

Human Pose Estimation from Egocentric Videos

Shunya Egashira

Graduate School of Engineering, Kyushu Institute of Technology, 1-1 Sensuicho, Tobata, Kitakyushu, 804-8550, Japan

Yui Tanjo¹

¹Faculty of Engineering, Kyushu Institute of Technology, 1-1 Sensuicho, Tobata, Kitakyushu, 804-8550, Japan
Email: Egashira.shunnya602@mail.kyutech.jp, tanjo@cntl.kyutech.ac.jp

Abstract

According to a survey conducted by the Ministry of Health, Labour and Welfare in 2019, about 30% of patients suffer from back pain and stiff shoulders. Although researches on pose estimation have been conducted for a long time, they cannot be used for daily pose estimation, because they need fixed cameras to capture target/subject motion. To solve this problem, the present paper, proposes a novel pose estimation method from egocentric videos using Epipolar Geometry. It computes three rotational angles, i.e., pitch, yaw and roll, from the egocentric motion videos to evaluate differences from his/her normal motion. In the experiment, three egocentric videos were used to verify the performance and effectiveness of the proposed method and reasonable/satisfactory results were obtained.

Keywords: Egocentric, Posture Estimation, Epipolar Geometry, Rotation Angles

1 Introduction

According to the Ministry of Health, Labor and Welfare's 2009 National Survey of Basic Living Conditions [1], the prevalence of back pain and shoulder stiffness is about 30%. [2] [3] Since these problems interfere with daily life of a person, it is important to consciously correct his/her posture in daily life. Therefore, posture estimation, which allows one to check one's posture status, has been used in the research on posture estimation in animation motion production and virtual reality games. One of the posture estimation techniques is motion capture. Motion capture is a method of creating a 3D model by placing a monocular camera[4] or multiple cameras[5][6][7] around a person to be photographed and capturing the subject's movements. Methods for detecting the position of the subject's joints include attaching markers to the subject [8] and deep learning estimation [9].

However, these methods have disadvantages in that they require time for environmental preparation such as equipment and can only be performed in a limited area. There have also been studies on posture estimation using deep learning[10], gait posture estimation using video from a third viewpoint using a fixed camera[11], and gait posture estimation using epipolar geometry[12] as previous researches. However, the problems are that it is difficult to determine the cause of the problem because of the use of deep learning, and it is also impossible to estimate the posture of a human daily life because it can



Fig.1. The chest-mounted camera

only take pictures in a fixed environment. Moreover, it is difficult to estimate a posture with small movement, because it is normally used to estimate walking posture.

Based on these backgrounds, this paper proposes a method to acquire images of the surrounding landscape from a chest-mounted camera(Fig.1) and estimate the angle using epipolar geometry using the feature points obtained from the images.

2 Methodology

2.1. Camera Calibration

The camera contains lens distortion, which affects the accuracy of the feature point correspondence and makes it impossible to correctly map between the two images. In order to eliminate the distortion, the camera is calibrated. The camera parameters include internal parameters, external parameters, and distortion coefficients. A checkerboard is used to estimate the camera parameters.

2.2. Feature Point Detection and Matching

A-KAZE is a robust feature point extraction algorithm invariant to scaling, rotation, and lighting change. Using A-KAZE, the feature points are extracted and described, and they are matched between two image frames.

While the Lucas-Kanade (LK) tracker is commonly used in target tracking, its use in feature matching is comparatively rare. After detecting feature points in one image, the corresponding matching region in the other image must be identified to establish feature point pairs. (Fig.2) The LK tracker is subsequently employed to locate this matching region.

The removal of the outliers is performed by applying RANSAC to hypothetically matched point pairs.



Fig.2. Feature point matching: (a) frame1 (b) frame3

2.3. Modeling with homographic matrices

A homography matrix is a matrix that represents a projective matrix between two images, as well as a constraint between the corresponding points.

If the x and y coordinates of the point in the image of the previous frame and those of the corresponding point in the image of the next frame are denoted by \mathbf{x}_1 , \mathbf{y}_1 , \mathbf{x}_2 , \mathbf{y}_2 , respectively, the relationship is calculated using a homographic matrix.

$$\begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \quad (1)$$

where $h_{11} \dots h_{33}$ are the elements of the homography matrix.

When estimating the homography matrix using the corresponding points obtained from the feature point matching described above, RANSAC is used to remove the outliers.

2.4. Removing Outliers with RANSAC

The RANSAC algorithm is shown below.

1. Randomly extract a certain number of data from the entire data and perform model estimation.
2. Count and memorize outlier data.
3. Repeat steps 1-2 multiple times.
4. Determine the model with the least outliers as the best model.

In the final projection transformation model, the feature points pairs that correspond to the outlier are removed. If the two images are correlated, the position and the pose of the camera is estimated using epipolar geometry.

2.5. Estimation of the fundamental matrix

All correspondences between the two images satisfy Eq. (2).

$$\tilde{\mathbf{m}}_{t+1,i}^T \mathbf{F} \tilde{\mathbf{m}}_{t,i} = 0 \quad (2)$$

$\tilde{\mathbf{m}}_{t,i}$ and $\tilde{\mathbf{m}}_{t+1,i}$ are the uniform coordinates of the image coordinates of the i th camera at time t and $t+1$, respectively, and \mathbf{F} is the fundamental matrix. They are expressed as shown in Eq. (3).

$$\mathbf{F} = \begin{pmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{pmatrix}, \quad \tilde{\mathbf{m}}_{t,i} = \begin{pmatrix} u_{t,i} \\ v_{t,i} \\ 1 \end{pmatrix}, \quad \tilde{\mathbf{m}}_{t+1,i} = \begin{pmatrix} u_{t+1,i} \\ v_{t+1,i} \\ 1 \end{pmatrix} \quad (3)$$

In this study, we use the 8-point algorithm [13] to find the fundamental matrix \mathbf{F} . Equations (2) and (3) are expanded to $f_{11} \dots f_{33}$, which is summarized as Eq. (4)

$$\begin{pmatrix} u_{t,i}u_{t+1,i} & v_{t,i}u_{t+1,i} & u_{t+1,i} & u_{t,i}v_{t+1,i} & v_{t,i}v_{t+1,i} & v_{t+1,i} & u_{t,i} & v_{t,i} & 1 \end{pmatrix} \begin{pmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{21} \\ f_{22} \\ f_{23} \\ f_{31} \\ f_{32} \\ f_{33} \end{pmatrix} = 0 \quad (4)$$

Equation (4) has 8 unknown parameters due to the indeterminate scale, and if there are 8 or more sets of corresponding points, the basic matrix \mathbf{F} can be estimated. Therefore, Eq. (4) is applied to the corresponding points of N set ($N \geq 8$) and is expressed as Eq. (5).

$$\mathbf{M} \mathbf{f} = 0 \quad (5)$$

$$\mathbf{M} = \begin{pmatrix} u_{t,1}u_{t+1,1} & v_{t,1}u_{t+1,1} & u_{t+1,1} & u_{t,1}v_{t+1,1} & v_{t,1}v_{t+1,1} & v_{t+1,1} & u_{t,1} & v_{t,1} & 1 \\ u_{t,2}u_{t+1,2} & v_{t,2}u_{t+1,2} & u_{t+1,2} & u_{t,2}v_{t+1,2} & v_{t,2}v_{t+1,2} & v_{t+1,2} & u_{t,2} & v_{t,2} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_{t,N}u_{t+1,N} & v_{t,N}u_{t+1,N} & u_{t+1,N} & u_{t,N}v_{t+1,N} & v_{t,N}v_{t+1,N} & v_{t+1,N} & u_{t,N} & v_{t,N} & 1 \end{pmatrix} \quad (6)$$

$$\mathbf{f} = (f_{11} \quad f_{12} \quad f_{13} \quad f_{21} \quad f_{22} \quad f_{23} \quad f_{31} \quad f_{32} \quad f_{33})^T \quad (7)$$

Since the right-hand side of Eq. (5) is 0, it is considered to have invariance with respect to scale. To do this, the Euclidean norm of \mathbf{f} is normalized to 1 and the fundamental matrix \mathbf{F} is estimated. However, since the condition that the rank of the fundamental matrix is 2 is not satisfied, the singular value decomposition of the fundamental matrix is performed, so that its rank becomes 2.

First, the obtained fundamental matrix is decomposed using the following equation.

$$\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (8)$$

Here \mathbf{U} and \mathbf{V} are orthogonal matrices, and $\mathbf{\Sigma}$ is an accusative matrix such as Eq. (9), in which the singular values satisfy $\sigma_1 > \sigma_2 > \sigma_3$.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix} \quad (9)$$

Of the three diagonal terms, the smallest singular value is 0, and \mathbf{F} obtained by recalculating equation (8) is the final fundamental matrix.

2.6. Estimating the Essential Matrix

The essential matrix is a matrix containing information about the motion of the camera between two images, and it is expressed by Eq. (10) using the fundamental matrix \mathbf{F} and the internal parameter \mathbf{K} obtained in Section 2.1.

$$\mathbf{E} = \mathbf{K}^T \mathbf{F} \mathbf{K} \quad (10)$$

The essential matrix \mathbf{E} is expressed by the following equation using the rotation matrix \mathbf{R} of the camera and the translation vector \mathbf{t} .

$$\mathbf{E} = [\mathbf{t}]_X \mathbf{R} \quad (11)$$

where $[\mathbf{t}]_X$ is the skewed symmetric matrix of \mathbf{t} defined by

$$[\mathbf{t}]_X = \begin{pmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{pmatrix} \quad (12)$$

2.7. Camera position and pose estimation

The essential matrix is decomposed into a rotation matrix \mathbf{R} and a translation vector \mathbf{t} as shown above.

The rotation matrix \mathbf{R} and the translation vector \mathbf{t} are expressed as follows.

$$\mathbf{R} = \begin{cases} \mathbf{U}\mathbf{W}\mathbf{V}^T \\ \mathbf{U}\mathbf{W}^T\mathbf{V}^T \end{cases} \quad (13)$$

$$[\mathbf{t}]_X = \begin{cases} \mathbf{U}\mathbf{\Sigma}\mathbf{W}\mathbf{U}^T \\ \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T\mathbf{U}^T \end{cases} \quad (14)$$

where

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (15)$$

From Eq. (13) and Eq. (14), each of \mathbf{R} and $[\mathbf{t}]_X$ has two sets of solutions. It is therefore necessary to determine the correct combination of \mathbf{R} and $[\mathbf{t}]_X$ from among them. For this purpose, three-dimensional reconstruction of the corresponding points is performed using those \mathbf{R} and $[\mathbf{t}]_X$, and a combination of \mathbf{R} and $[\mathbf{t}]_X$ is chosen which

makes the restored points all positive in front of the camera.

2.8. Estimating Angles

The rotation matrix \mathbf{R} obtained in the previous section is decomposed into three matrices using Eq.(16).

$$\mathbf{R} = \mathbf{R}_{roll} \mathbf{R}_{pitch} \mathbf{R}_{yaw} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \quad (16)$$

$$\mathbf{R}_{roll} = \begin{bmatrix} \cos \theta_{roll} & -\sin \theta_{roll} & 0 \\ \sin \theta_{roll} & \cos \theta_{roll} & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (17)$$

$$\mathbf{R}_{yaw} = \begin{bmatrix} \cos \theta_{yaw} & 0 & \sin \theta_{yaw} \\ 0 & 1 & 0 \\ -\sin \theta_{yaw} & 0 & \cos \theta_{yaw} \end{bmatrix} \quad (18)$$

$$\mathbf{R}_{pitch} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta_{pitch} & -\sin \theta_{pitch} \\ 0 & \sin \theta_{pitch} & \cos \theta_{pitch} \end{bmatrix} \quad (19)$$

Here, roll, pitch, and yaw are the axes shown in Fig. 3, and θ_{roll} , θ_{pitch} , and θ_{yaw} are the angles around respective axes.

Using Eqs. (16), (17), (18) and (19), the three angles are computed by

$$\theta_{roll} = \arctan \frac{-r_{12}}{r_{22}}$$

$$\theta_{pitch} = \arcsin (r_{32})$$

$$\theta_{yaw} = \arctan \frac{-r_{31}}{r_{33}}$$

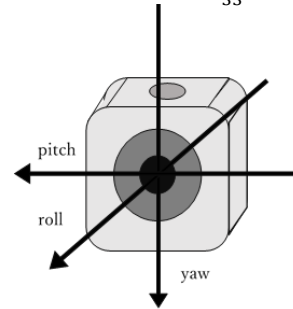


Fig.3. Camera Coordinate System

3 Experimental Results

In the experiment, we estimated walking posture of three types of behavior patterns of four subjects. The input image is obtained by cutting out the video at 3-frame intervals. We find the changes in each axis of the camera coordinate system. Assuming that the attitude angle of the initial state is 0 [deg], the rotation angle of the camera relative to the initial state of the camera is estimated by integrating it. The output obtained from the 9-axis sensor is integrated by the RTQF algorithm, and

the value is evaluated comparing with the true value. The following RMSE (Root Mean Squared Error) is used for the evaluation.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - X_i)^2}$$

where N is the number of data, and x_i and X_i represent the estimated and true values with $i=1,2,\dots, N$, respectively.

Since the value of the 9-axis sensor is an absolute angle, the initial value of each axis stopped by the sensor is 0 [deg], which is a relative angle.

The experimental results are shown in Table 1, which are the average evaluation results for three types of behavior patterns of four subjects.

Table 1. Indicators of a person's posture by movement

	Types of behavior	Pitch [deg]	Yaw [deg]	Roll [deg]
Average	Sit	3.74	3.98	1.79
	Bend to the side	4.47	8.29	8.36
	Twisting	2.26	11.16	5.77

4 Conclusion

In this paper, we proposed a method of estimating self-posture of a person using MY VISION which employs an ego-camera. In the proposed method, two consecutive images were acquired at 3-frame intervals, and, from these images, the posture of the camera wearer was estimated using epipolar geometry. As the result of the experiment, the average RMSE was 4.74 [deg], and the effectiveness of the proposed method was verified.

5 References

1. Ministry of Health, Labour and Welfare: Overview of the 2019 Basic Survey on National Living Conditions, p.19, 2019.
2. Heisei Medical Association: Posture and Mental Health, 2020.
3. Health Net Sunk. "The way you sit, and the way you walk will extend your healthy life expectancy", 2017.
4. Dushyant. D, Sridhar. S, Sotnychenko. O, Rhodin. H, Shafiei. M, Seidel. H. P, Xu. W, Casas. D, Theobalt. C.: "VNect: real-time 3d human pose estimation with a single rgb camera", ACM Transactions on Graphics, Vol.36, Issue 4, Article No. 44, 2017.
5. S. Ishikawa., J. K. Tan., Kim H.s: "3-D recovery of a non-rigid object from a single camera view employing multiple coordinates representation ", Proceedings 2013 2nd IAPR Asian Conf. on Pattern Recognition: Recent Advances in Computer Vision and Pattern Recognition (RACVPR) , pp. 946–950,2013

6. J. K. Tan, S. Ishikawa.: "Deformable shape recovery by factorization based on a spatiotemporal measurement matrix", Computer Vision and Image Understanding, Vol.82, No.2, pp.101-109, 2001.
7. T. Ohashi, Y. Ikegami, K. Yamamoto, W. Takano, Y. Nakamura: "Video motion capture from the part confidence maps of multi-camera", Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.4226-4231, 2018.
8. M. Loper, N. Mahmood, J. M. Black: "MoSh: motion and shape capture from sparse markers", ACM Transactions on Graphics (TOG), Article No. 220, pp1-13, 2014.
9. Z. Cao, T. Simon, E. S. Wei, Y. Sheikh: "Realtime multi-person 2d pose estimation using part affinity fields", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.7291-7299, 2017.
10. J. K. Tan, T. Kurosaki: "Estimation of self-posture of a pedestrian using MY VISION based on depth and motion network", Journal of Robotics, Networking and Artificial Life, Vol.7, No.3, pp.152-155, 2020.
11. M. Ooba., Y. Tanjo.: "A method of recognizing human walk motion from multiple directions", International Journal of Innovative Computing, Information and Control, Vol.20, No.4, pp.1245-1256, 2024.
12. Z. Liu, J. K. Tan.: "Analysis of human walking posture using a wearable camera", International Journal of Innovative Computing, Information and Control, Vol.19, No.3, pp.805-819, 2023.
13. R.I. Hartley: "In Defence of the 8-point Algorithm", Proc. of IEEE Intl. Conf. on Computer Vision, pp.1064-1070,1995

Authors Introduction

Mr. Shunya Egashira



He received his Bachelor's degree in Engineering in 2023 from the Faculty of Engineering, Kyushu Institute of technology in Japan. He is currently a master student in Kyushu Institute of Technology, Japan. He is now interested in human posture analysis using an ego camera.

Prof. Dr. Yui Tanjo



Dr. Tanjo is currently Professor with the Department of Mechanical and Control Engineering, Kyushu Institute of Technology. Her current research interests include ego-motion analysis by MY VISION, three-dimensional shape/motion recovery, human detection, and its motion analysis from video. She was awarded SICE Kyushu Branch Young Author's Award in 1999, the AROB Young Author's Award in 2004, the Young Author's Award from IPSJ of Kyushu Branch in 2004, and the BMFSA Best Paper Awards in 2008, 2010, 2013 and 2015. She is a member of IEEE, The Information Processing Society, The Institute of Electronics, Information and Communication Engineers of Japan.