A Method of Recognizing Body Movements Based on a Self-viewpoint Video

Iichirou Moribe
Graduate School of Engineering, Kyushu Institute of Technology, 1-1 Sensui, Tobata-ku, Kitakyushu, 804-8550, Japan

Yui Tanjo
Faculty of Engineering, Kyushu Institute of Technology, 1-1 Sensui, Tobata-ku, Kitakyushu, 804-8550, Japan

Email: tanjo@.cntl.kyutech.ac.jp

Abstract

The most critical human sensory function resides in vision. This paper focuses on utilizing visual information, specifically self-perspective footage, to identify individual movements. Existing researches require third-party filming to recognize human body movements and states. The proposed method, on the other hand, simply attaches a camera to the human head and enables the recognition of the subject's actions. Consequently, it becomes easier to monitor daily movements of a human and gather his/her data on body kinetics. This approach would be beneficial in scenarios involving individuals engaging in risky behavior or, during a certain emergency, providing valuable assistance.

Keywords: My VISION, Posture estimation, Optical flow, HSV conversion

1. Introduction

Vision is the most crucial function among human sensory organs. Human intakes a vast amount of information through vision. Therefore, self-perspective footage can potentially serve as the primary source of information from an individual. In fact, it has been evident that visual information is more potent in perceiving human posture and bodily movements compared to non-visual information [1]. Therefore, it was considered possible to recognize the filmmaker's physical state by analyzing self-perspective footage.

Maintaining physical health has become more commonplace, exemplified by the widespread use of smartwatches. Smartwatches offer various functions such as recording heart rate, blood oxygen levels, and sleep patterns. By utilizing these features to log daily physical conditions, it becomes possible to promptly recognize deviations from normal bodily states.

Hence, this study aims to develop a method for analyzing and recording an individual's activities from a self-perspective video. Similar research includes studies on self-posture estimation [2], methods utilizing single-eye images based on HOG features [3], and methods employing deep learning [4]. However, these methods focus on action estimation or recognition rather than understanding a person's activity. Additionally, the techniques using HOG-based single-eye images or employing deep learning require capturing individuals externally, which is entirely different from the self-perspective video approach in this paper.

2. Methodology

This section describes feature extraction methods. The features to be extracted are the values of the hue and the norm of the optical flow on an image.

2.1. *Detection of area similarity*

This subsection describes the method for detecting area similarity. By dividing the image into multiple blocks and comparing the H (Hue) histogram features of specific blocks, area similarity is determined.

2.1.1 *HSV Conversion*

First, the RGB color image, which serves as the input image, is converted from the RGB color space to the HSV color space. The RGB color space represents a color space with red, green, and blue as coordinate axes, while the HSV color space represents a color space with hue, saturation, and value as coordinate axes.

2.1.2 *Creation of H(Hue) histogram*

In the proposed method, a histogram based on the Hue (H) is constructed. The values of S (Saturation) and V (Value) are not taken into account. The H histogram, as depicted in Fig. 1, represents the hue values on the horizontal axis and the frequency count on the vertical

axis. Although the original range of hue values is 1 to 360, for expedited processing, it is scaled to 1 to 180.

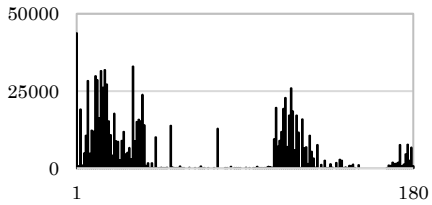


Fig. 1. H (Hue) Histogram

2.1.3 *Calculation of similarity*

The H histogram is used to determine the similarity between two images. Here, the similarity is calculated using the following formula. The meanings of the characters in Eq. (1) are given in Table 1.

$$S\ (A,B) = \frac{\sum_{i=0}^{N} \min(h_i^A,\ h_i^B)}{\sum_{i=0}^{N} h_i^A} \qquad (1)$$

Table 1. Meanings of the characters in Eq. (1)

N	Number of pixels in the image
h_i^x	The ith element of histogram x
$\min(x_A, x_B)$	Minimum value of x_A and x_B

2.2 *Derivation of optical flow*

This subsection describes the derivation of optical flow. The optical flow is the movement vector of the camera, or the subject obtained by mapping feature points in two consecutive frame images extracted from the video.

2.2.1 *Feature point extraction and description*

With respect to two consecutive frame images, the feature points in the first frame image are focused on. After extracting the feature points, the feature values of each feature point are described, and the feature points are matched in the two frame images. The Shi-Tomasi corner detection method is used for feature point extraction and feature point description [5]. The image from which the feature points were extracted is shown in Fig. 2.

Fig. 2. Input image with feature point extraction

2.2.2 *Feature point matching*

The Lucas-Kanade method is used for matching feature points [6]. The Lucas-Kanade method is one of the leading methods for deriving optical flow and is computationally less expensive than the method that finds all pixels in the image by searching for them.

2.2.3 *Removal of outliers*

The optical flow is obtained by the feature point matching, but the optical flow obtained from the actual video contains many outliers. Outliers are false flows obtained by matching different feature points and need to be removed for accurate analysis. Outliers are removed by applying RANSAC to modelling with homography matrices.

An example of the optical flow obtained by the above process is shown in Fig. 3. Note that the green points in the figure represent the end points of the optical flow.

Fig. 3. Derived optical flow.

2.3 *Normal state recognition*

In this subsection, the method of normal state recognition is first described. Next, the methods for classifying states are described.

2.3.1 *Feature extraction methods*

The input image is separated into nine blocks and each block image is assigned a number as shown in Fig. 4. Focusing on blocks 2, 5 and 8, the three block images are HSV-transformed and H histograms are created. The H histograms are compared to obtain the similarity between block 8 and block 2, and between block 8 and block 5, respectively. The similarity between block 8 and block 2 and between block 8 and block 5 is calculated by comparing the H histograms.

Block 1	Block 2	Block 3
Block 4	Block 5	Block 6
Block 7	Block 8	Block 9

Fig. 4. Separated input image

Furthermore, in the method, the number and norm of optical flows are the key features. First, the average value of the norm of the optical flow is calculated with each frame, and then the maximum value of the norm in the whole video is calculated. This value is used to evaluate the intensity of the motion of the entire video. The average value of the norm of the optical flow is calculated by the following equation. The meanings of the characters in Eq. (2) are shown in Table 2.

$$V_f = \frac{1}{N_f}\sum_{i=1}^{N_f} V_i \tag{2}$$

Table 2. Meanings of the characters in Eq. (2)

N_f	Total number of optical flows obtained between the fth frame and the f+1th frame
V_i	Norm of each optical flow

2.3.2 *Prior statistics*

Since the estimation of each state is done by comparison with the normal state, it is necessary to know in advance what properties each state has in relation to the normal state. For the three states, feature extraction is carried out using the above procedure with the three videos.

The results of the validation of the properties of the normal state are presented in Table 3. The results of the validation of the properties of the 'looking down' state are presented in Table 4. The results of the validation of the properties of the 'stumble' state are presented in Table 5.

Table 3. Results of the validation of the properties of the normal state

Scene	$\bar{S}$	V_{max}
Video 1	0.53	118
Video 2	0.40	109
Video 3	0.51	124

Table 4. Results of the validation of the properties of the 'looking down' state

Scene	$\bar{S}$	V_{max}
Video 1	1.29	113
Video 2	1.19	110
Video 3	1.06	75.6

Table 5. Results of the validation of the properties of the 'stumble' state

Scene	$\bar{S}$	V_{max}
Video 1	0.61	271
Video 2	0.54	181
Video 3	0.45	194

From these results, it can be read that $\bar{S}$ is greater in the 'looking down' state among the three states. This means that in the 'looking down' state, the similarity of block 2, block 5 and block 8 is greater. It can also be seen that V_{max} in the 'stumble' state is the largest among the three states. This indicates that the motion in the 'stumble' state is the most intense.

2.3.3 *State analysis*

In the proposed method, the motion states are classified into three states. Classification 1: If the value of S satisfies Eq. (3), the image is classified as a 'looking down' image. If the value of S does not satisfy Eq. (3), the image is classified as 'other' image. 'other' can be either 'normal' or 'stumble'. The state of Classification 1 is determined using the following formula.

$$S - \bar{S} \geq TH1 \tag{3}$$

Here S is the similarity in the current state and $\bar{S}$ is the average of the similarity in the normal state. $TH1$ is the threshold value in Classification 1.

Sr is the proportion of the images in the image group that are classified as 'looking down' in Classification 1, and if the value of Sr satisfies Eq. (4), the image is classified as "looking down". If the value does not satisfy Eq. (4), it is assumed to be in the 'other' state and proceeds to the next state estimation. If the value of $Vmax$ is satisfies Eq. (5), the state is judged as the 'stumble' state. If none of the above conditions is satisfied, the state is judged as normal. Eq. (4) and Eq. (5) are shown below.

$$Sr - \bar{S_r} \geq TH2 \tag{4}$$

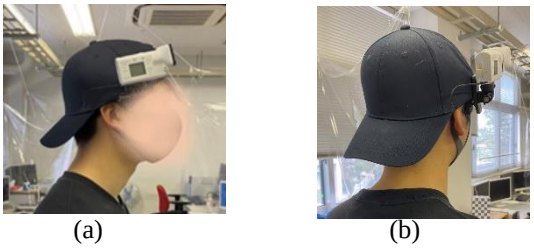
$$Vmax - \overline{V_{max}} \geq TH3 \tag{5}$$

Here Sr and $Vmax$ are the evaluated values of similarity and intensity of motion in the current condition, respectively. $\bar{S_r}$ and $\overline{V_{max}}$ are the average values of similarity and intensity of motion, respectively, in the normal condition. $TH2$ and $TH3$ are the threshold values for each condition.

3. Experimental Result

3.1. *Experimental methods*

An experiment is conducted with a camera worn on the head. Fig. 5 shows an image of the camera worn by a subject during the experiment.



(a) (b)

Fig. 5. Image of a worn camera. (a) side view,(b) rear view.

The details of the experimental method are as follows:

1. Feature extraction was carried out from five videos in the 'normal' state, and a database was created by averaging the feature values in each video.
2. Eight videos containing each of the three states - 'normal', 'looking down' and 'stumbling' - taken at different locations and at different dates and times were used as input videos to estimate each of the states.

The parameters during the experiment are shown in Table 6.

Table 6. Parameters' values

	Value	
Threshold 1	0.35	$TH1$ in Eq. (2-2)
Threshold 2	0.65	$TH2$ in Eq. (2-3)
Threshold 3	55	$TH3$ in Eq. (2-4)

3.2. *Result*

The results of the experiment are shown in Table 7. For each of the eight videos, the number of correct responses for the 'normal', 'looking down' and 'stumble' states were 6, 7 and 7 respectively.

Table 7. Results of the experiment.

state	normal	looking down	stumble
normal	6	0	2
looking down	1	7	0
stumble	1	0	7

4. Conclusion

In this paper, we proposed a method for estimating the activity state of the camera wearer from the self-viewpoint video obtained using a self-viewpoint camera. Three human walk states were detected by comparing H (Hue) histograms and using features calculated from the optical flow. In the experiment, three states were detected: 'normal', 'looking down' and 'stumble' states. The results showed that estimation was successful in more than 75% of all states. However, in the estimation of all states, misrecognition occurred in which a different state was estimated. Future issues include the estimation of states in case of sudden changes in vision and estimating more precise states corresponding to the angle of looking down.

References

1. D. Lee, E. Aronson: "Visual proprioceptive control of standing in human infants" , Attention Perception & Psychophysics, vol. 15, No.3, pp. 529-532, 1974.
2. 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.
3. K. Onishi, T. Takiguchi, Y. Ariki: "3D human posture estimation using the HOG features from monocular image" , 19th International Conference on pattern Recognition, pp. 1-4, 2008.
4. Z. Cao, T. Simon, S. E. 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.
5. J. Shi, C. Tomasi: "Good features to track" , IEEE Conference on Computer Vision and Pattern Recognition, 1994.
6. B. D. Lucas, T. Kanade: "An iterative image registration technique with an application to stereo vision" , Proceeding of Imaging Understanding Workshop, pp.121-130, 1981.

Authors Introduction

Mr. Iichirou Moribe

Mr. Moribe received his Bachelor's degree in Engineering in 2022 from the Control Engineering, Kyushu Institute of technology in Japan. He is currently a master student in Kyushu Institute of Technology, Japan.

Dr. Yui Tanjo

Dr. Tanjo is currently a 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 Award 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.