Self-Mounted Motion Capture System Using Mutual Projection of Asynchronous Cameras

Kazusa Ozaki, Fumihiko Sakaue and Jun Sato
Nagoya Institute of Technology, Nagoya, Japan

Keywords: Motion Capture Systems, Bundle Adjustment, Neural Network Representation, Asynchronous Stereo.

Abstract: In this research, we propose a method for restoring three-dimensional motion from time-series images captured by an asynchronous camera in order to realise a motion capture system using a camera attached to the body surface. For this purpose, we represent the motion trajectory of each marker using a neural network, and estimate the motion trajectory by optimising the neural network from the input images. It is also shown that stable 3D restoration can be achieved using a method called mutual projection, assuming that the cameras are reflecting each other. We show that it is possible to estimate 3D motion from asynchronous cameras with high accuracy.

1 INTRODUCTION

In recent years, motion capture technology (Moeslund et al., 2006), which measures and analyses human movements into numerical values, has been used in a wide range of fields. However, the current mainstream motion capture method, the optical motion capture method(Guerra-Filho, 2005), has high measurement accuracy, but the system tends to be large because it requires cameras to be set up around the area to be captured.

Other than optical motion capture systems, there are mechanical motion capture systems that use acceleration sensors or angular acceleration sensors for measurement(Roetenberg et al., 2009). Mechanical systems use sensors for measurement and can capture in various environments without the need to install equipment such as cameras and markers, but have lower accuracy than optical systems. In addition, the sensors used for capturing are affected by magnetic fields, making them unsuitable for locations with unstable magnetic fields.

In a previous study of motion capture in small-scale systems, a method for acquiring motion information by restoring the motion of a camera attached to the human body using structure from motion technology has been proposed(Shiratori et al., 2011). This method is relatively inaccurate compared to the actual measurement method, as it does not directly capture the markers, etc. There is also a method for capturing motion by attaching an omni-directional camera or other camera capable of capturing the target person(Miura and Sako, 2020). Although this method directly captures the object, there are many areas that are hidden by occlusion, resulting in low measurement accuracy.

As described above, methods for motion capture without using an external camera have been proposed, but each method has problems in terms of accuracy and stability. In this study, we propose a method to solve these problems by attaching a camera and a marker to the object and using the camera as a marker and a filming device. In this method, the camera attached to the body is regarded as a marker, while other markers and cameras are photographed and their 3D positions are recovered for motion measurement. In this case, a very strong geometric constraint called mutual projection can be used, which enables stable estimation and restoration of the camera position.

However, when using a stereo camera system such as the one proposed in this system to perform restoration, the cameras need to take pictures synchronously, as the corresponding points taken at the same time are required. However, as described above, the cameras need to be connected to each other for synchronous shooting, which makes it unsuitable for a method in which the cameras are attached to the human body. For this reason, this study presents a method that can stably realise 3D restoration even from asynchronous cameras. This method focuses on the motion trajectory of each marker and restores parameters related to the trajectory, enabling stable stereo restoration even

Ozaki, K., Sakaue, F. and Sato, J.
Self-Mounted Motion Capture System Using Mutual Projection of Asynchronous Cameras.
DOI: 10.5220/0012473500003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 4: VISAPP*, pages 147-153
ISBN: 978-989-758-679-8; ISSN: 2184-4321

from images captured by an asynchronous camera. In this way, a method for stable motion restoration from images taken by a group of cameras attached to the human body is presented.

2 EPIPOLAR GEOMETRY AND STEREO RECONSTRUCTION

In this section, we will explain the camera model used in this research and a method for restoring 3D shapes based on image information obtained from the camera based on epipolar geometry.

2.1 Epipolar Geometry

First, 3D restoration using the stereo camera system used in this study is described. In this study, the 3D information is recovered from the image information obtained from the cameras. For this purpose, the epipolar geometry(Hartley and Zisserman, 2003) is used, which can represent multiple cameras. Let us consider the case where two cameras capture a point $\mathbf{X} = [X\ Y\ Z]^\top$ in 3D space and the points $\mathbf{m} = [u\ v]^\top$ and $\mathbf{m}' = [u'\ v']^\top$ on the image are obtained. The following relation holds between these two points using the fundamental matrix $\mathbf{F}$ that represents the relative relationship between cameras.

$$\widetilde{\mathbf{m}}'^\top \mathbf{F} \widetilde{\mathbf{m}} = 0 \tag{1}$$

where $\widetilde{\mathbf{m}}', \widetilde{\mathbf{m}}$ is the homogeneous representation of $\mathbf{m}', \mathbf{m}$ and $\tilde{\mathbf{m}} = [\mathbf{m}^T\ 1]^T$. Also, $\mathbf{F}$ is the fundamental matrix, which represents the relative relationship between the two cameras. Images taken by two cameras, the corresponding points will always satisfy this equation. In addition, $\mathbf{F}$ contains information about the position and orientation between the cameras. Therefore, this epipolar geometry can be used to determine the relative attitude information between the cameras.

2.2 3D Reconstruction by Stereo Camera System

Next, the camera projection matrix $\mathbf{P}$ is obtained from the $\mathbf{F}$ matrix, and the 3D reconstruction is performed using this matrix. The following relationship is established between the camera matrix $\mathbf{P}$ and the points $\mathbf{X} = [X\ Y\ Z]^\top$ in 3D space and $\mathbf{m} = [u\ v]^\top$ on the image.

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2}$$

where $p_{11} \sim p_{34}$ are components of the camera projection matrix. where p_{11} to p_{34} are the elements of the camera matrix. Eliminating λ from the equation (2) and summarising for $[X\ Y\ Z]^\top$, we obtain the following.

$$\begin{bmatrix} p_{31}u - p_{11} & p_{32}u - p_{12} & p_{33}u - p_{13} \\ p_{31}v - p_{21} & p_{32}v - p_{22} & p_{33}v - p_{23} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} p_{14} - p_{34}u \\ p_{24} - p_{34}v \end{bmatrix} \tag{3}$$

This equation shows that two constraint equations for $\mathbf{X}$ can be obtained from one camera if the camera matrix and the point $\mathbf{m}$ on the image are known. Thus, a 3D point $\mathbf{X}$ can be recovered if the corresponding points taken by two or more cameras are available.

2.3 Bundle Adjustment

In 3D shape reconstruction with actual camera images, it is often not possible to find a suitable solution due to various noise effects. This is because not only the image points used for reconstruction, but also the parameters in the camera matrix are strongly influenced by noise. Therefore, in many cases, a method called bundle adjustment is used to optimise the camera parameters and the 3D restoration result to obtain a more accurate estimation of the restoration result. Bundle adjustment(Triggs et al., 2000) is a method for optimising multiple parameters in a batch in order to improve the estimation accuracy. This optimisation is achieved by estimating the camera matrix $\mathbf{P}$ and the three-dimensional point $\mathbf{X}$ in such a way that the error in the reprojection of the estimated 3D shape onto the image plane, i.e. the reprojection error, is minimised, as described above in the following formula.

$$E = \frac{1}{2} \sum_i \sum_j \{(u_i^j - \bar{u}(\mathbf{P}_j, \mathbf{X}_i))^2 + (v_i^j - \bar{v}(\mathbf{P}_j, \mathbf{X}_i))^2\} \tag{4}$$

where u_i^j, v_i^j are the coordinates of the observation point obtained by imaging the three-dimensional point $\mathbf{X}_i$ with camera j, $\bar{u}(\mathbf{P}_j, \mathbf{X}_i)$, $\bar{v}(\mathbf{P}_j, \mathbf{X}_i)$ are the coordinates of the three-dimensional point $\mathbf{X}_i$ projected onto the image plane with coordinates obtained by projecting the 3D point $\mathbf{X}_i$ onto the image plane by the camera $\mathbf{P}_j$. As this reprojection error is a nonlinear function, some reasonable initial values are required for its minimisation. For this reason, in general 3D reconstruction methods, an initial estimate of $\mathbf{P}$ is made from the $\mathbf{F}$ matrix obtained based on epipolar geometry, and bundle adjustment is carried out using this as the initial value.

3 MUTUAL CAMERA PROJECTION IN EPIPOLAR GEOMETRY

This section describes a stereo reconstruction method using mutual projection.

3.1 Mutual Camera Projection

By using the bundle adjustment presented in the previous section, the position and orientation of the camera and the 3D shape can be estimated simultaneously. However, in this research, both the camera and the marker are mounted on the human body for measurement, so the camera position varies significantly compared to normal scenes. Therefore, an important issue in this research is to stabilise the camera position. In order to solve this problem, we utilise the mutual projection system (ITO and SATO, 2002).

The self-attached motion capture system proposed in this study requires a camera mounted on the body surface to capture other markers. Therefore, wide-angle cameras that can capture a very large area are used. As described above, not only the markers whose positions are to be measured, but also the camera for taking the images will be incorporated in the images taken by each camera. Considering that the epipole in the epipolar geometry coincides with the point at which the optical centre of the camera is captured, it can be seen that in this situation the epipole can be directly obtained from the observed image. Since the epipoles contain the position information of the cameras, the relative positions of the cameras can be determined very stably by obtaining them directly.

When two cameras are projected onto each other's image plane as epipoles $\mathbf{e}$,$\mathbf{e}'$ as in Fig.1, the following relationship is established between the basis matrix $\mathbf{F}$ and the epipoles $\mathbf{e}$,$\mathbf{e}'$ $\mathbf{e}$ and $\mathbf{e}'$.

$$\mathbf{F}\widetilde{\mathbf{e}} = 0 \tag{5}$$

$$\mathbf{F}^{\top}\widetilde{\mathbf{e}}' = 0 \tag{6}$$

where $\widetilde{\mathbf{e}}$,$\widetilde{\mathbf{e}}'$ are homogeneous representations of $\mathbf{e}$,$\mathbf{e}'$. Since this epipole places a strong constraint on the basis matrix $\mathbf{F}$, the relative attitude information between the two cameras can be obtained stably by calculating $\mathbf{F}$ using the directly observed epipole.

3.2 Bundle Adjustment Using Mutual Projection

The mutual projection in the estimation of the $\mathbf{F}$ matrix described above allows a stable estimation of the camera parameters. This is optimised by bundle adjustment to perform 3D reconstruction. In this case, the strong constraints obtained by the mutual projection can also be used in the bundle adjustment. As mentioned above, in mutual projection the camera position is captured directly. Therefore, when performing bundle adjustment, the reprojection error can be calculated for the estimated camera position in addition to the reprojection error of the 3D points. Considering this, bundle adjustment using mutual projection can be defined as minimising the reprojection error of the reconstructed points as well as the reprojection error of the epipoles by the following equation.

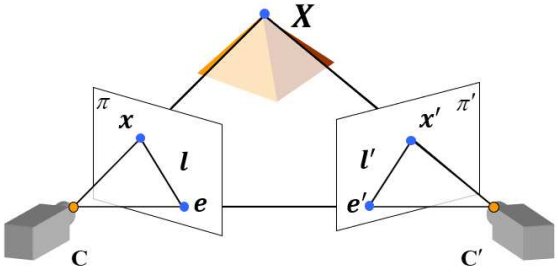


Figure 1: Epipolar geometry and epipole.Points e and e' are epipoles of cameras c and c'.

$$E'(\mathbf{P},\mathbf{X}) = E + \frac{1}{2}\sum_{j}\sum_{k\neq j}(\mathbf{e}_j - \bar{\mathbf{e}}(\mathbf{P}_k,\mathbf{T}_j))^2 \tag{7}$$

where $\mathbf{e}$ is the epipole observed by the camera and $\bar{\mathbf{e}}(\mathbf{P},\mathbf{T})$ are the coordinates of the estimated 3D position of the camera $\mathbf{T}$ projected onto the image plane.

The reprojection error calculated in this way can directly optimise the information on the camera position. Therefore, the estimation of the camera position is more accurate than when estimating the camera position only from the relation of the corresponding points.

4 STEREO RECONSTRUCTION USING ASYNCHRONOUS CAMERAS

This section describes a method for stereo reconstruction using asynchronous cameras.

4.1 3D Trajectory Reconstruction Based on Parameter Representation of 3D Trajectories

At last, 3D reconstruction using asynchronous cameras is described. All the methods described above assume that multiple cameras are capturing the same

scene, i.e. that they are acquiring information at the same time. However, multiple cameras running independently often capture images asynchronously. Therefore, such an assumption is no longer valid when synchronous camera systems are not used. This makes proper 3D reconstruction difficult when using the epipolar geometry that is common in asynchronous cameras.

In order to solve this problem, a method has been proposed to transform the trajectory of the corresponding points into frequency space and restore them as points in frequency space(Kakumu et al., 2013). This method focuses on the trajectory as a whole, rather than on each 3D point, and estimates the frequency components that represent the trajectory. This enables 3D reconstruction with an asynchronous camera. However, in this method, the reconstruction is carried out using an affine camera model so that the 2D projected points and the 3D points can be represented in a linear relationship. This makes it difficult to apply when the cameras are located very close to each other, as is the case in this study.

Here, viewing the frequency components recovered by this method as parameters for parametrically constructing the 3D trajectory, the reconstruction of the 3D trajectory can be considered as the estimation of parameters for constructing the trajectory. In this case, as long as the necessary constraints for estimating the parameters are obtained, 3D reconstruction can be achieved appropriately even when images taken at the same time are not available. In this study, 3D reconstruction from an asynchronous camera is performed using such a parameter representation of the trajectory.

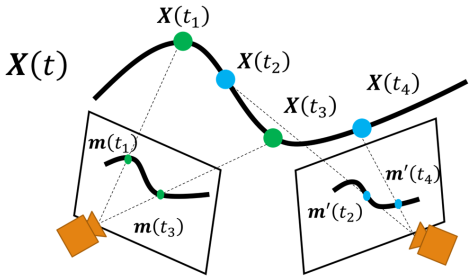


Figure 2: Example of the 3D trajectory and projected 3D points.X is a 3D point, x is a projection point, and t is time.

4.2 Representation of 3D Trajectories Using Neural Networks

A typical representation of the parametric representation of a 3D trajectory is the interpolation method using spline interpolation, etc. In the method, a 3D trajectory can be constructed from multiple basis points. Therefore, the estimation of the 3D trajectory in this method is equivalent to the estimation of the basis points. However, when using such an interpolation method, the 3D trajectory that can be represented by the chosen interpolation method is limited. In addition, if the corresponding points cannot be observed due to occlusion or other reasons, appropriate estimation will not be possible.

Therefore, this study adopts the representation of trajectories using neural networks. This method focuses on the fact that neural networks are general-purpose functions that can represent various functions, and uses this functional representation to represent trajectories. In other words, when a certain time t is input, the neural network is trained as a function that outputs a 3D point at that time. This learning is achieved by minimising the reprojection error at each camera and each time, defined as follows.

$$E'' = \frac{1}{2}\sum_{t}\sum_{i}\sum_{j}[\{(u_i^{j,t} - \bar{u}(\mathbf{P}_j^t, \mathbf{X}_i^t))^2 + (v_i^{j,t} - \bar{v}(\mathbf{P}_j^t, \mathbf{X}_i^t))^2\} + \sum_{k\neq j}\{(\mathbf{e}_j^t - \bar{\mathbf{e}}(\mathbf{P}_k^t, \mathbf{T}_j^t))^2\}] \tag{8}$$

where $\mathbf{X}_i^t$ is the 3D point obtained when time t is input to the neural network. Also, $\bar{u}$ and $\bar{v}$ are the projected points obtained by projecting the 3D point by the camera matrix. By minimizing the loss function, we can obtain a neural network that represents the 3D trajectory of the observed points taken by asynchronous cameras.

Note that when using a neural network to represent an arbitrary function, it is known that if variables such as time are input directly, it becomes difficult to represent high-frequency components. To avoid this, it is necessary to map these variables to a higher-order space in advance using positional encoding. This method is also used in this study, and t is input to the neural network after being mapped to a higher-order space. In addition, appropriate initial values are required for this non-linear minimisation. For this reason, in this study, 2D points are interpolated in advance to create a set of pseudo-synchronised corresponding points. The interpolated values are used for synchronous bundle adjustment. The results obtained are optimised using the method described above to estimate the final reconstruction result.

Furthermore, the parameter representation using such a neural network is applicable not only to the 3D points to be restored, but also to all parameters including the camera position. Therefore, in this research, the same representation is used for these parameters, and the camera position and 3D trajectory are estimated by minimising the reprojection error shown by

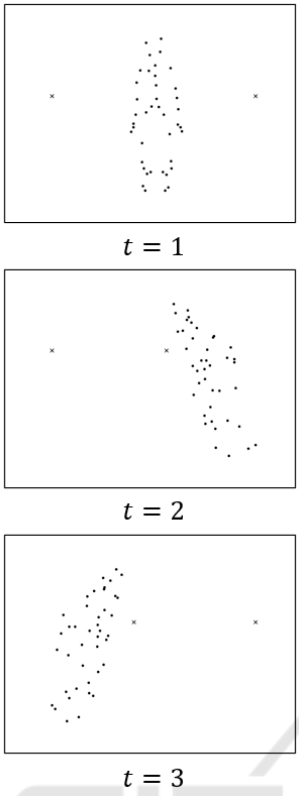


Figure 3: Examples of input images.The cross mark is an epipole.

the equation (8). This enables stable estimation of the 3D trajectory even from asynchronous cameras.

5 EXPERIMENTAL RESULTS

In this section, we present the results of simulation-based restoration using the proposed method.

5.1 Environment

The results of 3D reconstruction from images captured by an asynchronous camera using the proposed method are presented. In this experiment, the motion of doing a standing long jump was selected from the CMU Graphics Lab Moiton Capture Database(Carnegie Mellon University, 2003), and the image taken by the virtual camera was used. The scenes were taken by three cameras at different positions at different times, as shown in Fig.4. The red dots in the figure represent the 3D point cloud and the surrounding rectangles represent the cameras. In order to reproduce the asynchronous situation where the cameras were not synchronised, the 3D point cloud was taken at a time when each camera was off by

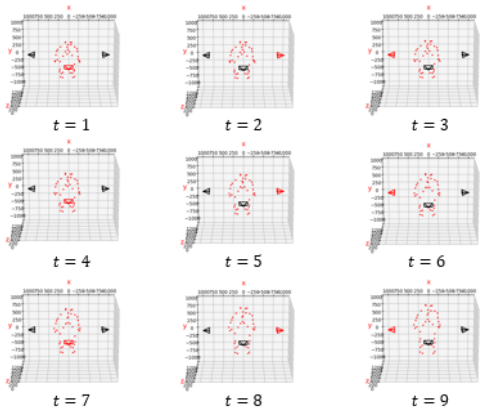


Figure 4: Target 3D points.The red dots in the figure represent the 3D point cloud and the surrounding rectangles represent the cameras.The red camera is the camera that is projecting at that time.

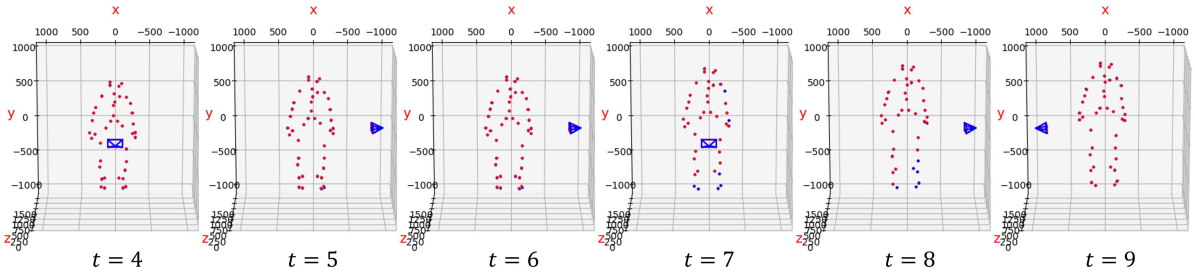
three frames. The camera indicated in red in Fig.4 is the camera that is projecting at that time.

The images used for the actual reconstruction are shown in Fig3. The images show the first image taken by each camera. The black dots are the projected points of the 3D point cloud and the X marks the epipoles of the cameras. From these images, the 3D reconstruction was carried out using the proposed method. For comparison, the following methods were used for the interpolation of asynchronous cameras, with and without mutual projection restoration was performed using neural networks, linear interpolation and cubic spline interpolation, respectively.

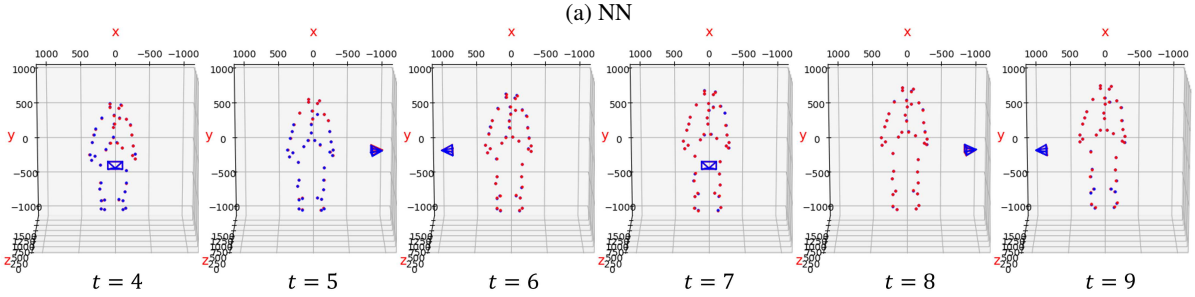
We experimented with a NN structure consisting of only one fully connected layer with 256 units, with the input being 4-dimensional by positional encoding and the 3-dimensional point X being the output. In addition, the results of the neural network are optimized using the equation (8) after initial learning of the network using the spline results.

5.2 Results

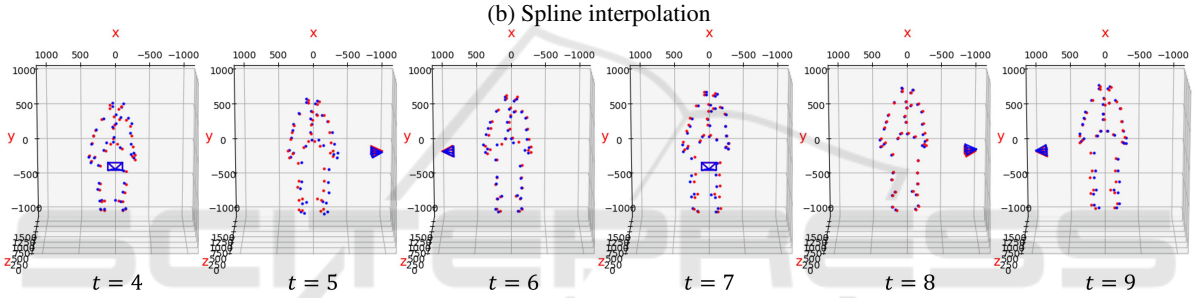
The results of the restoration using each method are shown in Fig.5. In all of these results, the restoration is carried out using mutual projection. In the result images, the true values are shown in red and the recovered results in blue. In addition, Table 1 shows the restoration error (RMSE) with and without mutual projection and when the restoration is carried out using each interpolation method. These results show that the 3D points reconstructed using the neural network are more accurate than those reconstructed using linear interpolation or spline interpolation. It can also be seen that the results of reconstruction using mutual projection are more accurate for all interpola-



(a) NN



(b) Spline interpolation



(c) Linear Interpolation

Figure 5: Examples of 3D recunstruction.In all of these results, the restoration is carried out using mutual projection. The true values are shown in red and the recovered results in blue.

Table 1: Reconstruction errors (RMSE) by each method (mm).

	w/ epipoles	w/o epipoles
Linear interpolation	22.661	24.366
Cubic spline polation	6.881	6.970
NN	5.572	5.722

Table 2: Reconstruction errors (RMSE) by each method from 10 movoments (mm).

	w/ epipoles	w/o epipoles
Linear interpolation	19.921	21.057
Spline interpolation	4.649	4.940
NN	4.313	4.596

tion methods. These results confirm that the use of mutual projection enables high-precision 3D restoration even with an asynchronous camera.

The results of calculating the average restoration error from 10 movements in the dataset are shown in Table 2. These results confirm that the proposed method can achieve highly accurate restoration.

6 CONCLUSION

In this study, a self-attached motion capture system using mutual projection in an asynchronous camera is proposed as a method to realise small-scale motion capture without location constraints. To this end, a method for stable 3D restoration even with asynchronous cameras using mutual projection is presented.

REFERENCES

Carnegie Mellon University (2003). CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/.

Guerra-Filho, G. (2005). Optical motion capture: Theory and implementation. *RITA*, 12(2):61–90.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press.

ITO, M. and SATO, J. (2002). Robust computation of epipolar geometry from mutual projection of cameras. *IEICE transactions on information and systems*, 85(3):600.

Kakumu, Y., Sakaue, F., Sato, J., Ishimaru, K., and Imanishi, M. (2013). High frequency 3d reconstruction from unsynchronized multiple cameras. In *BMVC*.

Miura, T. and Sako, S. (2020). 3d human pose estimation model using location-maps for distorted and disconnected images by a wearable omnidirectional camera. *IPSJ Transactions on Computer Vision and Applications*, 12:1–17.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126.

Roetenberg, D., Luinge, H., Slycke, P., et al. (2009). Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV, Tech. Rep*, 1:1–7.

Shiratori, T., Park, H. S., Sigal, L., Sheikh, Y., and Hodgins, J. K. (2011). Motion capture from body-mounted cameras. In *ACM SIGGRAPH 2011 papers*, pages 1–10.

Triggs, B., McLauchlan, P. F., Hartley, R. I., and Fitzgibbon, A. W. (2000). Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, ICCV '99, pages 298–372, London, UK, UK. Springer-Verlag.