (estimated cameras as green rectangles), then, knowing that these cameras are located in the same position, we move the individual skeletons into the same space. In this way, we get their position relative to each other

framework for the automatic video synchronization, which is ideally suited to address this problem.

The main contributions of this paper are summarized as:

- Taking advantage of the fact that the predicted camera parameters refers to the same camera, the individual skeletons are rotated into a common space.
- Only the possibilities provided by multiple three uncalibrated cameras are used, highlighting the potential of leveraging multiple perspectives to overcome the lack of information about the camera parameters.
- Throughout the procedure, we are utilizing 2D ground truth rather than 3D ground truth. This presents a significant advantage, as acquiring 3D ground truth is frequently quite challenging in real-world scenarios.

Chapter 2

Related Work

In this section, we review the literature related to some concept of this paper, those literature all related to multiple view 3D human pose estimation.

2.1 Multi-View Single-Person Pose Estimation

Multi-View Single-person estimation refers to the task of estimating the 3D pose of single person from multiple camera views. Utilizing multiple camera views can provide complementary information of a 3D scene, and thus mitigate projective ambiguities. This approach is particularly effective in resolving issues of human skeleton occlusion.

Considerable research has been conducted to address the task of estimating the 3D pose of a single person in the context of multi-view scenarios. Remelli et al. [8] proposed a novel multi-camera fusion technique that leverage 3D geometry in latent space to efficiently and jointly reason about different views and drastically improve the accuracy of 2D detectors. Iqbal et al. [9] present a weakly-supervised approach to monocular 3D pose estimation using unlabeled multi-view data and images annotated with 2D poses. Rhodin et al. [10] presents a novel approach to 3D human pose estimation from single images, leveraging multiple views at training time instead of relying solely on large, manually-annotated datasets. This method uses a multi-view consistency loss and a regularization loss, encouraging prediction of consistent 3D poses across views and maintaining classifier behavior. A method was proposed by [11] to estimate 3D human poses from multiple calibrated cameras. This method first uses a CNN-based multi-view feature fusion technique to enhance the accu-

racy of 2D pose estimation, then employs a recursive pictorial structure model to convert multi-view 2D poses into 3D poses. Skakov et al. [12] introduces a learnable triangulation technique, which adapts per-view confidence weights for standard triangulation, and a volumetric approach that consolidates multi-view images to conduct 3D pose estimation within a 3D volumetric space.

2.2 Multi-View Multi-Person Pose Estimation

Multi-view Multi-person pose estimation aims to localize 3D skeleton joints for each person instance in a scene from multi-view camera inputs. There are various methods for multi-person 3D pose estimation, which involves identifying the 3D positions of multiple people’s body parts in space over time.

One common approach involves breaking the problem down into two steps: first, grouping the 2D joint positions in each frame, and then using that information to reconstruct the 3D poses. There are two significant obstacles: associating joints of the same person and 2D poses of the same person across different views. Previous approaches have used various strategies, including bottom-up [13] or top-down [14, 15] grouping and pictorial structure models [16] or triangulation for 3D pose reconstruction. However, these methods possess constraints, as incorrect correspondences can impact accuracy. Then newer methods have been developed to address these issues.

One such approach involves projecting perview 2D joint heatmaps into a 3D space and detecting people and predicting their 3D poses directly in that space. Another approach involves encoding both 2D and 3D pose features and decoding 3D poses of all individuals in the scene. Zhang et al. [17] proposed a method directly regresses the multi-person 3D poses without relying on intermediate tasks. The method represents skeleton joints as learnable query embeddings and uses a hierarchical scheme and a novel geometrically guided attention mechanism to improve accuracy. Chen et al. [18] propose a method to overcome the limitations in dense crowd circumstances caused by the mismatch of human joints due to the simplicity of cues provided by the Euclidean distances between joints and epipolar lines, as well as the lack of robustness in the naive least squares minimization formulation, involves utilizing a graph model for efficient cross-view matching and a maximum a posteriori estimator for 3D human pose reconstruction. Ershadi-Nasab et al. [19]

proposes a fully connected pairwise conditional random field to estimate the number of humans in the scene and constructs search spaces for each person, with the 3D pose approximated using the loopy belief propagation inference algorithm. This method rely on 3D ground truth data during training. A new adversarial framework was proposed by [20] for estimating the association of multiple humans in an uncalibrated multi-view camera system, By using a generator and discriminator network trained by 3D ground truth, the method improves pose estimation accuracy, addresses occlusion and self-occlusion issues, and does not require prior calibration information.

2.3 Non-rigid Structure from Motion

Non-rigid structure from motion is a technique designed to reconstruct the 3D structure of an object that is undergoing deformation based on the 2D correspondences that are observed from several different viewpoints [21]. There are some limitations of current NRSfM regard to the number of images and shape variability they can handle, which hinder their practical use in computer vision applications. To slove this problem, Kong et al. [1] proposes a new deep neural network that can recover camera poses and 3D points solely from 2D image coordinates, without the need for 3D supervision.

Several approaches have been proposed to address multi-body pose estimation using the NRSfM method. Kumar et al. [22] proposed a new framework leverages subspace clustering to simultaneously reconstruct and segment each 3D trajectory, which involves reconstructing and segmenting multiple 3D structures of non-rigid objects or deformable scenes from images. However these methods do not use the possibilities provided by multiple cameras. Another method proposed by Huang et al. [23] introduces a framework for directly recovering human motions and extrinsic camera parameters from sparse multi-view videos, using physics-geometry consistency and a latent motion prior for joint optimization. This method does not rely on camera calibration data or spatial data, while still requiring consecutive frames for camera calibration. In contrast to this approaches, our method is capable of handling a variety of materials and is adapted to constantly changing camera parameters.

Chapter 3

Proposed Method

In this chapter, we concentrate on the fundamental principles underlying our Proposed method.

3.1 Motivation to the Proposed Method

The motivation for our proposed methodology comes from the current constraints of multi-person multi-view pose estimation research. One constraints is the needed of 3D ground truth. Despite there are a large mount of the state of art model in this field, most existing models still necessitate 3D ground truth. This requirement not only implicates substantial resource expenditure but also presents a considerable challenge due to the difficulty of accurately capturing 3D spatial points of the human body.

Furthermore, current methodologies largely depend on a high quality muti view camera set ups. This necessitates knowledge of specific camera calibration, which means should include both intrinsic and extrinsic camera parameters. Procuring these parameters is a time-intensive task, further complicated by the fact that, in some instances, these parameters may be hard to acquire.

Therefore, we initiated an exploration aimed at addressing these inherent issues. Through the method proposed by [1], we discovered the potential of utilizing a multi-layer sparse dictionary through Non-Rigid Structure from Motion (NRSfM). This approach enables us to train models without the dependency on 3D ground truth. Dabhi et al. [2] demonstrated the possibility of employing a limited (2-3) number of

uncalibrated camera views to successfully facilitate the 3D reconstruction of a single object.

These articles inspired us to think about how we could take advantage of these advantages to achieve multi-person, multi-body pose estimation. Consequently, this led to the inception of our proposed method.

3.2 Problem Description

Through the application of the MV NRSfM method, we successfully got pose estimates and predicted camera parameters specific to an individual, but these results are under different coordinate systems, which meaning that the relative pose between them remains unknown. We intend to determine the relative poses of N non-rigid object sets in a space observed with 3 synchronized cameras. An object set contains M objects, depending on the given frame it can be 2 or more objects. These objects are described as key-points. The planar coordinates of the key-points (we can also say it as joint points of the human skeletons) of M objects are obtained per frame. We have to construct the spatial key-points of the objects based on these plane coordinates, in order to be able to determine their relative position. Formalized we have $\{W_1^{(1)}, \dots, W_K^{(N)}\}$ observations and each W is a group of objects $\{V_1, \dots, V_M\}$ where $V \in \mathbb{R}^{p \times 2}$ and represents the object with the plane coordinates of the key points. Our task is to reconstruct the spatial representation of the object groups $\{G^{(1)}, \dots, G^{(N)}\}$. In this context, each G represents a group of objects. $\{S_1, \dots, S_M\}$ where $S \in \mathbb{R}^{p \times 3}$

3.3 Non-rigid Structure from Motion

Non-rigid structure from motion is a method of restoring time-varying 3D structures from their projections. The foundations of this solution based on factorization were laid by Tomasi and Kanade [24] on non-rigid objects and were later extended to non-rigid bodies by Bregler [21]. The basis of the procedure was that these objects can be described by a linear combination of the elements of a compact set containing basic shapes. Kong et al. [1] traced this problem back to dictionary learning, where basic forms encoded with a sparse representation can be considered as dictionary words and Dabhi et al. [2] showed that in multiple camera cases it is possible to

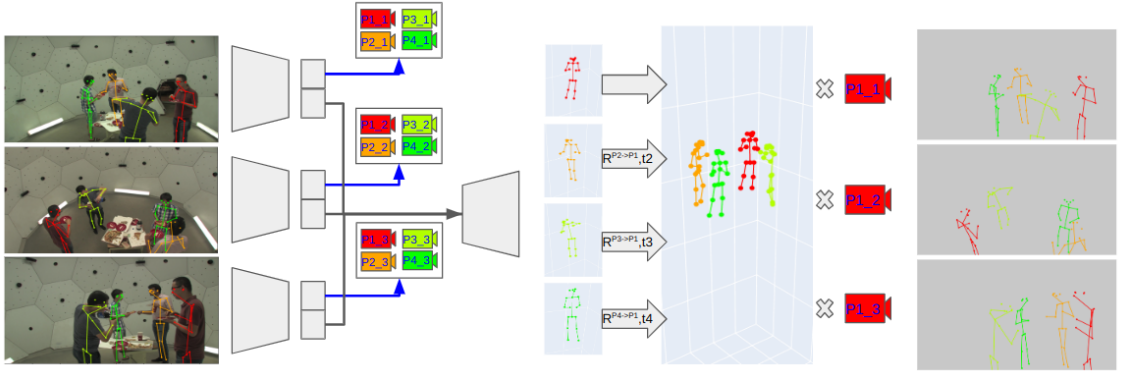


Figure 3.1: **Pipeline of the proposed method.** The method can be divided into two phases. In the first phase, three image inputs are processed through an encoder, yielding a set of 2D projections, these 2D projections are merged to estimate the 3D structures(MV NRSfM method). While in the second phase, we ascertain relative poses through the computation of rotation and translation matrices. The Procrustes algorithm is utilized to calculate relative rotation while relative translation is optimized from the minimization of the re-projection error.

compete with the accuracy of expensive, calibrated camera systems. Our solution is based on this idea: although the joint position of several people cannot be solved on a statistical basis, the structure of individual people can be obtained by applying dictionary learning. After that, only these spatial structures need to be found in their proper position in a common space.

3.4 Single Pose Estimation

We followed Dabhi et al. [2] method to estimate the pose of individual objects. We developed our PyTorch implementation.

Encoder-Decoder Architecture For the encoder and decoder architecture they use a convolutional network and share the convolution kernels between the encoder and decoder. The convolution kernels perform the role of dictionaries. The 2D projections are the input of the encoder that extracts the block sparse code from the corresponding views. A linear layer, called Rotation Factorization (RF) layer, factorizes it into the camera matrix $\mathbf{R}_k \in \mathbb{R}^{2 \times 3}$ for camera k , and the block sparse representation of the spatial structure. These representations are combined into a common representation, from which the decoder produces the spatial representation of the objects $\mathbf{S} \in \mathbb{R}^{p \times 3}$

Loss Function The learning objective is to minimize re-projection error between

input 2D points $\mathbf{V}$ and re-projected 2D points $\mathbf{SR}$ while simultaneously encouraging $\mathbf{R}$ matrix to retain its orthonormality. We achieve this by hardcode the singular values of $\mathbf{R}$ to be exact using Singular Value Decomposition (SVD). So our loss function looks like this

$$\mathcal{L}_{\text{Reproj}} = \frac{1}{NK} \sum_{k=1}^K \sum_{n=1}^N \|\mathbf{v}_k^{(n)} - \mathbf{s}^{(n)} \tilde{\mathbf{R}}_k^{(n)}\|_F \quad (3.1)$$

Where $\tilde{\mathbf{R}} = \mathbf{U}\mathbf{V}^T$ and $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{R}$

3.5 Procrustes Transformation

In the field of computer vision, the Procrustes Transformation is a crucial tool used for aligning shapes and images. The transformation involves three main steps: translation, scaling, and rotation, or a combination of these transformations.

- Translation: The shape is moved so that its centroid (geometric center) coincides with the origin of the coordinate system.
- Scaling: The size of the shape is uniformly adjusted so that the sum of the squares of the distances of the points from the origin is equal to one.
- Rotation: The shape is rotated to minimize the sum of the squared distances between corresponding points in the two shapes.

The goal of these operations is to minimize the difference between two shapes (or datasets), thus allowing for accurate comparisons and analyses. The transformation does not distort the shapes, it merely brings them into alignment, preserving their local properties. This characteristic makes the Procrustes Transformation a useful tool in computer vision for tasks such as object recognition, pose estimation, and image registration, where the relative arrangement of points is more important than their absolute positions.

In our proposed method, we use Procrustes algorithm to estimate an orthonormal matrix for aligning the predicted 3d points to the ground truth, which ensures the global optimality of the solution.

3.6 Orthogonal Projection

Orthogonal projection is a fundamental concept in linear algebra and geometry. It refers to the process of projecting a point or a vector from a higher dimensional space onto a lower dimensional subspace, such as a line, a plane, or a hyperplane. Orthogonal projection plays a crucial role in the field of computer vision. In computer graphics and vision, it's often necessary to map 3D world points to 2D image points, especially when rendering 3D models onto a 2D screen or when performing object recognition. The process of mapping these points involves orthogonal projection, where the 3D points are projected onto a 2D plane (image plane) perpendicularly.

In our proposed method, after we obtain the rotated predicted 3D shapes and predicted camera matrix, We then apply orthogonal projection to map these 3D points back onto a 2D plane. In this way we can utilize the error between 2D ground truth and reprojected 2D points.

3.7 Multi Pose Estimation

With the poses associated with each object, we can determine the rotation and displacement by applying the re-projection error function, which gives their relative position in the common space. We do this in two steps

Relative Rotation

There are N cameras, and the extrinsic camera parameters can be estimated by the MV NRSfM method. Moreover, pose parameters can be obtained for each skeleton model, separately. The global coordinate systems are different for each skeleton, but the relative poses between cameras are the same. For example, the relative pose between the r -th and o -th cameras are as follows:

$$R_{r \rightarrow o} = R_o R_r^T$$

as $R_r^{-1} = R_o^T$.

If we pick up two skeletons, e.g. the original one o and the reference one r , then there is a global rotation transforms the corresponding camera rotations: $R_i^o = R^{r \rightarrow o} R_i^r$, for all $i \in (1, \dots, N)$. Both the relative pose and the transformation between skeletons can be solved using the Procrustes method. Let us consider the second problem class when $R^{r \rightarrow o}$ should be estimated. In the least squares since,

the algebraic

$$R_{r \rightarrow o} = R_o R_r^T$$

problem can be written as

$$R^{r \rightarrow o} = \arg_R \min \|R^o - R^{r \rightarrow o} R^r\|^2,$$

where $\|\cdot\|^2$ is the square of the so-called Frobenius norm of a matrix. This is the special case of the so-called point registration problem [25], it can be optimally solved by the the singular value decomposition of matrix H : $H = USV^T$ where

$$H = (R^o)^T R^r,$$

and

$$R_r = \begin{bmatrix} R_1^r \\ R_2^r \\ \vdots \\ R_N^r \end{bmatrix}, R_o = \begin{bmatrix} R_1^o \\ R_2^o \\ \vdots \\ R_N^o \end{bmatrix}.$$

The optimal solution is then $R^{r \rightarrow o} = UV^T$.

Relative Translation

After we have rotated the objects into a common space, we determine the displacement vector $\mathbf{t} \in \mathbb{R}^3$ for optimization.

We take $\mathbf{S}^{(r)}$ and $\mathbf{S}^{(o)}$ from the same object group as previously determined by single pose estimator. Then, as discussed in the previous point, we define $\mathbf{R}^{(o \rightarrow r)}$, which rotates $\mathbf{S}^{(o)}$ into the space of $\mathbf{S}^{(r)}$. Next, we need to determine $\mathbf{t}^{(o \rightarrow r)}$, which moves $\mathbf{S}^{(o)} \mathbf{R}^{(o \rightarrow r)}$ to the correct location relative to $\mathbf{S}^{(r)}$. This can be achieved by minimizing the following re-projection error function by optimizing $\mathbf{t}^{(r)}$:

$$\mathcal{L}_{\text{pr}} = \frac{1}{K} \sum_{k=1}^K \|\mathbf{V}^{(r)} \oplus \mathbf{V}^{(o)} - \mathbf{S}^{(r)} \mathbf{R}_k^{(r)} \oplus (\mathbf{S}^{(o)} \mathbf{R}^{(o \rightarrow r)} + \mathbf{t}^{(r)}) \mathbf{R}_k^{(r)}\|_F \quad (3.2)$$

Chapter 4

Ablation Studies

Before introducing our proposed method, we conducted several experiments using other techniques. In this section, we outline the first experiment we conducted, which aimed at achieving multi-view pose estimation for multiple body using traditional triangulation methods.

4.1 Introduction

In this experiment, we utilized the [1] method to estimate the 3D skeleton and camera parameters. We then projected the 3D skeleton to 2D using the OnP method [26], and used the resulting 2D coordinates for our multi-person pose estimation, applying stereo 3D triangulation reconstruction and Bundle Adjustment algorithms. With this experiment, we also obtained good 3D reconstruction of the human skeleton. However, throughout this process, camera calibration was still necessary to acquire the K matrices, which were then used to calculate the essential and fundamental matrices required for the stereo triangulation reconstruction.

4.2 Methodology

4.2.1 Eight-point Estimation Method of Fundamental Matrix

Normalize Data To maintain numerical stability and ensure the reliability of the method, we normalize the data before applying the eight-point method to compute the fundamental matrix. This normalization involves a translation process where

the origin is moved to the center of gravity, and a scaling process where the average point distances are adjusted to be $\sqrt{2}$ within the two sets of points.

Set up a system of Equations

For each pair of matching points p_1 in the first image and p_2 in the second image, we can create an equation of the form:

$$p_2^T F p_1 = 0$$

Where F is the 3x3 Fundamental Matrix, and p_1 and p_2 are homogeneous coordinates $(x, y, 1)$ of the corresponding points. This can be rewritten as a linear equation in the entries of F :

$$\begin{bmatrix} p_{2x} \\ p_{2y} \\ 1 \end{bmatrix}^T \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} p_{1x} \\ p_{1y} \\ 1 \end{bmatrix} = 0$$

Solve the system of Equations

We now have a system of homogeneous linear equations of the form $Af = 0$, where A is the matrix of coefficients obtained from the previous step and f is a vector of the entries of F . This system can be solved by applying singular value decomposition (SVD) to A : $A = UDV^T$. The solution corresponds to the singular vector associated with the smallest singular value, which is the last column of V , v_9 . This vector v_9 can then be reshaped into a 3x3 matrix to obtain the estimated fundamental matrix F .

Denormalize Fundamental Matrix

After we got our fundamental matrix, we need to denormalize it to get back a Fundamental Matrix that is correct for the original, non-normalized image points. This step is important to ensure the resulting Fundamental Matrix can be correctly used with the original image points in subsequent processing. We use the following formula for denormalize :

$$\hat{F} = T_2^T F T_1$$

Here T_1 and T_2 represent the scaling factors obtained during the normalization process.

4.2.2 Triangulation Reconstruction and Bundle Adjustment

Essential matrix and Decompose it

Once the fundamental matrix is obtained and the intrinsic parameters (K matrices) are known from camera calibration (in this case, we're using ground truth camera values here), we can generate the essential matrix using the following formula:

$$E = K^T F K$$

After obtaining the essential matrix, we can decompose it using Singular Value Decomposition (SVD): $E = UDV^T$. This allows you to derive two potential rotation matrices: $R_1 = UWV^T$, $R_2 = UW^TV^T$ (W is a 3x3 matrix) and two possible translation vectors based on the matrices U and V resulting from the SVD process.

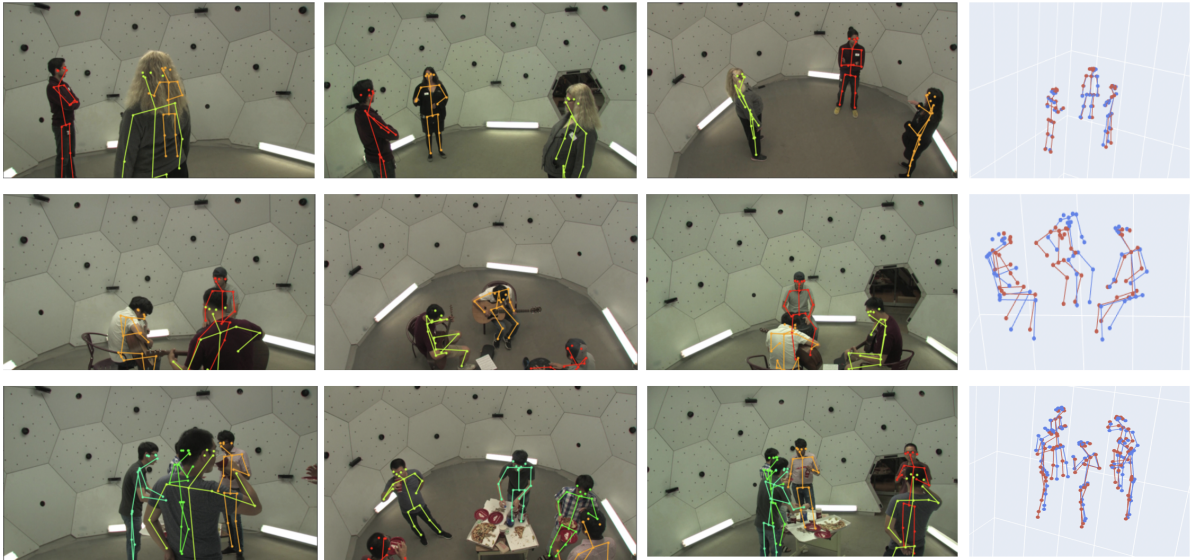


Figure 4.1: The left 3 images are the input, the right is the estimated Pose

Triangulation Reconstruction

Once we've obtained the rotation and translation matrices, we compute the projection matrices for both cameras as well as the fundamental matrices associated with each camera setup. Subsequently, for every point, the Hartley and Sturm triangulation method, along with the linear triangulation technique, is employed to estimate the point's 3D location. Then we use the Bundle adjustment to verify if the estimated 3D point is visible to both cameras. If this condition is met, the counter for the counter corresponding to that particular camera setup is incremented. Finally, we got the estimated 3D coordinates of each point in the skeleton.

4.3 Results and Conclusion

We’ve also implemented this method for evaluation using the CMO Panoptic dataset, and the results obtained from this process are presented here. Seeing in Figure 4.1

This experimental findings suggest that this method is quite effective in 3D reconstruction tasks, though it does exhibit certain limitations:

- We observed that the method performs optimally when the image include more human skeletons(which means more 2D keypoints can be used to calculate). However, the stability of the reconstruction diminishes when the image has fewer human skeletons.
- This method still necessitates camera calibration to derive the K-matrix, which is essential for 3D reconstruction.

After analyzing the outcomes of this experiment, we aim to address the identified limitations. Consequently, we are proposing a new method to overcome these challenges.

Chapter 5

Experiment

The purpose of our work was to determine the relative position of people. Thus, during the evaluation, the main aspect was to minimize the deviation of the group members compared to the group. The goal is not to minimize the absolute error, but neither is it to examine individual members of the group.

5.1 Datasets and Parameter Settings

In this section, we show the Experimental details and evaluation results on different multi view multi person datasets.

CMU Panoptic is a large scale Multi-View datasets specifically targeting 3D human pose estimation and activity recognition, which contains 60 hours of data with 3D poses and tracking information captured by 31 synchronized high-quality cameras and 480 VGA cameras from multiple camera viewpoints. The datasets also includes corresponding ground truth 3D poses and annotations. During the training, we sampled 3 cameras and in the reconstruction part we used fixed 3 cameras following [3].

In the Wild We acquired the datasets by capturing a video using three synchronized cameras and subsequently utilized this video in our proposed method. In our experiments we used one Samsung Galaxy S22 Ultra as leader and two Samsung Galaxy S20 Plus smartphones as clients and located them at different height level. Camera positions and orientations have been chosen in order to minimize the case of self-occlusions. We used ultra-wide camera sensors of all devices at 30 fps. Due to the selected camera sensors, perspective distortion occurs at the

edges of the frames, but it can be eliminated using the distortion coefficients to get them we made chessboard based camera calibration. Synchronous recordings were taken by OpenCamera-Sensors Android application, managed by the leader smartphone, then the actual synchronization happened during post-processing based on RecSync Android’s [27] provided scripts. We found that to achieve the proposed sub-millisecond synchronization(reference) all recording devices and settings should be identical. We used the ViTPose [6] algorithm to detect 2D landmarks, which also perfectly detects self-occluded key points.

Parameter Settings In order to optimize the hyperparameters for our single pose estimation network architecture, we employ hyperparameters optimization algorithm. This assists in finding the best weight for the model. Given that our method doesn’t depend on 3D ground truth, we utilize 2D ground truth for validation during the optimization process. In our hyperparameter optimization process, we focused on optimizing three types of hyperparameters: dictionary size (experimenting with {7, 8, 9}), learning rate (testing values {0.0001, 0.0002, 0.0003}), and exploration gamma (with values {0.99, 0.98, 0.97}) (Seeing in Table 5.1). During the training phase, we employ the Adam optimizer and apply an exponential decay with the exploration gamma to adjust the learning rate.

During the hyperparameter optimization stage, we divided the parameters to be optimized into two parts. One part was trained using three NVIDIA GeForce RTX 2080 Ti GPUs, while the other part was also trained on Azure servers using three K80 GPUs.

5.2 Evaluation Metrics

Our solution was motivated by a special problem, which is the examination of the social behavior of several people using the analysis of body poses. This requires a special evaluation method, which is different from the ones used before.

Mean Per-joint Position Error (MPJPE) is a widely used metric in the field of 3D human pose estimation for performance evaluation. It measures the average Euclidean distance between the predicted and ground-truth 3D joint positions in the dataset, after aligning the root joints of the estimated and groundtruth 3D pont set. The MPJPE is expressed in millimeters, with lower values indicating higher

Table 5.1: Hyperparameters Optimazation Results, from this table we can find that the best hyperparameters is learning rate is 0.0002, gamma is 0.97, and the dictionary size is 7

hyperparameters	size 7	size 8	size 9
lr 0.0001 & gama 0.99	0.28870	0.29434	0.28415
lr 0.0001 & gama 0.98	0.27777	0.29962	0.28124
lr 0.0001 & gama 0.97	0.28058	0.29012	0.27919
lr 0.0002 & gama 0.99	0.16349	0.27811	0.32828
lr 0.0002 & gama 0.98	0.15229	0.27427	0.31928
lr 0.0002 & gama 0.97	0.13971	0.28870	0.31161
lr 0.0003 & gama 0.99	0.26927	0.27850	0.31391
lr 0.0003 & gama 0.98	0.24394	0.25183	0.33386
lr 0.0003 & gama 0.97	0.26367	0.28293	0.30111

accuracy. Since we only have a normalized prediction available, we cannot use this metric

Procrustes Aligned Mean Per-joint Position Error (PA-MPJPE) is a variant of the MPJPE metric. Similar to MPJPE, it measures the average Euclidean distance between the predicted and ground-truth points, after rigid alignment using the Procrustes Analysis. The PA-MPJPE is also expressed in millimeters. Since this adjustment is done individually in multi person case, we would lose the information that our entire research is aimed at.

Procrustes Aligned Group Mean Per-joint Position Error (PAG-MPJPE)

Due to the above problems, we introduce the third version of MPJPE, where the Procrustes Analysis is implemented for the whole group of people. In other words, we consider the entire group of people as a single set of points

$$E_{MPJPE}(f, \mathcal{S}) = \frac{1}{N_S} \sum_{i=1}^{N_S} \|m_{\mathbf{f}, \mathcal{S}}^{(f)}(i) - m_{\mathbf{gt}, \mathcal{S}}^{(f)}(i)\|_2 \quad (5.1)$$

5.3 Main Results

Following MVP [17], We use the ‘160906 pizza1’, ‘160422 haggling1’, ‘160906 ian5’, and ‘160906 band3’ as our evaluation sequence for panoptic.

5.3.1 Results on CMU Panoptic Dataset

Before evaluating the model’s performance, we involve two steps: hyperparameter optimization and fine-tuning on the evaluation dataset. We identify and obtain the most suitable parameters and weights for the model using the entire dataset. After optimizing the hyperparameters, we proceeded to fine-tune the model using a designated evaluation dataset. This step involves adjusting the model’s parameters and weights based on the evaluation dataset to further refine its performance. After incorporating this fine-tuning step, our subsequent evaluation results showed an improvement of 10mm.

Before we can calculate the metrics, it is necessary to denormalize the data, as our method utilizes normalized inputs and outputs. By denormalizing the data and returning it to the original reference system, then we can know the relative pose in the real world. In this way, we can obtain a more accurate evaluation of the results. We evaluate the single pose estimation use PA-MPJPE metric on our own implementation based on MV NRSfM method, and the multi-person pose estimation use PAG-MPJPE metric on our proposed method (Seeing in Table 5.2). To obtain the PA-MPJPE metric and PAG-MPJPE, we evaluate the MVP method in the same way, we got the results(Seeing in Table 5.3). Additionally, we present some results of our approach when applied to the Panoptic dataset(Seeing in Figure 5.1).

Table 5.2: Evaluation on Panoptic (in mm), This table displays the assessment of our multi-body estimation method, as well as the evaluation of single-body estimation using our own implementation of the MV NRSfM approach.

Method	Metrics	haggling	band	ian	pizza	Mean
MV NRSfM. [1]	PA-MPJPE	48.80	35.15	38.33	44.26	41.64
MVMB NRSfM(ours)	PAG-MPJPE	78.71	128.74	132.69	86.97	106.78

Table 5.3: Evaluation on Panoptic sequences (in mm), This table demonstrates the impressive performance of our method, even with a limited number of uncalibrated cameras, when utilizing only 2D Ground Truth.

Method	PA-MPJPE	PAG-MPJPE	Use 3D GT	Use 2D GT	Use Camera parameters	Number of Views
MVP [17]	14.43	20.35	✓	✗	✓	5
MV NRSfM [1]	41.64	-	✗	✓	✗	3
MVMB NRSfM(ours)	-	106.78	✗	✓	✗	3

5.3.2 Results on In the wild Dataset

The task of the participants in the experiment was to move in the space surrounded by the cameras and from time to time turn towards other and then turn away. Our aim was to determine whether these events can be detected by our method. In order to test the limits of the method, we asked colleagues with different body types to the task. In solving the task, we first trained a model on the Panoptic database for the landmarks estimated by ViTPose, and then tuned this model to the given sequences with the keypoint also estimated by ViTPose. The results seeing in Figure 5.2.

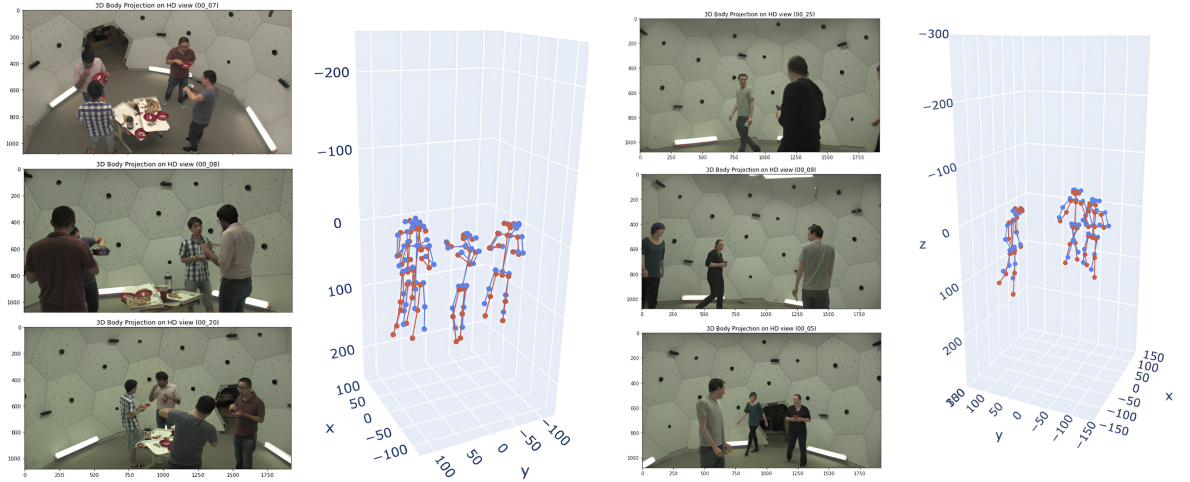


Figure 5.1: On the left, there are three randomly chosen input images taken from Panoptic’s synchronized cameras. On the right, we present the prediction and ground truth generated by our method, with the ground truth depicted in red and the prediction in blue

5.3.3 Discussion

This study, while providing valuable insights, is subject to several limitations that should be considered when interpreting the findings. One limitation of the current study is the handling of scale when dealing with two bodies with a significant difference in height and volume (For instance, in the Panoptic dataset’s sequence *ian5*, which features a video with both children and adults, our method experiences the worst error when evaluating this particular sequence). This issue arises when calculating the relative position, as we first normalize both bodies simultaneously

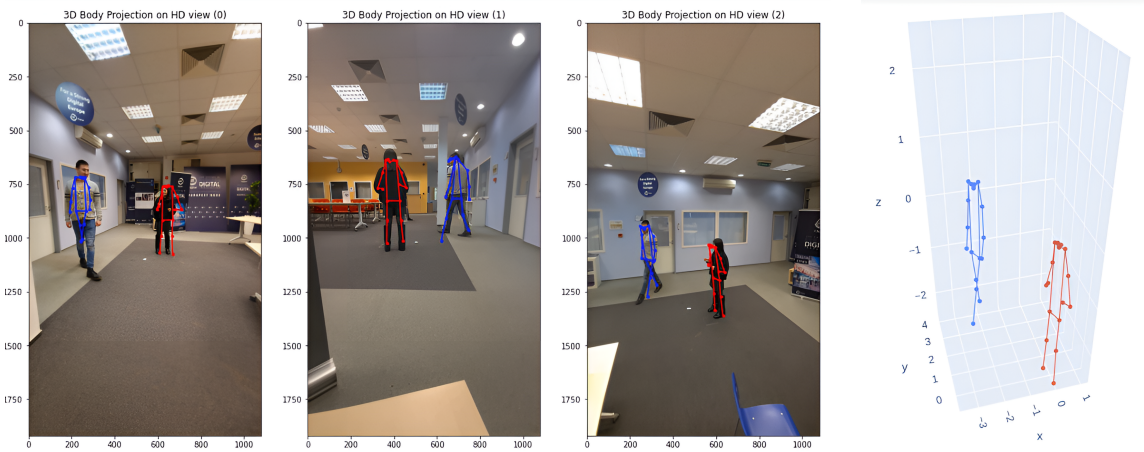


Figure 5.2: Left 3 images: The input images with the re-projected relative pose
Right: the predicted relative 3D pose.

(using the neck key point for position calculation and normalization). Afterward, we use the same scale to revert to the initial reference system when the calculation is completed, resulting in substantial errors for bodies with considerable height and volume differences.

Chapter 6

Conclusion

We have introduced a simple yet efficient novel technique utilizes the fixed relative position of the camera to compute the rotation and translation matrices, which in turn allows for the calculation of the relative pose between multiple skeletons. Unlike previous methods, our approach does not require both the camera parameters and 3D ground truth, making it a valuable contribution to the field.

While our proposed method demonstrates promising results, there are several directions for future research that can further enhance its performance and applicability. One potential direction for future work is to replace the 2D ground truth required by our method with an existing estimated 2D pose model. Numerous approaches have already demonstrated promising results in this area. By leveraging such models, our method could be adapted to function without any ground truth, thereby increasing its practical applicability and ease of use.

Acknowledgements

We would like to express our sincere gratitude to Professor András Lőrincz, Professor Levente Hajder and Gábor Baranyi for their invaluable assistance and guidance throughout the course of this project. Thanks to Szlatincsán Adrián and Xiang Linyun for helping us to record the video of our experimental. Thanks to the NIPG group, MiLAB, Bosch and HumanE-AI for providing the resources.

Bibliography

- [1] Chen Kong and Simon Lucey. “Deep non-rigid structure from motion”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1558–1567.
- [2] Mosam Dabhi et al. “High Fidelity 3D Reconstructions with Limited Physical Views”. In: *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, pp. 1301–1311.
- [3] Hanbyul Joo et al. “Panoptic studio: A massively multiview system for social motion capture”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3334–3342.
- [4] Ke Sun et al. “Deep high-resolution representation learning for human pose estimation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5693–5703.
- [5] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked hourglass networks for human pose estimation”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*. Springer. 2016, pp. 483–499.
- [6] Yufei Xu et al. “Vitpose: Simple vision transformer baselines for human pose estimation”. In: *arXiv preprint arXiv:2204.12484* (2022).
- [7] Liqiang Yin et al. “Self-Supervised Human Pose based Multi-Camera Video Synchronization”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 1739–1748.
- [8] Edoardo Remelli et al. “Lightweight multi-view 3d pose estimation through camera-disentangled representation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 6040–6049.

- [9] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. “Weakly-supervised 3d human pose learning via multi-view images in the wild”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 5243–5252.
- [10] Helge Rhodin et al. “Learning monocular 3d human pose estimation from multi-view images”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8437–8446.
- [11] Haibo Qiu et al. “Cross view fusion for 3d human pose estimation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 4342–4351.
- [12] Karim Iskakov et al. “Learnable triangulation of human pose”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 7718–7727.
- [13] Leonid Pishchulin et al. “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation”. In: (Nov. 2015).
- [14] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. “Lcr-net++: Multi-person 2d and 3d pose detection in natural images”. In: *IEEE transactions on pattern analysis and machine intelligence* 42.5 (2019), pp. 1146–1161.
- [15] Hao-Shu Fang et al. “RMPE: Regional Multi-person Pose Estimation”. In: Oct. 2017, pp. 2353–2362. DOI: 10.1109/ICCV.2017.256.
- [16] Vasileios Belagiannis et al. “3d pictorial structures revisited: Multiple human pose estimation”. In: *IEEE transactions on pattern analysis and machine intelligence* 38.10 (2015), pp. 1929–1942.
- [17] Jianfeng Zhang et al. “Direct multi-view multi-person 3d pose estimation”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 13153–13164.
- [18] He Chen et al. “Multi-person 3d pose estimation in crowded scenes based on multi-view geometry”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III* 16. Springer. 2020, pp. 541–557.

- [19] Sara Ershadi-Nasab et al. “Multiple human 3d pose estimation from multiview images”. In: *Multimedia Tools and Applications* 77 (2018), pp. 15573–15601.
- [20] Sara Ershadi-Nasab, Shohreh Kasaei, and Esmail Sanaei. “Uncalibrated multi-view multiple humans association and 3D pose estimation by adversarial learning”. In: *Multimedia Tools and Applications* 80 (2021), pp. 2461–2488.
- [21] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. “Recovering non-rigid 3D shape from image streams”. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*. Vol. 2. IEEE. 2000, pp. 690–696.
- [22] Suryansh Kumar, Yuchao Dai, and Hongdong Li. “Multi-body non-rigid structure-from-motion”. In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE. 2016, pp. 148–156.
- [23] Buzhen Huang et al. “Dynamic multi-person mesh recovery from uncalibrated multi-view cameras”. In: *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, pp. 710–720.
- [24] Carlo Tomasi and Takeo Kanade. “Shape and motion from image streams: a factorization method.” In: *Proceedings of the National Academy of Sciences* 90.21 (1993), pp. 9795–9802.
- [25] K. S. Arun, T. S. Huang, and S. D. Blostein. “Least-squares fitting of two 3-D point sets”. In: *PAMI* 9.5 (1987), pp. 698–700.
- [26] Carsten Steger. “Algorithms for the orthographic-n-point problem”. In: *Journal of Mathematical Imaging and Vision* 60.2 (2018), pp. 246–266.
- [27] Azat Akhmetyanov et al. “Sub-millisecond Video Synchronization of Multiple Android Smartphones”. In: *2021 IEEE Sensors*. IEEE. 2021, pp. 1–4.

List of Figures

1.1	The method we propose simultaneously estimates each pose (reference in blue and other in red) and the corresponding camera parameters (estimated cameras as green rectangles), then, knowing that these cameras are located in the same position, we move the individual skeletons into the same space. In this way, we get their position relative to each other	5
3.1	Pipeline of the proposed method. The method can be divided into two phases. In the first phase, three image inputs are processed through an encoder, yielding a set of 2D projections, these 2D projections are merged to estimate the 3D structures(MV NRSfM method). While in the second phase, we ascertain relative poses through the computation of rotation and translation matrices. The Procrustes algorithm is utilized to calculate relative rotation while relative translation is optimized from the minimization of the re-projection error. .	11
4.1	The left 3 images are the input, the right is the estimated Pose . . .	17
5.1	On the left, there are three randomly chosen input images taken from Panoptic’s synchronized cameras. On the right, we present the prediction and ground truth generated by our method, with the ground truth depicted in red and the prediction in blue	23
5.2	Left 3 images: The input images with the re-projected relative pose Right: the predicted relative 3D pose.	24

List of Tables

5.1	Hyperparameters Optimazation Results, from this table we can find that the best hyperparameters is learning rate is 0.0002, gamma is 0.97,and the dictionary size is 7	21
5.2	Evaluation on Panoptic (in mm), This table displays the assessment of our multi-body estimation method, as well as the evaluation of single-body estimation using our own implementation of the MV NRSFM approach.	22
5.3	Evaluation on Panoptic sequences (in mm), This table demonstrates the impressive performance of our method, even with a limited number of uncalibrated cameras, when utilizing only 2D Ground Truth. .	22