Hand Waving Gesture Detection using a Far-infrared Sensor Array with Thermo-spatial Region of Interest

Chisato Toriyama[1], Yasutomo Kawanishi[1], Tomokazu Takahashi[2], Daisuke Deguchi[3], Ichiro Ide[1], Hiroshi Murase[1], Tomoyoshi Aizawa[4] and Masato Kawade[4]

[1]*Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*

[2]*Faculty of Economics and Information, Gifu Shotoku Gakuen University, 1-38, Nakauzura, Gifu-shi, Gifu, Japan*

[3]*Information Strategy Office, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*

[4]*Corporate R&D, OMRON Corporation, 9-1, Kizugawadai, Kizugawa-shi, Kyoto, Japan*

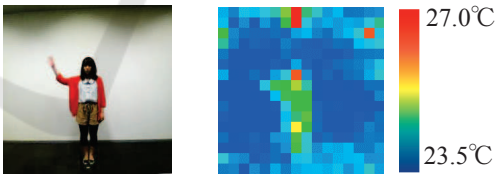
Keywords: Far-infrared Sensor Array, Gesture Detection.

Abstract: We propose a method of hand waving gesture detection using a far-infrared sensor array. The far-infrared sensor array captures the spatial distribution of temperature as a thermal image by detecting far-infrared waves emitted from heat sources. The advantage of the sensor is that it can capture human position and movement while protecting the privacy of the target individual. In addition, it works even at night-time without any light source. However, it is difficult to detect a gesture from a thermal image sequence captured by the sensor due to its low-resolution and noise. The problem is that the noise appears as a similar pattern as the gesture. Therefore, we introduce "Spatial Region of Interest (SRoI)" to focus on the region with motion. Also, to suppress the influence of other heat sources, we introduce "Thermal Region of Interest (TRoI)" to focus on the range of the human body temperature. In this paper, we demonstrate the effectiveness of the method through an experiment and discuss its result.

1 INTRODUCTION

Gesture has been drawing attention as a means of user interfaces. For example, with a gesture interface, we can easily control appliances by performing gestures intuitively using our own body. Especially, hand waving is one of the simplest and the most intuitive gesture. Among operations for controlling appliances, switching on/off are the most basic ones. Thus, in this paper, we aim to detect a hand waving gesture which can be used for switching on/off appliances.

Gesture interfaces need to capture human body motions to detect the target gesture. Human body motions can be obtained by either contact devices or non-contact devices. In the case of contact devices, users need to wear them. An example of a contact device is a ring with multiple sensors (Jing et al., 2012). On the other hand, in the case of non-contact devices, we do not need to wear them. Cameras such as RGB-D cameras (Mahbub et al., 2013) and visible-light cameras (Lee and Kim, 1999) are mainly used as non-contact devices. Therefore, we can use a gesture interface with our own body as long as we are in the shooting range of a non-contact device.

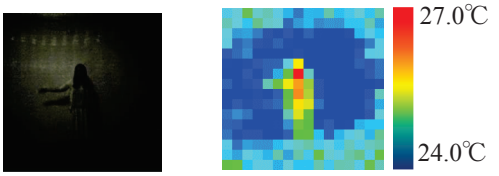


(a) Visible-light image (b) Low-resolution thermal image

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

Practically, visible-light cameras have a drawback. We cannot always set cameras anywhere and/or anytime because they involve privacy issue and they do not work well in the dark. As for the privacy issue, if an user is always observed by a camera in his/her private area, the user may feel uncomfortable. As we can see from the captured image shown in Figure 1 (a), we can easily identify the individual from the image and what the user is doing, so it may not be acceptable. On the other hand, an example of a

Toriyama, C., Kawanishi, Y., Takahashi, T., Deguchi, D., Ide, I., Murase, H., Aizawa, T. and Kawade, M.
Hand Waving Gesture Detection using a Far-infrared Sensor Array with Thermo-spatial Region of Interest.
DOI: 10.5220/0005718105450551
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 545-551
ISBN: 978-989-758-175-5



(a) Visible-light image (b) Low-resolution thermal image

Figure 2: Examples of an output at night-time.

Figure 3: A 16 × 16 far-infrared sensor array.

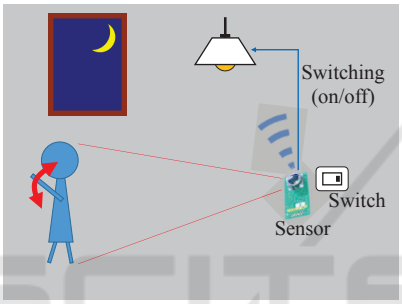


Figure 4: Example of the application of the proposed method.

captured image at night-time is shown in Figure 2 (a). Although it is difficult to identify the user, now it becomes also difficult to observe the gesture.

To avoid these problems, a far-infrared sensor array (Ohira et al., 2011) can be a good choice. A 16 × 16 far-infrared sensor array is shown in Figure 3. Although the sensor captures noisy image, it captures the spatial distribution of temperature as a thermal image by detecting far-infrared waves emitted from heat sources. Therefore, it works even at night-time without any light source. Examples of its output are shown in Figure 1 (b) and Figure 2 (b). As we can see, they represent the spatial distribution of temperature as a low-resolution image. Since the images only show the rough shape of a body with no texture, we cannot easily identify the individual. Therefore, as illustrated in Figure 4, we can use the sensor to switch on/off a room light by waving our hand toward it without privacy concerns.

In order to realize the gesture interface, we need to detect a gesture from an image sequence. In this paper, we propose a method for hand waving gesture detection using a far-infrared sensor array. The detection process segments the image sequence and classify each subsequence whether it is a hand waving gesture or not. To accurately classify the gesture using the low resolution and noisy sensor, the gesture and background clutter should be distinguished. Therefore, as contributions, we introduce the following concepts:

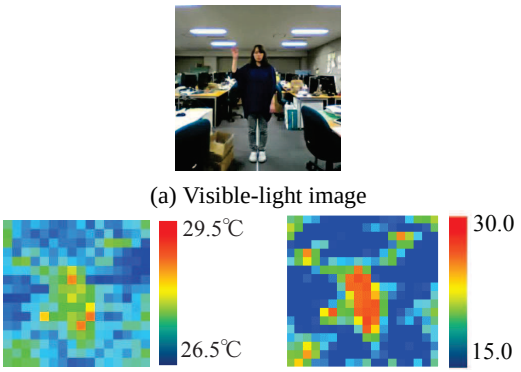
- "Thermal Region of Interest (TRoI)" that focuses on the range of human body temperature to emphasize the human body.
- "Spatial Region of Interest (SRoI)" that focuses on the region with target motion to eliminate the others.

2 RELATED WORKS

Hosono *et al.* proposed a method for human tracking using a far-infrared sensor array (Hosono et al., 2015). This method tracks a human in low-resolution thermal images, but it does not target gesture recognition.

There are some researches related to vision-based gesture recognition using a visible-light camera. Fujii *et al.* proposed a method that focused on the change of arm directions during a gesture (Fujii et al., 2014). They extrapolated arm directions from joint points of the human body captured by Microsoft's Kinect sensor (Shotton et al., 2013). Mohamed *et al.* proposed a method of tracking a hand trajectory (Alsheakhali et al., 2011). This method detected its user's hands based on skin tone and motion information. However, in case of far-infrared sensor arrays, it is difficult to detect the joint position clearly because the captured thermal image is in very low-resolution. In addition, the far-infrared sensor array cannot capture color information. Thus, it is difficult to apply these methods directly to images captured from the far-infrared sensor array.

Takahashi *et al.* and Cutler *et al.* proposed a method of detecting a periodic motion from low-resolution images by the Discrete Fourier Transform (Takahashi et al., 2010) (Cutler and Davis, 1998). The former method applies it to a time series of intensity values and the latter method applies the segmented object's self-similarity. However, in case of far-infrared sensor arrays, it is difficult to detect the periodic gesture because noise appears in the captured images. Therefore, these methods are not suitable for being applied to far-infrared sensor arrays.



(a) Visible-light image

(b) Thermal image before weighting

(c) Thermal image after weighting

Figure 5: Example of the effect of focusing on TRoI.

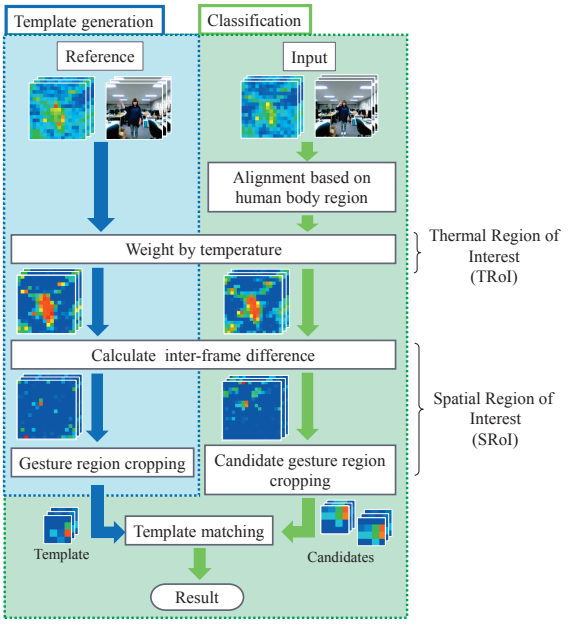


Figure 6: Process flow of the proposed method.

3 HAND WAVING DETECTION WITH THERMO-SPATIAL REGIONS OF INTEREST

There are two difficulties to detect a hand waving gesture with a far-infrared sensor array.

One is to separate the hand waving gesture from noisy images. This noise is caused by heat sources in the background except for the user's body. An example of the output image when there are heat sources in the background is shown in Figure 5 (b). To emphasize the human body, we introduce "Thermal Region of Interest (TRoI)". The TRoI emphasizes the difference between the user's body and the background. Pixel values are weighted according to the user's body temperature.

The other is to localize the motion region from the human body region in an image of the far-infrared sensor array because it only captures the rough shape of the body. To localize the motion region that includes an arm for hand waving detection, we introduce "Spatial Region of Interest (SRoI)". The SRoI restricts the spatial region for detection.

3.1 Process Flow of the Proposed Method

As a reference sequence, we assume that an image sequence of a hand waving gesture of an user in given. The proposed method detects the hand waving gesture by matching the reference sequence with an input image subsequence segmented by the temporal window scan. The input image subsequence is classified according to whether it is a hand waving gesture or not. The process flow is illustrated in Figure 6. It consists of template generation and classification.

In the template generation, to emphasize the human body, pixel values of the reference image sequence are weighted according to the TRoI. In addition, to eliminate other parts of the body than the arm, the gesture region is cropped as a template according to the SRoI.

In the classification, a template-matching-based detection process with a Dynamic Time Warping (DTW)-based distance metric is performed on each input sequence. DTW is performed between the reference sequence and a subsequence cut out from the input sequence. If the distance is smaller than a threshold, the process classifies the sequence as a hand waving gesture. Each process is described in detail in the following sections.

3.2 Template Generation

3.2.1 Thermal Region of Interest (TRoI)

To emphasize human body regions, the proposed method weights pixel values of the reference image sequence according to the human body temperature. The weighted value is defined as follows:

$$R'^{(j)}_{\boldsymbol{x}} = \begin{cases} \exp\left(-\frac{|R^{(j)}_{\boldsymbol{x}} - T_r|^2}{2}\right) R^{(j)}_{\boldsymbol{x}} & (R^{(j)}_{\boldsymbol{x}} < T_r) \\ R^{(j)}_{\boldsymbol{x}} & (\text{otherwise}) \end{cases} \quad (1)$$

where $R_{\boldsymbol{x}}^{(j)}$ is the value of the target pixel $\boldsymbol{x}$ in the j-th frame. T_r is the estimated value of the human body temperature, which is calculated as the upper quartile of pixel values sorted in ascending order in the human body region in the first frame. Here, the human body region in an image is bounded by a rectangle, which is annotated manually. This emphasizes the human body region while suppressing influence of other heat sources except for the human body.

Figure 5 (c) shows an example of an image weighted by Equation (1). We can see that the difference of the human body and the background becomes clearer.

3.2.2 Spatial Region of Interest (SRoI)

To localize the motion region, the proposed method extracts the inter-frame difference and crops the region to be used for the gesture classification. Let R'' denote the inter-frame difference value of two successive frames in the reference sequence. It can be written as follows:

$$R''^{(j)}_{\boldsymbol{x}} = R'^{(j)}_{\boldsymbol{x}} - R'^{(j-1)}_{\boldsymbol{x}} \tag{2}$$

The gesture region in the difference images are cropped as a template for gesture detection.

3.3 Classification

3.3.1 Normalization of Human Body Region

The size of a human body in an image captured by the far-infrared sensor array vary depending on the relative distance between the human body and the sensor. So we need to normalize the human body size of an input with the reference. When the human body size in an input image is smaller than the human body size in the reference image sequence, we expand it with bicubic image interpolation. On the other hand, when the input human body size is larger than the reference human body size, to suppress aliasing, it is expanded first and then shrunk to the same size as the reference human body size by downsampling.

3.3.2 Template Matching

The proposed method detects a hand waving gesture as follows:

1. Crop candidate gesture regions in the difference images from the input sequence.
2. Calculate the distance between the template and each of the candidate gesture regions.
3. Classify each candidate as a hand waving gesture if the distance is smaller than a given threshold.

An input sequence is preprocessed as same as the reference sequence, that is, it is emphasized by the human body temperature and inter-frame difference is calculated. Candidate gesture regions are cropped from the input sequence by the same process applied to the reference sequence. However, the exact location of the gesture region in the input sequence is not known. Therefore, several candidate gesture regions are cropped from the input sequence. The cropping position is determined based on the relative position between the human body region and the gesture regions in the reference sequence.

Here, the distance is calculated with a DTW-based distance metric. The distance $D(R'', I'')$ is defined as follows:

$$D(R'', I'') = \min_{c} \frac{g_c(R''^{(J)}, I''^{(K)})}{L} \tag{3}$$

where J and K denote the length of the reference and the input sequences respectively, $g_c(R''^{(J)}, I''^{(K)})$ denotes the distance between the template and the candidate c, and L denotes the path length based on the result of the DTW. We define $g_c(R''^{(j)}, I''^{(k)})$ as follows:

$$\begin{aligned} &g_c(R''^{(j)}, I''^{(k)}) \\ &= \min \begin{cases} g_c(R''^{(j-1)}, I''^{(k)}) + d(R''^{(j)}, I''^{(k)}) \\ g_c(R''^{(j-1)}, I''^{(k-1)}) + d(R''^{(j)}, I''^{(k)}) \\ g_c(R''^{(j)}, I''^{(k-1)}) + d(R''^{(j)}, I''^{(k)}) \end{cases} \end{aligned} \tag{4}$$

The distance between frames $R''^{(j)}$ and $I''^{(k)}$ is defined as follows:

$$\begin{aligned} &d(R''^{(j)}, I''^{(k)}) \\ &= \sum_{n=1}^{N} ||R''^{(j)}_{\boldsymbol{x}_n} - I''^{(k)}_{\boldsymbol{x}'_n}||^2 \\ &= \sum_{n=1}^{N} (||R''^{(j)}_{\boldsymbol{x}_n}||^2 - 2R''^{(j)}_{\boldsymbol{x}_n} I''^{(k)}_{\boldsymbol{x}'_n} + ||I''^{(k)}_{\boldsymbol{x}'_n}||^2) \end{aligned} \tag{5}$$

where N is the number of pixels in the gesture region. To make it robust to temperature variations of the human body and the background depending on capturing environments, pixel values of these images are normalized so that the average becomes 0 and the variance becomes 1. Therefore, Equation (5) is simplified as follows:

$$d(R''^{(j)}, I''^{(k)}) = \sum_{n=1}^{N} 2(1 - S(R''^{(j)}_{\boldsymbol{x}_n}, I''^{(k)}_{\boldsymbol{x}'_n})) \tag{6}$$

Table 1: 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, Role over, Pick up
Pose	Standing, Sitting	Standing, Sitting	Lying, Sitting, Relaxing
# of persons	6	5	3
# of datasets	11	13	8

where $S(R''^{(j)}_{\boldsymbol{x}_n}, I''^{(k)}_{\boldsymbol{x}'_n})$ is a Normalized Cross-Correlation (NCC) function defined as follows:

$$S(R''^{(j)}_{\boldsymbol{x}_n}, I''^{(k)}_{\boldsymbol{x}'_n}) = \frac{\sum_{n=1}^{N} R''^{(j)}_{\boldsymbol{x}_n} I''^{(k)}_{\boldsymbol{x}'_n}}{\sqrt{\sum_{n=1}^{N} (R''^{(j)}_{\boldsymbol{x}_n})^2 \times \sum_{n=1}^{N} (I''^{(k)}_{\boldsymbol{x}'_n})^2}} \quad (7)$$

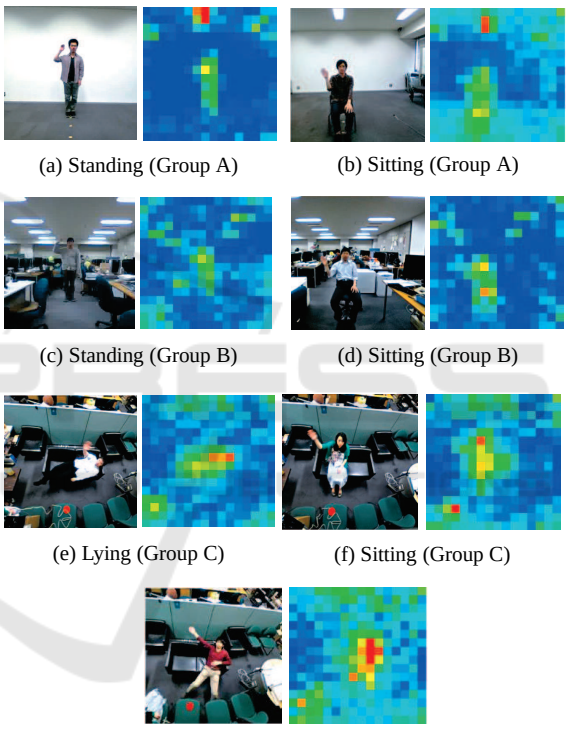
4 EXPERIMENT

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

4.1 Datasets

The target gesture in this experiment was “wave the right hand twice during approximately 4 seconds”. We collected 32 datasets, where a dataset consisted of a reference sequence and a number of input sequences. The reference sequence captured a subject performing the hand waving gesture. The input sequences were sampled from the video capturing a subject performing the gesture. Environment, subject, and his/her pose were varied and fixed among each dataset. They were divided into three groups by environments.

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



(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)

Figure 7: Examples of images from the datasets.

The details of the capturing conditions are summarized in Table 1, and examples from the datasets are shown in Figure 7.

4.2 Experimental Condition

In the experiment, we evaluated the performance of the gesture classification. To analyze the effectiveness of the thermal and spatial regions of interest, we compared the proposed method with its two Variations, 1, 2, and the BaseLine. To confirm the effectiveness of the proposed method, we compared it with

Table 2: Experimental results (Maximum classification rate).

	TRoI	SRoI	Data group A	B	C	All
Proposed method	✓	✓	0.79	**0.79**	**0.91**	**0.82**
Variation 1	-	✓	**0.82**	0.77	0.87	0.81
Variation 2	✓	-	0.77	0.70	0.84	0.76
BaseLine	-	-	0.79	0.74	0.87	0.79
Comparative method (DFT)			0.62	0.59	0.59	0.60

a Comparative method (Takahashi et al., 2010). The conditions of these methods are as follows;

- Proposed method: Using both SRoI and TRoI.
- Variation 1: Using only SRoI
- Variation 2: Using only TRoI
- BaseLine: Using neither SRoI nor TRoI
- Comparative method: Using Discrete Fourier Transform (DFT) (Takahashi et al., 2010)

Instead of using the SRoI, Variation 2 and BaseLine used the region including the whole body region for the matching. We used the maximum classification rate C as a criterion to evaluate each method, defined as follows:

$$C = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP}} \quad (8)$$

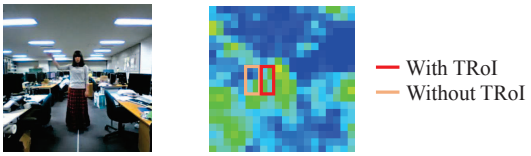
where #TP represents the number of true positives and #FP represents the number of false positives.

4.3 Results and Discussion

The results are shown in Table 2. It shows the average of the maximum classification rates for each group. As shown in this table, the proposed method achieved the best performance in almost all cases.

The SRoI worked effectively in Groups A and B. Although the input images which were captured by the far-infrared sensor array were noisy due to the air flow, the proposed method could reduce the noise by the SRoI. We confirmed that the distance was smaller when the proposed method successfully classified the target gesture. Therefore, we can say that it became easier to separate the target gesture with the others.

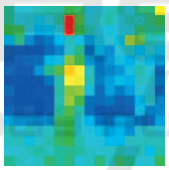
The TRoI was effective by combining it with the SRoI for all groups. This helped the proposed method determine gesture regions accurately. It seems that the influence of other heat sources was reduced and the human body region was emphasized by focusing on the TRoI. Example of images that the TRoI worked effectively is shown in Figure 8. In some cases, gesture regions were localized incorrectly where it was similar to the human body temperature. On the other hand, the TRoI made it possible to localize the gesture area correctly because it increased the influence



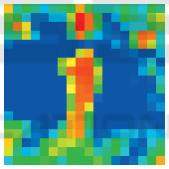
(a) Visible-light image (b) Difference in the position of the gesture area.

Figure 8: Example of images that the TRoI was effective.

(a)Visible-light image



(b) Thermal image before weighting



(c) Thermal image after weighting

Figure 9: Example of images that the TRoI was not effective.

of temperature changes around the human body temperature. The TRoI played a role that helps the SRoI.

Example of images that the TRoI was not effective are shown in Figure 9. Although we can see the arm in Figure 9 (a), after focusing on the TRoI, it became difficult to find the arm in Figure 9 (b). It seems that the arm region was weakened by the TRoI because the temperature difference between the arm and the body was larger than that for other subjects. We can say that the classification failed because the temperature of the arm region became similar to the background temperature.

The comparative method focused on the periodicity of the time series of the pixel value. However, the input images which were captured by the far-infrared sensor array were noisy. Therefore, the accuracy of the comparative method decreased because

it was easily affected by noise. Meanwhile, the proposed method was able to classify the hand waving gesture even if noise was included in the output images.

5 CONCLUSION

In this paper, we proposed a hand waving gesture detection method using a far-infrared sensor array. The proposed method matched a reference sequence captured beforehand with an input sequence. We reduced the influence of other heat sources by the TRoI. We also reduced noise by the SRoI. Experimental results showed that the SRoI was effective in the reduction of noise. Furthermore, the TRoI was effective by combining it with the SRoI.

As future work, we will modify the TRoI to further improve the classification performance of the proposed method. We will also consider a method to improve the estimation of the human body temperature used in the TRoI. In addition, we need to track humans for gesture recognition. We expect to realize a practical gesture recognition system by combining the proposed method with a tracking method such as (Hosono et al., 2015).

ACKNOWLEDGEMENTS

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

REFERENCES

Alsheakhali, M., Skaik, A., Aldahdouh, M., and Alhelou, M. (2011). Hand gesture recognition system. In *Proc. Int. Conf. on Information & Communication Systems 2011*, pages 132–136.

Cutler, R. and Davis, L. (1998). View-based detection and analysis of periodic motion. In *Proc. 14th Int. Conf. on Pattern Recognition*, volume 1, pages 495–500.

Fujii, T., Lee, J. H., and Okamoto, S. (2014). Gesture recognition system for human-robot interaction and its application to robotic service task. In *Proc. Int. MultiConf. of Engineers and Computer Scientists 2014*, volume 1, pages 63–68.

Hosono, T., Takahashi, T., Deguchi, D., Ide, I., Murase, H., Aizawa, T., and Kawade, M. (2015). Human tracking using a far-infrared sensor array and a thermo-spatial sensitive histogram. In Jawahar, C. and Shan, S., editors, *Computer Vision – ACCV 2014 Workshops*, volume 9009 of *Lecture Notes in Computer Science*, pages 262–274. Springer International Publishing.

Jing, L., Zhou, Y., Cheng, Z., and Huang, T. (2012). Magic ring: A finger-worn device for multiple appliances control using static finger gestures. *Sensors*, 12(5):5775–5790.

Lee, H.-K. and Kim, J. H. (1999). An HMM-based threshold model approach for gesture recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(10):961–973.

Mahbub, U., Imtiaz, H., Roy, T., Rahman, M. S., and Ahad, M. R. (2013). A template matching approach of one-shot-learning gesture recognition. *Pattern Recognition Letters*, 34(15):1780–1788.

Ohira, M., Koyama, Y., Aita, F., Sasaki, S., Oba, M., Takahata, T., Shimoyama, I., and Kimata, M. (2011). Micro mirror arrays for improved sensitivity of thermopile infrared sensors. In *Proc. 24th IEEE Int. Conf. on Micro Electro Mechanical Systems*, pages 708–711.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., and Blake, A. (2013). Efficient human pose estimation from single depth images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840.

Takahashi, M., Irie, K., Terabayashi, K., and Umeda, K. (2010). Gesture recognition based on the detection of periodic motion. In *Proc. Int. Symposium on Optomechatronic Technologies*, pages 1–6.