Voting-based Hand-Waving Gesture Spotting from a Low-Resolution Far-Infrared Image Sequence

Yasutomo Kawanishi [*,1], Chisato Toriyama [**,§], Tomokazu Takahashi [†], Daisuke Deguchi [‡], Ichiro Ide [*],
Hiroshi Murase [*], Tomoyoshi Aizawa [⋆], Masato Kawade [⋆]

[*] *Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*
[**] *Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*
[†] *Faculty of Economics and Information, Gifu Shotoku Gakuen University, 1-38, Nakauzura, Gifu-shi, Gifu, Japan*
[‡] *Information Strategy Office, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*
[⋆] *Corporate R&D, OMRON Corporation, 9-1, Kizugawadai, Kizugawa-shi, Kyoto, Japan*
[1] kawanishi@i.nagoya-u.ac.jp
[§] Currently at Brother Industries, Ltd., Mizuho-ku, Nagoya-shi, Aichi, Japan

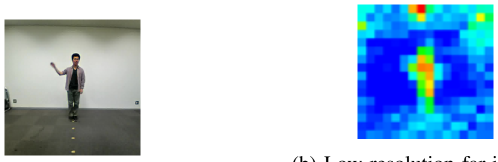
***Abstract*—We propose a temporal spotting method of a hand gesture from a low-resolution far-infrared image sequence captured by a far-infrared sensor array. The sensor array captures the spatial distribution of far-infrared intensity as a thermal image by detecting far-infrared waves emitted from heat sources. It is difficult to spot a hand gesture from a sequence of thermal images captured by the sensor due to its low-resolution, heavy noise, and varying duration of the gesture. Therefore, we introduce a voting-based approach to spot the gesture with template matching-based gesture recognition. We confirm the effectiveness of the proposed temporal spotting method in several settings.**

***Index Terms*—Far-infrared sensor array, gesture spotting, voting**

I. Introduction

Gesture is considered as one of the most intuitive means to control appliances in a daily scene. Accordingly, various gesture interfaces have been proposed [1]. As one simple application, we can consider turning on/off a room-light or a TV by a hand-waving gesture. In this paper, we propose a method to realize such a gesture interface by recognizing a reference gesture registered beforehand from a low-resolution far-infrared image sequence.

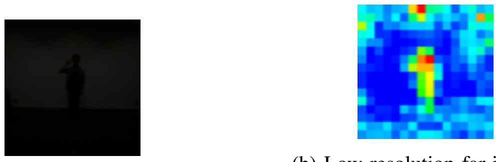
For sensing a gesture, a camera is usually used. There are some researches for vision-based gesture recognition using a visible-light camera. For example, Fujii et al. proposed a method that focused on the change of arm directions during a gesture [2]. This method extrapolates arm directions from human joints captured by Microsoft's Kinect sensor [3]. Meanwhile, Mohamed et al. proposed a method for tracking a hand trajectory [4]. This method detects user's hands based on skin tone and motion information. However, we cannot always make use of visible-light cameras anywhere and/or anytime because they do not work well in dark. As shown in Fig. 1 (a), we can observe a user and his/her gesture in light such as during the day-time or in a well-lit room. On the contrary, as shown in Fig. 2 (a), we cannot observe the user and his/her gesture in dark such as during the night-time or in a poorly-lit room.



(a) Visible-light image

(b) Low-resolution far-infrared image

Fig. 1. Examples of an output of a visible-light camera and a 16 × 16 far-infrared sensor array in light.



(a) Visible-light image

(b) Low-resolution far-infrared image

Fig. 2. Examples of an output of a visible-light camera and a 16 × 16 far-infrared sensor array in dark.

To avoid such a problem, using a far-infrared sensor array [5] can be a good solution. Images captured by a 16 × 16 far-infrared sensor array are shown in Figs. 1 (b) and 2 (b). Although the images are rather noisy, they can capture the spatial distribution of far-infrared intensity as a thermal image by detecting far-infrared waves emitted from heat-sources. Recently, several methods using such a sensor have been reported [6], [7]. With these methods, as illustrated in Fig. 3, we can turn on/off an appliance by waving our hands toward it.

In this paper, we propose a method for hand-waving gesture spotting using a far-infrared sensor array. Here, spotting stands for detecting a sequence of frames where a user is performing the registered hand-waving gesture. Since it is difficult to spot hand-waving gestures with varying durations, a voting-based

IEEE VCIP'18, Dec. 9 – Dec. 12, 2018, Taichung, Taiwan.
000-0-0000-0000-0/00/$31.00 ©2018 IEEE.

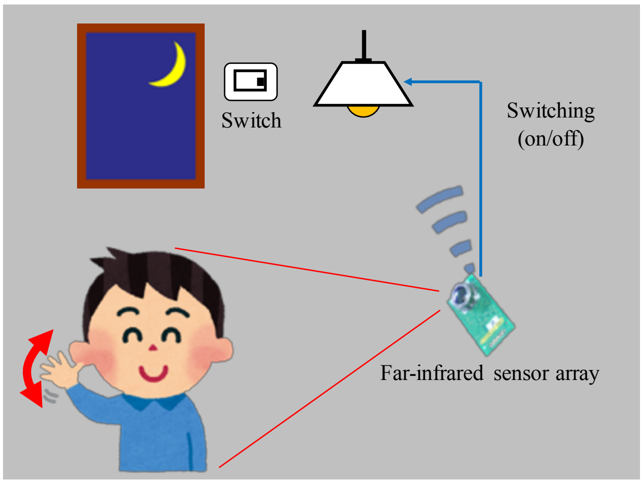


Fig. 3. Example of an application of the proposed method.

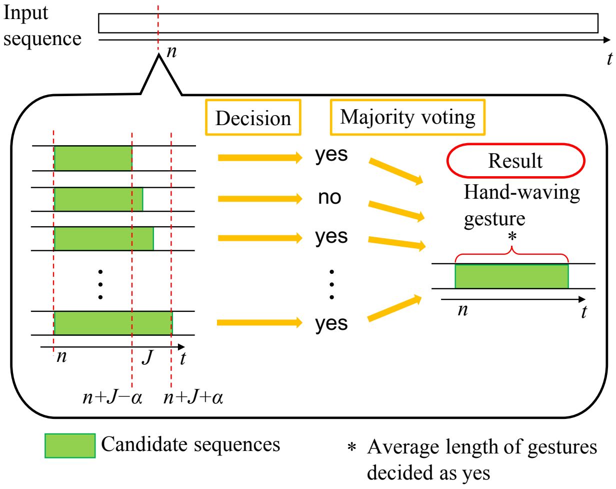


Fig. 4. Process-flow of the proposed method.

approach is employed. To accurately spot the gesture using the low-resolution and noisy sensor, the gesture and background clutter should be distinguished.

Our contributions are summarized as follows:

- To make it robust to varying durations of gestures, we propose a voting-based spotting framework using a Dynamic Time Warping (DTW)-based distance metric.
- To make it robust to noise, we use the concept "Spatial and Thermal Region of Interest (STRoI)" proposed by Toriyama et al. [8] for casting a vote or not for each candidate sequence.
- We show the superiority of the proposed method achieves the best performance through an evaluation performed on datasets captured in several conditions.

II. Voting-based Hand Gesture Spotting

When spotting a hand-waving gesture, it is difficult to accurately detect its beginning and end, since the duration of the registered gesture and that of the gesture that the user is actually performing are different. To deal with this problem, we introduce a voting-based framework with a Dynamic Time Warping (DTW)-based distance metric to robustly spot the hand-waving gesture.

Fig. 4 shows the process-flow of the proposed method at time n. First, candidate sequences with various durations starting from time n are cropped from a given input sequence. Then, for each candidate sequence, if it appears to include the hand-waving gesture, a vote is cast. If the majority of the votes are cast to any of the candidate sequences as the target gesture, the duration of the candidate sequences with votes are averaged. Finally, this average is output as the duration from time n as the spotted result. Note that the target hand-waving gesture is supposed to be registered as a reference sequence r prior to the gesture spotting process. Details of each process are described below.

A. Candidate Sequence Cropping

At time n, candidate sequences with various durations are cropped from the input sequence in order to allow handling gestures with various durations. Assuming that the length of the reference sequence is J, the length of the candidate sequences are set to $J-\alpha$ to $J+\alpha$ in a one-frame interval. As a result, $2\alpha+1$ candidate sequences $C=\{c_i\}_{i=1}^{2\alpha+1}$ are cropped.

B. Voting for a Candidate Sequence

For each candidate sequence $c \in C$, the decision whether c includes the target gesture or not is made by template matching. Since the length of c may differ from the reference sequence r, a distance metric which can handle input sequences with different lengths is required. Here, we use template matching with a DTW-based distance metric, which is proposed in Toriyama et al.'s work [8]. In their method, two input sequences are normalized and enhanced by focusing on "Spatial and Thermal Region of Interest (STRoI)" for accurate matching. The STRoI emphasizes the human body and moving regions.

By applying template matching, the decision result $d \in \{\text{yes}, \text{no}\}$ is obtained. Here, yes indicates that the candidate sequence c appears to include the target gesture and a vote is cast, and no, vice versa.

C. Aggregation of Voted Sequences

Here, let's assume that the decision results D, which consists of $2\alpha+1$ pairs of a candidate sequence $c_i \in C$ and its decision result d_i, have been obtained for time n. If the number of $d_i = \text{yes}$ is the majority (more than α), the merging process described as follows is preformed: First, the gesture candidates $E = \{c_i | d_i = \text{yes}, \forall(c_i, d_i) \in D\}$ are filtered. Then, the average length $\hat{J}$ of $c \in E$ is calculated. Finally, the sequence from time n to $n+\hat{J}$ is cropped as the output which includes the target gesture.

TABLE I
DATASETS USED IN THE EXPERIMENT.

Data group	A	B	C
Background heat source	—	✓	—
Sensor position	Front	Front	Above
Observation distance (reference)	150 cm	150 cm	200 cm
Observation distance (inputs)	90–270 cm	90–270 cm	200 cm
Input gesture	Hand wave, Stretch, Twist, Scratch one's head, Cross one's arms	Hand wave, Stretch, Twist, Scratch one's head, Cross one's arms	Hand wave, Stretch, Twist, Roll over, Pick up
Pose	Standing, Sitting	Standing, Sitting	Lying, Sitting, Relaxing
# of persons	6	5	3
# of datasets	11	13	8

III. EVALUATION

To evaluate the performance of the proposed method, we conducted an experiment on frame sequences captured using a far-infrared sensor array (Thermal sensor D6T-1616L by OMRON Corp.) in 10 fps. The sequences included several persons waving his/her hand or not. We describe below the specification of the dataset, experimental conditions, and then report and discuss the results from the experiment.

A. Datasets

We used the same datasets as in Toriyama et al.'s work [8]. A brief introduction of the datasets is as follows.

The target gesture in this experiment was "waving the right hand twice during approximately 4 seconds". It consists of 32 datasets, where a dataset consists of a reference sequence and a number of input sequences. They are divided into three groups by the capturing environments.

- Group A: Simple situation from the front
- Group B: Cluttered situation from the front
- Group C: Captured from the ceiling

Here, the cluttered situation had several heat sources in the background.

Details of the capturing conditions are summarized in Table I, and several examples from the datasets are listed in Fig. 5.

B. Experimental Condition

We evaluated the performance of the gesture spotting. To confirm the effectiveness of the voting-based approach, we compared it with a Comparative method [9]. We also compared the Proposed method with a Baseline method to analyze the effectiveness of the STRoI. The conditions of these methods are as follows:

- Proposed method: Voting-based method using STRoI.
- Baseline method: Voting-based method without STRoI.
- Comparative method: Using Discrete Fourier Transform (DFT) [9]

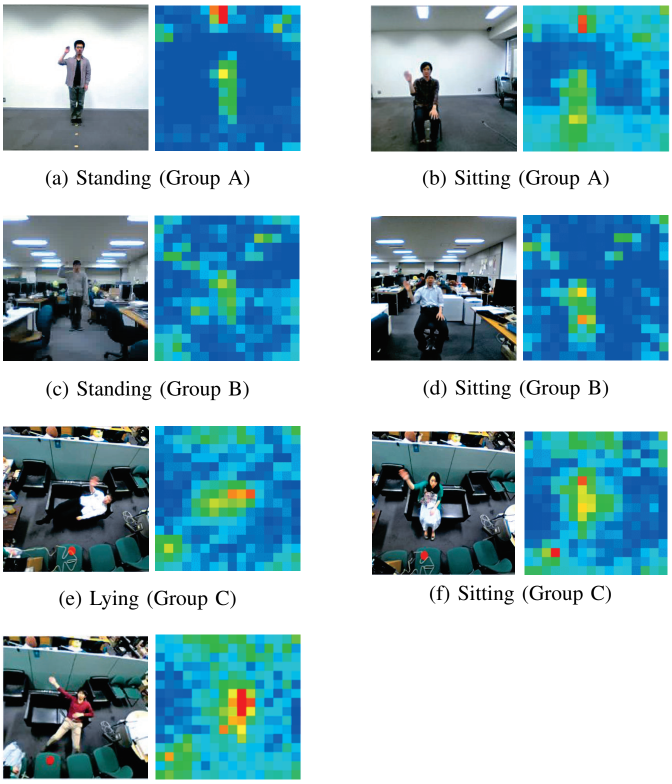

(a) Standing (Group A)
(b) Sitting (Group A)
(c) Standing (Group B)
(d) Sitting (Group B)
(e) Lying (Group C)
(f) Sitting (Group C)
(g) Relaxing (Group C)

Fig. 5. Sample images from the datasets.

The Proposed and the Baseline methods take the voting-based approach, while the Comparative method does not. The Proposed method used STRoI to emphasize the human body and motion, while the Baseline method used the region including the entire body region for the decision.

For evaluation, we used precision and recall of the spotted result of the target gesture.

For each sequence including a true gesture, given a spotted result, we judged whether the spotting is correct or not by the following two metrics: *correct spotting ratio over spotted results* t_{sr} (1) and *correct spotting ratio over the ground truth* t_{gt} (2);

$$t_{\mathrm{sr}} = \frac{|b|}{|b+c|}, \tag{1}$$

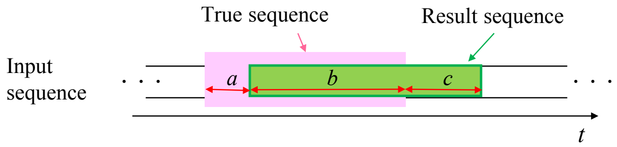


Fig. 6. Ground truth and a spotted result.

TABLE II
EXPERIMENTAL RESULTS (PRECISION).

	Data group			
	A	B	C	All
Proposed method	**0.57**	**0.57**	**0.77**	**0.62**
Baseline method	0.50	0.47	0.33	0.44
Comparative method (DFT)	0.26	0.24	0.38	0.28

$$t_{\text{gt}} = \frac{|b|}{|a+b|}, \tag{2}$$

where $a + b$ is the sequence which the true gesture is being performed and $b + c$ is the sequence of the spotted result, respectively, as shown in Fig. 6. Here, $|\cdot|$ indicates the duration of a sequence. For the correct spotting ratio over spotted results, we judge the result as correct when $t_{\text{sr}} > 0.5$. For the correct spotting ratio over the ground truth, we judge the result as correct when $t_{\text{gr}} > 0.5$.

Finally, The precision and recall are calculated as follows:

$$\text{precision} = \frac{\text{\# of the result sequences where } t_{\text{sr}} > 0.5}{\text{\# of all of the result sequences}} \tag{3}$$

$$\text{recall} = \frac{\text{\# of the result sequences where } t_{\text{gr}} > 0.5}{\text{\# of all of the ground-truth sequences}} \tag{4}$$

C. Results and Discussion

The results are shown in Tables II and III. They show the precision and the recall of the gesture spotting results for each group. As shown in these tables, the Proposed method achieved the best performance in all cases.

Compared to the Comparative method, the proposed voting-based approach performed better. We consider that this was because, although the input images captured by the far-infrared sensor array were noisy due to the air flow, the STRoI could reduce the noise and perform better. Since the Proposed method is voting-based, gestures are matched correctly even when the duration of the gestures are different from the target gesture. Meanwhile, the Comparative method that focuses on the periodicity of the time series of the pixel value was easily affected by noise.

IV. CONCLUSION

In this paper, we proposed a voting-based hand-waving gesture spotting method using a far-infrared sensor array. The proposed method detects a reference gesture, which is captured beforehand, from an input sequence. Since the duration of the gesture is different, we introduced a voting-based framework using a DTW-based distance metric. Since the sensor output is noisy and in low-resolution, we used the concept of "Spatial and Thermal Region of Interest (STRoI)". Experimental results showed that the voting approach was effective in spotting gestures with various durations. Furthermore, the STRoI was effective by combining it with the proposed method.

TABLE III
EXPERIMENTAL RESULTS (RECALL).

	Data group			
	A	B	C	All
Proposed method	**0.52**	**0.58**	**0.69**	**0.59**
Baseline method	0.26	0.35	0.19	0.28
Comparative method (DFT)	0.20	0.17	0.17	0.18

As future work, we will introduce a more effective preprocessing method to improve the spotting performance of the proposed method. Combining our method with a deep learning framework, which is actively developing recently, is also another direction of our future work.

ACKNOWLEDGMENT

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research.

REFERENCES

[1] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 3, pp. 311–324, May 2007.

[2] T. Fujii, J. H. Lee, and S. Okamoto, "Gesture recognition system for human-robot interaction and its application to robotic service task," in *Proc. Int. MultiConf. of Engineers and Computer Scientists 2014*, vol. 1, Hong Kong, SAR, China, Mar. 2014, pp. 63–68.

[3] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, Dec. 2013.

[4] M. Alsheakhali, A. Skaik, M. Aldahdouh, and M. Alhelou, "Hand gesture recognition system," in *Proc. 2nd Int. Conf. on Information & Communication Systems*, 2011, pp. 132–136.

[5] M. Ohira, Y. Koyama, F. Aita, S. Sasaki, M. Oba, T. Takahata, I. Shimoyama, and M. Kimata, "Micro mirror arrays for improved sensitivity of thermopile infrared sensors," in *Proc. IEEE 24th Int. Conf. on Micro Electro Mechanical Systems*, Cancun, Q. Roo, Mexico, Jan. 2011, pp. 708–711.

[6] T. Hosono, T. Takahashi, D. Deguchi, I. Ide, H. Murase, T. Aizawa, and M. Kawade, "Human tracking using a far-infrared sensor array and a thermo-spatial sensitive histogram," in *Computer Vision —ACCV 2014 Workshops*, ser. Lecture Notes in Computer Science, vol. 9009. Singapore: Springer, 2015, pp. 262–274.

[7] T. Kawashima, Y. Kawanishi, D. Deguchi, I. Ide, H. Murase, T. Aizawa, and M. Kawade, "Action recognition from extremely low-resolution thermal image sequence," in *Proc. 14th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Lecce, Puglia, Italy, Aug. 2017, pp. 1–6.

[8] C. Toriyama, Y. Kawanishi, T. Takahashi, D. Deguchi, I. Ide, H. Murase, T. Aizawa, and M. Kawade, "Hand waving gesture detection using a far-infrared sensor array with thermo-spatial region of interest," in *Proc. 11th Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 4, Rome, Italy, Feb. 2016, pp. 545–551.

[9] M. Takahashi, K. Irie, K. Terabayashi, and K. Umeda, "Gesture recognition based on the detection of periodic motion," in *Proc. 2010 Int. Symposium on Optomechatronic Technologies*, Toronto, ON, Canada, Oct. 2010, pp. 1–6.