Action Recognition from Extremely Low-Resolution Thermal Image Sequence

Takayuki Kawashima, Yasutomo Kawanishi, Ichiro Ide, Hiroshi Murase
Graduate School of Information Science, Nagoya University, Aichi, Japan
kawashimat@murase.m.is.nagoya-u.ac.jp, {kawanishi, ide, murase}@is.nagoya-u.ac.jp

Daisuke Deguchi
Information Strategy Office, Nagoya University, Aichi, Japan
ddeguchi@nagoya-u.jp

Tomoyoshi Aizawa, Masato Kawade
Corporate R&D, OMRON Corporation, Kyoto, Japan
{aizawa, kawade}@ari.ncl.omron.co.jp

Abstract

*This paper proposes a Deep Learning-based action recognition method from an extremely low-resolution thermal image sequence. The method recognizes daily actions by humans (*e.g. *walking, sitting down, standing up, etc.) and abnormal actions (*e.g. *falling down) without privacy concerns. While privacy concerns can be ignored, it is difficult to compute feature points and to obtain a clear edge of the human body from an extremely low-resolution thermal image. To address these problems, this paper proposes a Deep Learning-based action recognition method that combines convolution layers and an LSTM layer for learning spatio-temporal representation, whose inputs are the thermal images and their frame differences cropped by the gravity center of human regions. The effectiveness of the proposed method was confirmed through experiments.*

1. Introduction

Since the number of elderly people living alone is increasing in recent years, it is becoming necessary to provide them with support. Medical professionals believe that analyzing the human behavior and finding changes in the Activities of Daily Living (ADL [18]) are important for detecting physical and mental health problems before they become critical. Since it is difficult for nurses or families to monitor elderly people constantly, automatic monitoring systems are expected to be developed for recognizing their daily activities. For an elderly person who lives alone, it is also required to find his/her abnormal actions such as falling down.

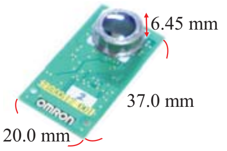
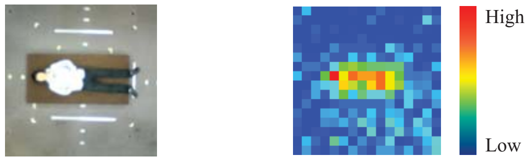


Figure 1. 16×16 far-infrared sensor array.



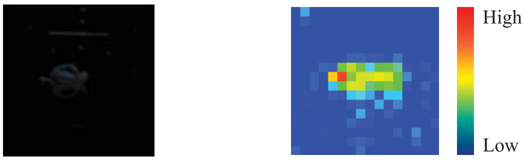
(a) Visible-light image (b) Extremely low-resolution thermal image

Figure 2. Example of images captured by a visible-light camera and a 16×16 far-infrared sensor array.

To realize such monitoring systems, we are considering the usage of a far-infrared sensor array [13]. Figure 1 shows a 16×16 far-infrared sensor array. It is a very low price sensor integrating multiple thermopile infrared sensors into a grid. Although the output image of the far-infrared sensor array is quite noisy due to its mechanism, it can capture spatial distribution of temperature as a thermal image by detecting far-infrared waves emitted from heat sources. It also works even at night-time without any light source. Examples of its output are shown in Figure 2 (b) and Figure 3 (b). As we can see, the images are in extremely low-resolution (16×16 pixels), so individuals cannot be easily identified. Therefore, it is possible to use the sensor for monitoring a person constantly day and night without privacy concerns.

While privacy concerns can be ignored, visual information obtained from extremely low-resolution images is lim-

978-1-5386-2939-0/17/$31.00 ©2017 IEEE



(a) Visible-light image (b) Extremely low-resolution thermal image

Figure 3. Example of images captured at night-time.

ited. Therefore, it is difficult to compute feature points and to obtain a clear edge of the human body from it. Moreover, the pixel values will be easily affected by factors such as the motion of a person and the distance between the sensor and the human body. Therefore, most conventional action recognition methods using a visible-light camera are not suitable for being applied to extremely low-resolution thermal image sequences.

To overcome the above-mentioned problems, this paper proposes a recognition method for both daily human actions (*e.g.* walking, sitting down, standing up, etc.) and abnormal actions (*e.g.* falling down) from extremely low-resolution thermal image sequences obtained from a far-infrared sensor array mounted on the ceiling. Since abnormal actions will occur anywhere in the room, the method needs to recognize actions regardless of positions. In order to recognize actions from an extremely low-resolution thermal image sequence, focusing on shapes of human body and its temporal variation should be important. Therefore, this paper tries to overcome these problems by a CNN+LSTM (Long Short Term Memory [7]) approach extending the Long-term Recurrent Convolutional Network (LRCN) model [4].

In the rest of this paper, related works are described in Section 2, characteristics of extremely low-resolution thermal image sequences are described in Section 3, details of the proposed method are presented in Section 4, and evaluation results are described in Section 5. Finally, Section 6 concludes this paper.

2. Related works

There are only few works on action recognition using a far-infrared sensor array. Toriyama *et al.* [16] proposed "Thermal Region of Interest" and "Spatial Region of Interest" to recognize a hand waving gesture, but they did not consider other body actions. Hevesi *et al.* [5] referred to the change in sensor output values to recognize household activities. However, they assumed only activities related to specific areas in the room. This approach that considered positional information of the activities is effective but could not recognize position-independent actions such as falling down.

On the other hand, there are many works related to vision-based action recognition using a visible-light camera [1]. Especially, Deep Learning-based methods have been proposed and yielded high performance in recent years. In some works [10, 15], CNN is used in order to learn spatial features of actions. For example, Ji *et al.* [8] proposed 3D-CNN to learn spatio-temporal features over an image sequence directly. However, methods using a visible-light camera require a relatively high-resolution image sequence as input, and optical flow or trajectory features [17] are usually used. Since it is difficult to compute feature points from extremely low-resolution images captured by a far-infrared sensor array, these methods cannot be applied to the thermal image sequences. Donahue *et al.* [4] introduced LRCN model that combines CNN and LSTM to learn long-term dependencies. This model is end-to-end trainable and pre-trained on a larger object recognition dataset. However, if images captured by a ceiling sensor are directly input to the network, there is a possibility that the network learns the position dependency rather than the motion feature. This would be noticeable when the number of learning data were small. Since it is impossible to repeat the pooling process or apply a large-size kernel to an extremely low-resolution image, the conventional CNN architectures are not suitable for our purpose.

There are several works that consider privacy through the use of extremely low-resolution visible-light videos. Ryoo *et al.* [14] introduced the concept of "Inverse super resolution" and generated a set of low-resolution images from a high-resolution image for effective learning of CNN. Chen *et al.* [2] proposed ideas of using two-stream CNN and sharing filters between extremely low-resolution and high-resolution images during training for effective learning. However, since both methods require high-resolution images in training, we cannot apply these methods to low-resolution images from a far-infrared sensor array. Dai *et al.* [3] use an l_1 nearest-neighbor classifier to discriminate action sequences. However, a large amount of training data is required to obtain accurate results by the nearest-neighbor method, or else the shift of one pixel greatly influences the result of template matching in extremely low-resolution images.

3. Characteristics of extremely low-resolution thermal image sequences

Extremely low-resolution thermal image sequences obtained from a far-infrared sensor array have several characteristics (*c.f.* Figure 4).

- The edge of the human body does not appear clearly.
- The motion of a person changes the pixel values of both the human body and its surrounding region.
- When the distance between the sensor and the human body changes, the pixel values also change.
- A pixel value changes depending on the occupancy

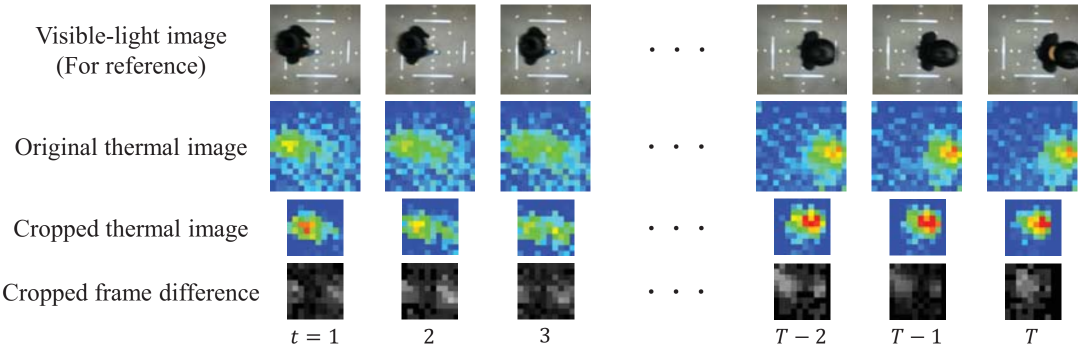


Figure 4. Example of a thermal image sequence.

area ratio of the human body in the observation range of a thermopile infrared sensor.

Although it is difficult to define and calculate hand-crafted features considering these characteristics, it is important to focus on the rough shape of the human body and its temporal variation to recognize actions. We also consider that the motion of a person can be represented by frame difference.

4. Action recognition from an extremely low-resolution thermal image sequence

In this paper, we propose a Deep Learning-based action recognition method from an extremely low-resolution thermal image sequence. The following points are considered for this:

- To cope with extremely low-resolution thermal image sequences, the network structure is modified to a specialized CNN for extremely low-resolution inputs.
- To learn spatio-temporal representations, convolution layers for extracting appearance features and an LSTM layer for learning temporal dependencies are combined.
- To extract position invariant features, input images are cropped by the gravity center of human regions, and the network input are extended by adding frame differences as an additional channel to extract motion features from extremely low-resolution thermal images.

Details are described in the following sections.

4.1. Pre-processing of an extremely low-resolution thermal image sequence

To recognize actions independent of the position they occurred, the proposed method firstly aligns images by the gravity center of human regions. Then, the pixel values of images are normalized to make it robust against temperature change of the environment.

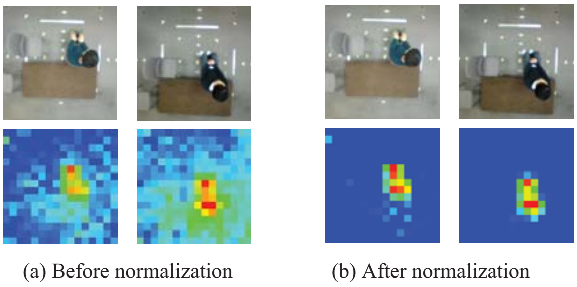

(a) Before normalization
(b) After normalization

Figure 5. Effect of normalization.

First, for each image sequence, a sequence of binary silhouettes are extracted by Gaussian Mixture Model (GMM)-based background subtraction [9]. Then, the gravity center of the silhouette is calculated from each image of the binary silhouette sequence. They are considered as the positions of the human body region in each image.

Next, each thermal image sequence $\{I_t\}_{t=0}^{T}$ is normalized. Here, T is the length of a sequence (different depending on sequence), and $I_t(i,j)$ represents the pixel value at pixel (i,j) of image I_t as:

$$\widehat{I}_t(i,j) = \frac{I_t(i,j) - m}{M - m}, \tag{1}$$

where M and m are the maximum and the minimum pixel values over an entire image of each sequence, respectively. Figure 5 shows examples of the effect of normalization. The normalization gives robustness against environmental variations, and makes a human body and the environment more clearly separable. Frame difference images $\{D_t\}_{t=1}^{T}$ are calculated from the normalized image sequence as follows:

$$D_t(i,j) = |\widehat{I}_t(i,j) - \widehat{I}_{t-1}(i,j)|, \tag{2}$$

Finally, a fixed-size region around the gravity center of a human region is cropped from each image in a sequence. In more detail, an image of $R \times R$ pixels around the gravity center is cropped. Similarly, an image of $R \times R$ pixels is cropped from each frame difference in the sequence. Here,

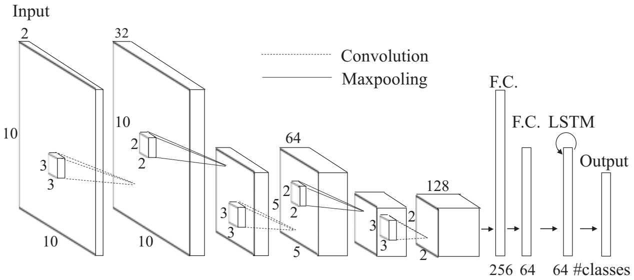


Figure 6. Proposed network architecture. The kernel size and the number of channels or the units of each layer are shown. Note that in each convolution layer, padding is applied.

$R = 10$ is used so that the entire human body region should be included in each image. This cropping can reduce the influence of cluttered background. These cropped thermal and frame difference image sequences are input to the network. Figure 4 shows examples of each sequence.

4.2. Deep learning-based action recognition

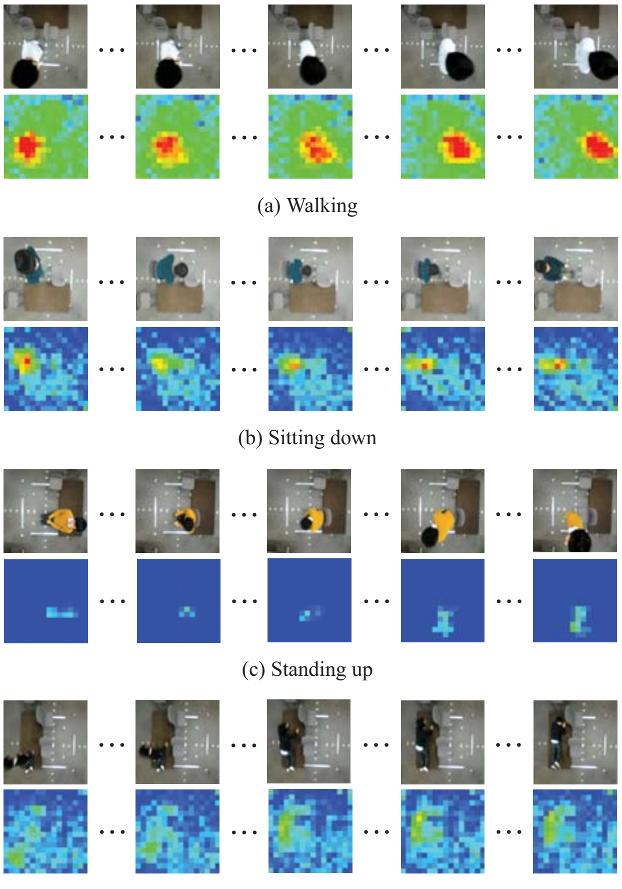
In order to learn spatio-temporal representations, convolution layers and an LSTM layer are combined. As explained section 2, since the conventional CNN architectures are not suitable for extremely low-resolution images, we propose a redesigned network architecture using three convolution layers, two fully-connected layers, and an LSTM layer as show in Figure 6. In each convolution layer, the stride of the kernels is 1. Moreover, in convolution layers 1 and 2, maxpooling is applied with a stride of 2. Rectified Linear Units (ReLU) [12] is used as the activation function for the convolution and fully-connected layers. Also, sigmoid function is used as the activation function for the LSTM layer, and a softmax function is used for the output layer.

As shown in Figure 6, there are two input channels for the network. A thermal and a frame difference image pair is input to the network at each time step, and the network also predicts an action class at each time step. The output of the last frame is adopted as the final classification result, since it takes into account the temporal variations over all frames of a sequence.

Cross-entropy loss function is used to train the network. Kernels, weights, and biases are initialized with random values. Adam [11] is used as the optimization algorithm. The dropout [6] with a ratio of 0.2 is also applied to the two fully-connected layers for enhancing the generalization capability.

5. Experiment and discussions

To confirm the effectiveness of the proposed method, an experiment was conducted using actual extremely low-resolution thermal image sequences. The sequences were captured using a 16×16 far-infrared sensor array (Thermal sensor D6T-1616L by OMRON Corp.). The frame rate was 10 fps. The dataset and the experimental conditions are described below, followed by report and discussion on the results.



(a) Walking

(b) Sitting down

(c) Standing up

(d) Falling down

Figure 7. Examples of human actions from the dataset.

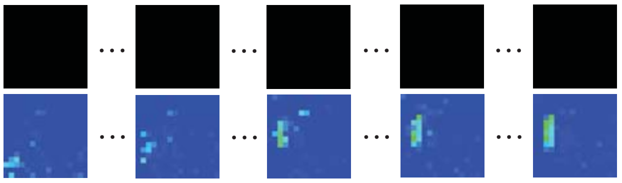


Figure 8. Example of "Falling down" in the dark.

5.1. Extremely low-resolution thermal image sequence dataset

Figure 7 shows examples of human actions from the dataset[1]. It was captured by the sensor mounted on the ceil-

[1] This dataset will be opened to the public.

	Overall	No action	Walking	Sitting down	Standing up	Falling down
Proposed	**91.07**%	94.35%	97.50%	74.72%	89.17%	93.06%
Thermal only	85.75%	92.78%	83.89%	74.44%	75.28%	88.33%
Diff. only	82.34%	98.80%	96.11%	42.22%	68.06%	73.61%
Baseline	75.95%	87.87%	74.72%	54.44%	68.89%	70.00%
Dai *et al*. [3]	58.25%	82.96%	42.78%	36.11%	27.50%	52.50%

Table 1. Experimental results (CCR).

ing (220 cm above the floor) of a room. The dataset includes four actions: "Walking", "Sitting down", and "Standing up" as daily actions, and "Falling down" (while subjects perform Walking or Standing up) as an abnormal action. In addition, in order to evaluate whether the proposed method can learn the temporal change of a motion, the dataset includes "No action" (*i.e.* standing, sitting, or lying) class. These action samples were collected from nine subjects, each of whom repeated each action forty times (For "No action", standing, sitting, and lying were performed forty times each). Overall, the dataset contained 2,520 sequences. Half of them were performed in daytime and the others at night in the dark. Figure 8 shows examples of "Falling down" in the dark. To prevent from making an action dependent from the position that it occurred, these actions were performed in various directions and at various positions in the room. The number of frames included in each sequence was between 10 and 50.

5.2. Experimental condition

In order to analyze the effectiveness of combining thermal images and frame differences, the proposed method was compared with the baseline and two variations of the proposed method ("Thermal only" and "Diff. only"). Furthermore, the proposed method was compared with the method by Dai *et al*. [3] that uses a nearest-neighbor classifier. In the experiment, the same background subtraction method was used for all methods for fair comparison. The conditions of the methods are summarized as follows:

- Proposed: Input of the network is both the cropped thermal and the frame difference image sequences.
- Thermal only: Input is only the cropped thermal image sequence (without frame difference).
- Diff. only: Input is only the cropped frame difference sequence (without thermal image).
- Baseline: Input is the entire 16×16 thermal image sequence (without cropping nor frame difference).
- Dai *et al*. [3]: Multi-templates and nearest-neighbor.

We performed leave-one-person-out cross-validation and computed the Correct Classification Rate (CCR) to measure the performance. During the network training, data augmentation was performed by shifting ±1 pixel in each of the x, y axis directions.

5.3. Results and discussions

Table 1 shows the experimental results. As shown in this table, the proposed method achieved significantly good overall performance; more than 30% higher than Dai *et al*. [3]. We consider that the proposed network could learn both appearance and motion features for discriminating actions.

By comparing "Thermal only" and "Diff. only", it can be observed that "Thermal only" tends to show higher performance on "Sitting down," "Standing up", and "Falling down". We consider that using thermal images is effective for actions which can be discriminated by changes in appearance or posture.

It can also be observed that "Diff. only" tends to show higher performance on "No action" and "Walking". We consider that this is because using frame differences is effective for actions which can be discriminated by motion features.

Here, "Proposed" is superior to both "Thermal only" and "Diff. only" methods for most action classes. Therefore, by combining thermal and frame difference images, the network could learn features that can not be obtained by using each feature individually.

5.4. Effectiveness of the combination of CNN and LSTM

In order to analyze the effectiveness of combining convolution layers and an LSTM layer, three network architectures were compared. The first architecture is a network where the LSTM layer of the proposed network is replaced by a fully-connected layer. The number of units and the activation functions are the same. In this architecture, T images in a sequence are individually classified and scores of the actions are calculated. Then, the final classification is done by averaging the scores. The second architecture is a network where the convolution layers of the proposed network are replaced by fully-connected layers. The number of units of the fully-connected layers is set all to 256. In this architecture, an image of 10×10 pixels for two channels is

Network architecture	CCR
CNN (without LSTM)	67.29%
LSTM (without CNN)	46.94%
CNN + LSTM (Proposed)	**91.07**%

Table 2. Results from each network architecture.

input to the network as a 200-dimensional vector. The third architecture is the proposed network architecture as shown in Figure 6.

As shown in Table 2, the proposed network architecture achieved the best performance. Furthermore, the performance was improved by combining convolution layers and an LSTM layer compared with the case without using each layer. This is because the convolution layers could learn appearance features, and the LSTM layer could learn temporal variations throughout the sequence. Therefore, the network could learn spatio-temporal representation by combining these layers for discriminating actions.

6. Conclusion

This paper proposed an action recognition method for extremely low-resolution thermal image sequences. To learn the spatio-temporal representation of a sequence, we proposed a Deep Learning-based method that combined convolution layers, whose inputs were the thermal images and their frame differences cropped by the gravity center of human regions, and an LSTM layer. Experimental results showed the effectiveness of the proposed approach.

As future works, we will consider a more suitable network architecture to improve the performance. We will also consider additional action classes that are similar to “Falling down” (*e.g.*, “Sitting up”, and “Lying down”). Furthermore, we expect to realize a practical action detection system by applying the the proposed method.

Acknowledgements

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research. We would like to thank the subjects who cooperated in creating the dataset.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, 2011.

[2] J. Chen, J. Wu, J. Konrad, and P. Ishwar. Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In *Proc. 2017 IEEE Winter Conf. Applicat. Comput. Vision*, pages 139–147, 2017.

[3] J. Dai, J. Wu, B. Saghafi, J. Konrad, and P. Ishwar. Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras. In *Proc. 2015 IEEE Conf. Comput. Vision and Patt. Recog. Workshops*, pages 68–76, 2015.

[4] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. 2015 IEEE Conf. Comput. Vision and Patt. Recog.*, pages 2625–2634, 2015.

[5] P. Hevesi, S. Wille, G. Pirkl, N. Wehn, and P. Lukowicz. Monitoring household activities and user location with a cheap, unobtrusive thermal sensor array. In *Proc. 2014 ACM Int. Joint Conf. Pervasive and Ubiquitous Comput.*, pages 141–145, 2014.

[6] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[7] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[8] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE Trans. Patt. Anal. Mach. Intell.*, 35(1):221–231, 2013.

[9] P. Kaewtrakulpong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *Video-based surveillance systems*, pages 135–144. Springer, 2002.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. 2014 IEEE Conf. Comput. Vision and Patt. Recog.*, pages 1725–1732, 2014.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proc. 27th Int. Conf. Mach. Learning*, pages 807–814, 2010.

[13] 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. 2011 IEEE Int. Conf. Micro Electro Mech. Syst.*, pages 708–711, 2011.

[14] M. S. Ryoo, B. Rothrock, and C. Fleming. Privacy-preserving egocentric activity recognition from extreme low resolution. *arXiv preprint arXiv:1604.03196*, 2016.

[15] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems 27*, pages 568–576, 2014.

[16] 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 Int. Conf. Comput. Vision Theory and Applicat.*, pages 545–551, 2016.

[17] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. 2013 IEEE Int. Conf. Comput. Vision*, pages 3551–3558, 2013.

[18] J. M. Wiener, R. J. Hanley, R. Clark, and J. F. Van Nostrand. Measuring the activities of daily living: Comparisons across national surveys. *J. Gerontology*, 45(6):S229–S237, 1990.