Journal of Robotics, Networking and Artificial Life
In Press, Corrected Proof
DOI: https://doi.org/10.2991/jrnal.k.201215.004; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

Human Motion Recognition Using TMRIs with Extended HOOF

Jing Cao[1], Youtaro Yamashita[1], Joo Kooi Tan[2,*]

[1]*Graduate School of Engineering, Kyushu Institute of Technology, Kitakyushu, Japan*
[2]*Faculty of Engineering, Kyushu Institute of Technology, Kitakyushu, Japan*

ARTICLE INFO

Article History
Received 10 November 2019
Accepted 17 June 2020

Keywords
Human motion
description
recognition
elderly care
crime prevention
MHI
triplet motion representation images
HOOF

ABSTRACT

In recent years, research on computer vision has shown great advancement and has been applied to a wide range of fields. Among them, automatic recognition of human motion is an important technology especially in crime prevention and elderly watching systems. Considering this trend, the paper proposes a novel method of human motion description and recognition using a motion history image-based method called triplet motion representation images and an extended feature descriptor called histograms of oriented optical flow which contains information on the direction and velocity of movement. One of the advantages of the proposed method over existent methods is that it solves a self-occlusive motion problem particularly in the depth direction which occurs when a single camera is used. The performance and effectiveness of the proposed method are verified by experiments.

1. INTRODUCTION

In modern society, aging trend is getting more and more serious. The increase in the number of elderly people, 65 years old or more, living alone is remarkable for both men and women in Japan. About 1.92 million men and about 4 million women among the total population of the elderly, 13.3% for men and 21.1% for women, live alone in 2015 [1]. Therefore, it is necessary to establish an elderly support/care system. According to Ministry of Health [2], 'fracture and fall' is the fourth leading cause of the need for protection for the elderly.

Nowadays, the Japanese elderly care system has been gradually improving. There are many care systems in the current market. For example, specialists regularly visit elderly homes to see his/her health status: They use contact sensors to monitor the usage of home appliances or even gas in the home to know the activities of a person in the house: 24-h monitoring systems are also popular. However, these systems are difficult to deal with emergency cases.

On the other hand, according to the 2017 public opinion poll on public security [3], 60.8% of the respondents answered that "domestic public security has deteriorated over the last 10 years." Surveillance cameras installed in a town contribute to increase the arrest rate for robbery on the street, but it is not a system to prevent crimes in advance.

In this paper, aiming at the development of a crime prevention and elderly watching system which will be used in an intelligent robot, we propose a novel human motion description and recognition method using a single camera. In conventional researches, Andrade et al. [4] proposed flow vectors for establishing optical flow distributions in regions of interest. Bobick and Davis [5] proposed Motion History Image (MHI) to describe historical motion using a single camera. Mitchelson and Hilton [6] proposed three-dimensional restoration of motion. However, most of the methods of motion recognition using a single camera such as MHI assume that the motion for recognition is on the plane perpendicularto the optical axis of the camera, and the motion toward the camera or leaving from the camera is not dealt with because of its self-occlusive nature. To describe a self-occlusive motion, Tan et al. [7] proposed a reverse MHI method and 3D-MHI [8]. But they had a drawback of computational cost. In this paper, we propose a method which can deal with the motions to and from a camera by expanding the MHI. The method is named Triplet Motion Representation Images (TMRIs) [8] with extended Histogram of Oriented Optical Flow (HOOF). It represents depth information, i.e. approaching or leaving from a camera, by three characteristic images. The performance of the proposed method is shown experimentally.

In the following, the proposed method is described in Section 2. Performed experiment and evaluation are presented in Section 3. The experimental results are discussed and the paper is concluded in Section 4.

2. PROPOSED METHOD

The proposed method is mainly separated into five steps. The first step is a foreground extraction. Since a background may change over time, extraction of a human region is performed using a background

*Corresponding author. Email: etheltan@cntl.kyutech.ac.jp

model that can cope with changes in the background. In the second step, optical flow is used to represent human movement. If the lines representing the optical flow in an image meet at a point, it is regarded as a Focus of Expansion (FoE), indicating that the motion contains the movement toward an observing camera or away from the camera. If no FoE is found, the motion is considered to be in a plane perpendicular to the optical axis of the camera. The third step is to extend the traditional MHI method into TMRIs to describe a human motion. The TMRIs are then transformed to Hu moments representation, providing a TMRI feature vector. In the fourth step, optical flow is extracted to know motion direction and its speed. For this purpose, HOOF is used. It gives a HOOF feature vector. Finally, in the fifth step, a classifier is trained using these two feature vectors. Motion recognition is performed by use of the classifier.

2.1. Foreground Extraction

In order to extract a human region from the temporal videos, successive background image inference is performed using a Gaussian Mixture Model (GMM) [9] for each pixel constituting the image. The influence of the external environment such as the light mutation in the background can be effectively ignored by the background image inference. In this study, the Expectation-Maximization (EM) algorithm is used to determine the means and variances of the initial parameters of GMMs in the background video that does not contain a moving object.

2.2. Computation of a FoE

In order to keep the detection of FoE stable, feature points are equally spaced on the outline of the extracted human region. In this study, LK tracker [10] is used to calculate the optical flow between successive frames. Figure 1a shows an example of optical flow of human movement. The optical flow lines are extended on the voting plane to find a FoE, as shown in Figure 1b. If the number of lines intersecting a single point exceeds a predetermined threshold, the point is regarded as a FoE.

Because of possible wrong detection of a FoE due to the deviation of the voting, this paper adopts a weighted voting method. It gives weight to the extended optical flow line. A larger weight is given to the pixels close to the line, whereas smaller weight is given to the pixels distant from it. A FoE exists if the result of the final vote is greater than a predetermined threshold.

2.3. Description of a Human Motion: TMRIs

The original two-dimensional (2-D) MHI is extended in the proposed method, so that it may describe a motion in a 3-D way. To realize this, we propose a motion description method using three kinds of motion history images called TMRIs. They are *newness*, indicating the original 2-D MHI, *density*, indicating the frequency of appearance in the past τ frames, and *depth*, showing the movement of the object in the depth direction. The definition of each image is given as follows:

$$H_\tau^{\text{new}}(x,y,t)=\begin{cases}\tau & \text{if } D(x,y,t)=1\\ \max(0,\ H_\tau^{\text{new}}(x,y,t-1)-1) & \text{otherwise}\end{cases}. \quad (1)$$

$$H_\tau^{\text{den}}(x,y,t)=\sum_{i=0}^{\tau} D(x,y,t-i). \quad (2)$$

$$H_\tau^{\text{dep}}(x,y,t)=\sum_{i=0}^{\tau}\left\{N_{\text{layer}}(t-i)\times D(x,y,t-i)\right\}. \quad (3)$$

Here $H_\tau(x, y, t)$ is the gray value at (x, y, t) and D is a binary image showing a foreground region.

$N_{\text{layer}}(t)$ is the number of layers to be overlapped at time t and defined by

$$N_{\text{layer}}=\begin{cases}\gamma\times L_{\text{ave}} & \text{if } V_{\max}>T_{\text{vote}}\\ 1 & \text{otherwise}\end{cases}. \quad (4)$$

Here $V_{\max}$ is the maximum value with voting results: T_{vote} is the voting threshold: γ is a constant used to determine the value of N_{layer} according to the size of optical flow: L_{ave} is the average length of optical flow.

The TMRIs w.r.t. walk to the right and walk to the back are shown in Figure 2.

The features of TMRIs are described using Hu moments [11]. Hu moments are invariant to image scaling, rotation and flipping, and defined by seven invariant features v_i (i = 1, 2, …, 7). Since they have large difference in scale, they are changed to logarithmic values by $s_i = \log(sign(v_i)\cdot v_i)$. The feature vector is then defined as $\boldsymbol{s} = (s_1, s_2, ..., s_7)$.

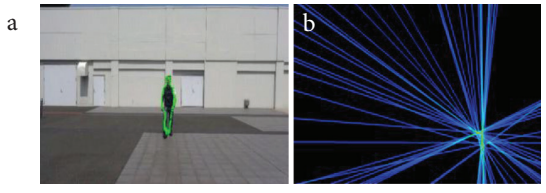


Figure 1 | Extension of optical flow lines: (a) Optical flow detected from a human motion, (b) FoE on a voting plane.

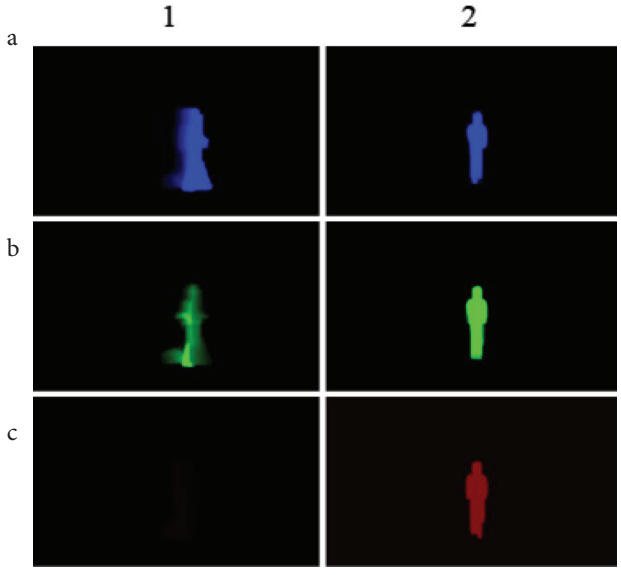


Figure 2 | Examples of TMRIs: (a) newness, (b) density, (c) depth: (1) walk to the right, (2) walk to the back.

Since TMRIs contain three kinds of motion history images and each image is described by Hu moments, they are integrated to form a 2-D vector as

$$V^{\mathrm{TMRI}} = (\boldsymbol{s}^{\mathrm{new}}, \boldsymbol{s}^{\mathrm{den}}, \boldsymbol{s}^{\mathrm{dep}}). \tag{5}$$

2.4. Description of a Human Motion: HOOF

The direction and the speed of a motion should also be extracted, since TMRIs does not represent them. For this purpose, the HOOF [12] method is used in an extended manner to describe the information in terms of the number of bins and the weights in calculating the histogram frequency. As shown in Figure 3, the direction of optical flow vectors are separated into n ($n = 30$ in the experiment) directions and the length of the optical flow of respective directions is accumulated to form a histogram representing the magnitude of optical flow with each direction.

Given an optical flow vector $\boldsymbol{v} = (x, y)^{\mathrm{T}}$, the angle $\theta = \tan^{-1}(y/x)$ and the length $L = \sqrt{x^2 + y^2}$ are calculated. If θ is the angle between bin a and bin b, the weights w_a, w_b are calculated by

$$w_a = \frac{\left|\theta - 2\pi/n \times (b-1)\right|}{2\pi/n}. \tag{6}$$

$$w_b = \frac{\left|\theta - 2\pi/n \times (a-1)\right|}{2\pi/n}. \tag{7}$$

Then, $L \times w_a$ and $L \times w_b$ are accumulated to bin a and bin b, respectively.

As defined in Equations (1)–(4), a motion is described using τ successive frames. Optical flow vectors are calculated from every other frame among the τ frames. All these optical flow vectors in the past τ frames are accumulated in a single histogram, and normalization is performed so that the total frequency becomes 1. The feature vector of HOOF having n bins is described by

$$\boldsymbol{V}^{\mathrm{HOOF}} = (f_1, f_2, \ldots, f_n). \tag{8}$$

Here, f_i ($i = 1, 2, \ldots, n$) is the frequency of the ith bin under the condition that $\|\boldsymbol{V}^{\mathrm{HOOF}}\| = 1$.

As $n = 30$ in the experiment, $\boldsymbol{V}^{\mathrm{HOOF}}$ is a 3-D feature vector.

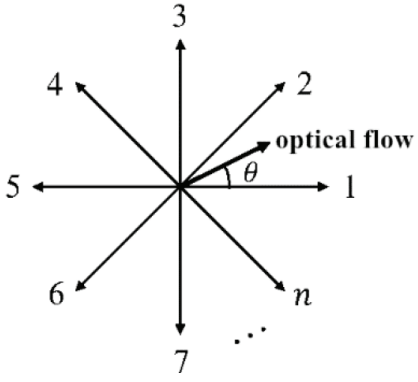


Figure 3 | The extended HOOF.

2.5. Motion Recognition

The feature vector $\boldsymbol{V}$ used in the proposed method is a 5-D vector given by

$$\boldsymbol{V} = (\boldsymbol{V}^{\mathrm{TMRI}}, \boldsymbol{V}^{\mathrm{HOOF}}). \tag{9}$$

A k-nearest neighbor classifier is used for the recognition. The class of motion l is denoted by C^l, the jth learning data in C^l by $\boldsymbol{v}_j^l$, and the input unknown motion is by v. Then the procedure for the k-nearest neighbor is given by

$$l^* = maj_l\left\{\arg_{v_j^l} k\min\left\{I(\boldsymbol{v}, \boldsymbol{v}_j^l)\forall l, \forall j, \boldsymbol{v}_j^l \in C^l\right\}\right\}. \tag{10}$$

Here l^* represents the recognized class: $maj_l\{S\}$ returns the class that appears the most in set S: k min$\{T\}$ is the k minimum numbers in the set T: I is the dissimilarity measure defined by $I(\boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{v} - \boldsymbol{w}\|$.

3. EXPERIMENTAL RESULTS

In order to verify the performance of the proposed method, experiments were carried out on the accuracy of detection of FoE and motion recognition. The experimental video contains 12 different motions; walk left/right/front/back, walk left-front/right-front, walk left-rear/right-rear and fall left/right/front/back. Each motion was done once by four people (univ. students). Among them, six motions are represented by TMRIs and shown in Figure 4. It is noted that, as described by Equations (1)–(4), a motion is given in the form of three images (newness, density and depth). They are changed into color images using R, G and B, respectively, and superposed to make a single image. It is shown in Figure 4.

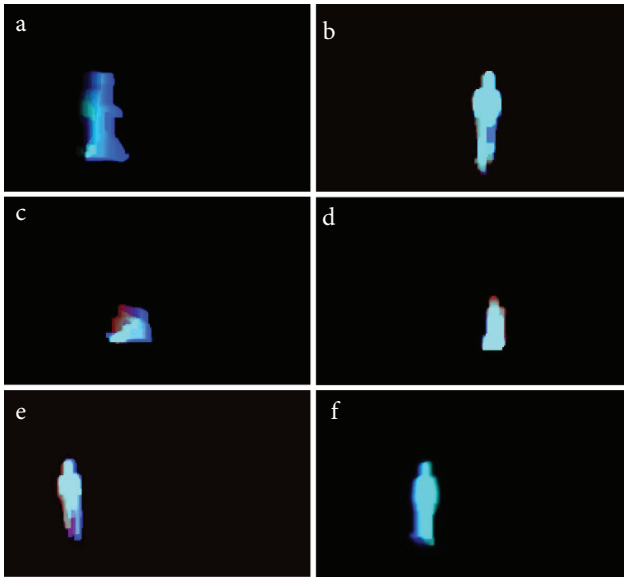


Figure 4 | Examples of motions described by TMRIs: (a) walk right, (b) walk back, (c) fall right, (d) fall back, (e) walk right front, (f) walk left rear.

Table 1 | Recognition rate of the motions

Motion	Recognition rate (%)			
	MHI	TMRIs	HOOF	TMRIs + HOOF (proposed)
Walk left	28.13	66.88	93.44	96.88
Walk right	20.31	50.31	75.31	85.94
Walk front	16.56	74.69	42.50	68.44
Walk back	10.00	71.56	87.50	94.69
Walk left front	12.19	36.56	92.19	97.50
Walk right front	37.19	63.75	70.00	87.19
Walk left rear	13.75	35.63	79.06	84.38
Walk right rear	25.94	45.63	82.19	91.25
Fall left	24.38	28.13	66.88	82.50
Fall right	25.63	33.13	54.06	68.44
Fall front	22.81	43.44	28.13	33.13
Fall rear	18.44	57.19	70.63	84.06
Average	21.28	50.57	70.16	81.20

3.1. Detection of a FoE

When the maximum count in the voting result exceeds a threshold (= 20 in this experiment), a FoE is detected.

The accuracy of detection is given by the following equation:

$$\text{Accuracy} = \frac{F_{\text{TP}} + F_{\text{TN}}}{F_{\text{ALL}}} \times 100\,(\%). \tag{11}$$

Here F_{ALL} is the total number of frames for which detection of a FoE was performed: F_{TP} is the number of frames in which the FoE is detected correctly for motion in depth direction: F_{TN} is the number of frames for which the FoE is not detected correctly for other motions.

The accuracy with all the motions was 64.3% in average.

3.2. Motion Recognition

In the experiment, TMRIs are computed using Equations (1)–(4) with $\tau = 10$. Given a motion video, 80 sets of 10 successive image frames are chosen in order from the frame where the motion starts. Since each of the four people performs 12 motions, and 80 sets of features, $\boldsymbol{V}$ of Equation (9), are yielded from each video, 3840 feature vectors are obtained in all. Leave-one-out cross validation is performed by separating the data into four groups to evaluate the accuracy of motion recognition. With respect to the k-nearest neighbor algorithm, k is set to 3. For comparison, motion recognition is done using the traditional MHI [5], $\boldsymbol{V}^{\text{TMRI}}$ and $\boldsymbol{V}^{\text{HOOF}}$. All these results are given in Table 1.

4. DISCUSSION AND CONCLUSIONS

In the proposed method, an arbitrary motion is described first by the TMRIs and then transformed into a feature vector consisting of Hu moments. An extended HOOF is also used to extract a feature vector on the direction and speed of each motion. The proposed method uses an overall feature vector V of Equation (9) composed of the above two feature vectors. An unknown motion is represented by V and recognized using a 3-nearest neighbor classifier. The accuracy of motion recognition by the proposed method was verified by experiments and an expected result (81.2% recognition rate) was obtained as shown in Table 1. The rate is much higher than the recognition rates of the three comparative experiments; use of MHI, and respective use of TMRIs and HOOF.

Although the recognition accuracy is acceptable in the performed experiment, the proposed method requires further improvements to make it have higher recognition rates and more stable performance by increasing training data, tuning used parameters, etc.

In most of the studies on human motion recognition, the direction of the motions they deal is almost perpendicular to a camera's optical axis. On the other hand, TMRIs proposed in this paper can describe the motion toward the depth (optical axis) direction under a single camera. This is the main advantage of the present method over other methods. Automating the task of human motion recognition is important and expected particularly by a future robot working together with a human. The goal of the present study is to realize a strong method of human motion recognition not influenced by the direction of motion. In this sense, the present study is worth developing further.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

ACKNOWLEDGMENT

This study was supported by JSPS KAKENHI Grant Number 16K01554.

REFERENCES

[1] Cabinet Office: 2017 White Paper on Aging Society (Overall version), 2017, pp. 2–6 (in Japanese). Available from: https://www8.cao.go.jp/kourei/whitepaper/w-2017/zenbun/29pdf_index.html.

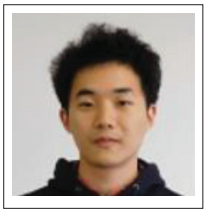
[2] Ministry of Health, Labor and Welfare: 2016 National Life Basic Survey, 2016, pp. 28–29 (in Japanese). Available from: https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa16/.
[3] Cabinet Office: Outline of the 2017 Public Opinion Poll on Security, 2017, pp. 5–20 (in Japanese). Available from: https://survey.gov-online.go.jp/tokubetu/h29/h29-chiang.pdf.
[4] E.L. Andrade, R.B. Fisher, S. Blunsden, Detection of emergency events in crowded scenes, Proceedings of IEEE International Symposium on Imaging for Crime Detection and Prevention, London, UK, 2006, pp. 528–533.
[5] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001), 257–267.
[6] J. Mitchelson, A. Hilton, Simultaneous pose estimation of multiple people using multiple-view cues with hierarchical sampling, Proceedings of the British Machine Vision Conference, BMVC, Norwich, UK, 2003, pp. 1–10.
[7] J.K. Tan, S. Okae, Y. Yamashita, Y. Ono, A method of describing a self-occlusive motion – a reverse motion history image, Int. J. Biomed. Soft Comput. Human Sci. 24 (2019), 1–7.
[8] Y. Yamashita, J.K. Tan, S. Ishikawa, Human motion description and recognition under arbitrary motion direction, 2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), IEEE, Kanazawa, Japan, 2017, pp. 110–115.
[9] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, Proceedings of 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, IEEE, Fort Collins, CO, USA, 1999, pp. 246–252.
[10] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vol. 2, Association for Computing Machinery, IJCAI, Vancouver, BC, Canada, 1981, pp. 674–679.
[11] M.K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Inform. Theory 8 (1962), 179–187.
[12] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions, 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, USA, 2009, pp. 1932–1939.

AUTHORS INTRODUCTION

Ms. Jing Cao

She received her B.E. from Republic of China and M.E. from the Graduate School of Engineering, Kyushu Institute of Technology, Japan in 2020. Her research includes computer vision, machine learning and human motion recognition.

Mr. Youtaro Yamashita



He received B.E. and M.E. from Kyushu Institute of Technology, Japan. His research includes image processing, human motion representation and recognition.

Prof. Dr. Joo Kooi Tan

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