2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

LFIR2Pose: Pose Estimation from an Extremely Low-resolution FIR image Sequence

Saki Iwata, Yasutomo Kawanishi, Daisuke Deguchi,
Ichiro Ide, Hiroshi Murase
Nagoya University
Aichi, Japan
Email: iwatas@murase.is.i.nagoya-u.ac.jp

Tomoyoshi Aizawa
OMRON Corporation
Kyoto, Japan

***Abstract*—In this paper, we propose a method for human pose estimation from a Low-resolution Far-InfraRed (LFIR) image sequence captured by a 16×16 FIR sensor array. Human body estimation from such a single LFIR image is a hard task. For training the estimation model, annotation of the human pose to the images is also a difficult task for human. Thus, we propose the LFIR2Pose model which accepts a sequence of LFIR images and outputs the human pose of the last frame, and also propose an automatic annotation system for the model training. Additionally, considering that the scale of human body motion is largely different among body parts, we also propose a loss function focusing on the difference. Through an experiment, we evaluated the human pose estimation accuracy with an original data set, and confirmed that human pose can be estimated accurately from an LFIR image sequence.**

I. Introduction

The aging society has become a problem in most developed countries. In addition, the aging rate is expected to increase in the future due to the declining birthrate, leading to an increase in elderly population that live alone. For the health and safety of such elderly people, it is necessary to monitor their physical functions and respond to emergencies. To this end, monitoring systems for the elderly people living alone are attracting attention.

While there are different types of monitoring systems such as wearable type [1][2][3] and sensor type [4][5], it is common to recognize the behavior of an elderly person using images taken from an indoor visible light camera. However, capturing high-resolution images in everyday life gives rise to privacy concerns.

With this, Low-resolution Far-InfraRed (LFIR) sensors as shown in Fig. 1 attract attention as a sensor to avoid privacy concerns— [6][7][8][9][10]. The infrared sensor array is a collection of multiple infrared sensors arranged in a grid, and can measure the temperature distribution in a specific area. Because of its low cost, it is also used as a human sensor in home appliances such as air conditioning. Since images taken with infrared sensor arrays are in low resolution, they can reduce privacy concerns. Another advantage is that it can detect heat sources even in the dark. Figure 2 shows an RGB image and its corresponding LFIR image and joint points.

There are works using LFIR that can recognize several types of behaviors [6][9]. However, in order to maintain the healthy life of an elderly person, it is necessary to evaluate the flexibility of the body parts after estimating the behavior class recognition, but these methods cannot recognize small but significant differences within each behavior. Therefore, we propose a pose estimation method from low-resolution images taken with an infrared sensor array with consideration of privacy, named *LFIR2Pose*.

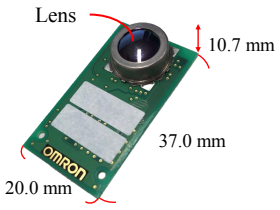
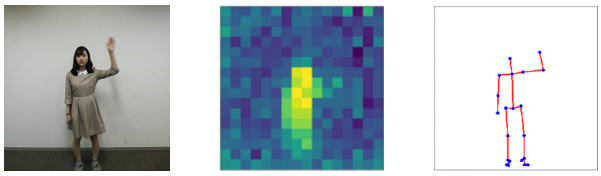


Fig. 1. Far-infrared sensor array.



(i) RGB image (ii) LFIR image (iii) Joint points

Fig. 2. Example of LFIR image and joint points.

However, 1) it is difficult to estimate an arbitrary human pose from the LFIR images. Also, for training the model, 2) it is required to prepare a significant number of annotated data but it is difficult to precisely annotate the joint point positions (ground truth) that compose the human pose from an LFIR image as shown in Fig. 2 (ii).

For problem 1), we limit the possible pose for each behavior, and build a model. Assuming that the behavior class can be obtained by an existing method such as [6][9], the variation of the human pose can be limited. Since the future motion is predictable within a behavior class even for LFIR images, the temporal information would be an effective key to estimate the human pose. Therefore, we use 3D Convolutional- Neural-Network (CNN) as the model to utilize temporal information

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

including motion information. Additionally, since there are joint points that move largely or not depending on the action, we introduce a loss function that weights each of the joint points, namely the *weighted joint point loss*. Meanwhile, for problem 2), we propose an automatic annotation system which utilizes an RGB image captured with an LFIR image synchronously.

Our contributions are as follows:

- LFIR2Pose model: A pose estimation framework which estimates human pose from LFIR images, with the following features.
  - 3D CNN: Utilize temporal information to estimate human pose.
  - Weighted joint point loss: Weighting each joint point depending on the standard deviations of their movements.
- Automatic annotation: Capturing both RGB and LFIR images, and annotating the ground truth for each LFIR image by applying OpenPose [11] to the corresponding RGB images.

Note that we assume that the behavior class is previously determined, and a model is constructed for each behavior class.

The rest of this paper is organized as follows: Sec. II introduces related work, Sec. III describes the pose estimation method from an LFIR sequence in detail, Sec. IV reports the experiment, and Sec. V concludes the paper.

II. Related Work

For human pose estimation, methods using high-resolution RGB images have been proposed. As applications using LFIR images captured by a far-infrared sensor array, human tracking [7], hand gestures recognition [8][10], and action recognition [6][9] have been proposed. Since the goal of this study is to estimate pose from LFIR images in this section, we summarize the related works in detail.

A. Human Pose Estimation

There are two approaches when modeling human poses: the bottom-up approach that extracts key point candidates in an image and connects them together to a human body, and the top-down approach that estimates the pose of each person after detecting the person. As an example of the top-down approach, Cascaded Pyramid Network (CPN) [12] is proposed. CPN consists of two networks: GlobalNet which extracts clear key points, and RefineNet which makes it possible to estimate key points that are difficult to find by upsampling and integrating features generated by the GlobalNet.

The method proposed by Xiao et al. [13] also takes the top-down approach. This is a simple yet effective baseline method using ResNet as a backbone. This is proposed in order to solve the difficulty of model comparison caused by the emergence of complicated pose estimation algorithms in recent years. Its network consists of ResNet and several deconvolution layers, and uses the simplest model that generates a heat map of joint points from low-resolution features.

On the other hand, OpenPose [11] is a representative example of the bottom-up approach. It calculates Part Confidence Maps (PCM), which indicate the likelihood of positions of the joint points as a heat map, and Part Affinity Fields (PAF), which indicate the connection between the joint points as vector fields, for an input image. Then, it estimates the pose of the person from the estimated joint point coordinates and their connections.

The method proposed by Raaj et al. [14] also uses PAF like OpenPose, but also utilizes temporal information: Temporal Affinity Fields (TAF) which represent the temporal movement of body parts. It realizes online pose tracking to link people using Spatio-Temporal Affinity Fields which consist of PAF and TAF.

Another method that takes the bottom-up approach is PersonLab [15]. It is a method that can perform human pose estimation and instance segmentation simultaneously and features offset estimation (regression) of joint points from each pixel. Concretely, it estimates a heatmap (a circle with a constant radius centered on the joint point called the Keypoint Disk), short-range offsets (a two-dimensional vector field that represents the coordinates of the joint point in the Keypoint Disk of each joint point), and mid-range offsets (a two-dimensional vector field that regresses the coordinates of the joint point corresponding to the joint point in the Keypoint Disk of each joint point type) by a CNN model and finally estimates the human pose.

Each of the methods introduced above targets high-resolution RGB images and achieves highly accurate estimation for complex poses of multiple people. However, it is difficult to apply the methods directly to LFIR images to estimate human pose due to their extremely low resolution.

B. Far-Infrared Sensor Array

The FIR sensor array used in this study is composed of 16×16 thermopile thermal infrared sensors placed just after the lens. The sensor generates an electromotive force when its temperature rises by receiving infrared rays. Because it can detect absolute values of the temperature, it can measure the surface temperature of objects in a non-contact manner. It is generally used as temperature monitoring sensors in factories and home appliances. Some works that use the sensor have been reported as follows.

A human body tracking method using an FIR sensor array has been proposed [7]. It introduces a thermo-spatial sensitive histogram to focus the target in an LFIR image sequence which allows accurate tracking by representing the target with multiple histograms and reducing the influence of the background pixels. Based on this representation, the proposed method tracks humans robustly to occlusions, pose variations, and background clutters.

In addition, to use the sensor as a user interface, hand gesture recognition methods have been proposed [8][10]. In order to reduce the problems caused by the characteristics of the FIR sensor such as low resolution, the methods focus on

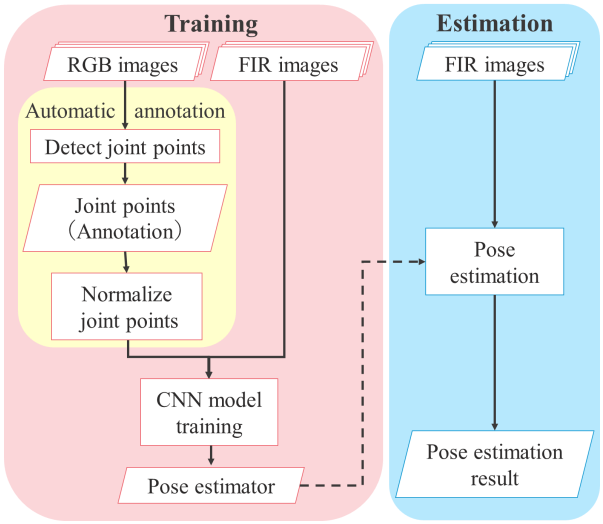


Fig. 3. Process flow of the proposed method.

the temperatures close to the human body and the region with significant motion.

As a study aimed at emergency response for the elderly people living alone, human behavior recognition methods [6] [9] have also been proposed. The method in [6] extracts features from 8×8 LFIR images and classifies them into five patterns: *stand*, *sit*, *lie low*, *stay in bed* (daily behaviors), and *tumble* (abnormal behavior), and detects emergency. It proposes the "Gradation method" considering the color of adjacent pixels. Similarly, the method in [9] realizes classification of behaviors into four patterns: *walk*, *sit*, *stand* (daily behaviors), and *fall down* (abnormal behavior). In this method, only the human area is cropped from LFIR images and their inter-frame difference is utilized. By extracting visual features using CNN and temporal information using Recurrent Neural Networks (RNN), it accomplishes highly accurate behavior classification.

However, these methods cannot detect small differences within the same behavioral class. By not only recognizing actions but also estimating the joint positions of a person in frames, we aim to understand actions in detail. This is the first attempt to estimate the human pose from extremely low-resolution FIR images.

III. LFIR2Pose

In this paper, we propose an accurate pose estimation method from an LFIR image sequence. Pose estimation from only one LFIR image is a hard task, since a target in the image is in extremely low-resolution. Therefore we design A) a 3D CNN model to handle time series of LFIR images, and B) the weighted point loss to focus on the motion information.

For training the CNN model, a number of training data with accurate annotation is required, but annotating the joint positions of the human body (ground truth) by referring only to each LFIR image is a difficult task for human. To this end, for collecting annotations automatically, C) we construct an automatic annotation system using both an FIR sensor array and an RGB camera.

Fig. 3 shows the process flow of the proposed method.

A. 3D CNN Model

We propose a 3D CNN model for the proposed pose estimation method LFIR2Pose as shown in Fig. 4 from LFIR images. In this model, by assuming that the behavior class is known in advance, pose variances are restricted and can utilize temporal information. This model takes an LFIR image sequence of N frames $X_i = (\mathbf{x}_{(i-(N-1))}, ...\mathbf{x}_{(i-1)}, \mathbf{x}_i) \in \mathbb{R}^{N \times 16 \times 16}$ as the input and outputs a $2J$-dimensional vector $\mathbf{y}_i \in \mathbb{R}^{2J}$ consisting of J joint points for the final frame of the input sequence. By using a 3D CNN, we can reduce the difficulty to estimate human poses from extremely low-resolution FIR images compared with by using only one image because temporal convolution can capture the human motion. Since LFIR images are in extremely low-resolution, applying pooling repeatedly loses much information of the input. Therefore, we do not insert any pooling layer after each convolution layer. Instead of the pooling layers, by inserting a global max pooling layer at the end of the convolutional layers, we can obtain a feature which is robust to the human position in LFIR images.

B. Weighted Joint Point Loss

Considering human behavior, such as the action of waving hands, the body joints from the arm to the hand move largely, while other joints hardly move. Estimation of the latter ones are easier than the former ones. Considering this, we would like to increase the loss for mis-estimation of largely moving joint points. Therefore, for the proposed model training, we use Mean Squared Error (MSE) with weights on each joint point defined as,

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_{2J}), \tag{1}$$

$$\sigma_j = \sqrt{\frac{1}{N_{\text{total}}} \sum_{i=1}^{N_{\text{total}}} (\mathbf{p}_{ij} - \overline{\mathbf{p}}_j)^2} \tag{2}$$

A weight vector $\mathbf{w}$ is determined per action by using joint points' ground truth $\mathbf{p}_i = (p_{i1}, p_{i2}, ..., p_{i2J})^T$, where i denotes the index of the frame and J denotes the number of joint points. By (1) and (2), the standard deviation of each joint point is scaled to $[1, \alpha]$ by referring to the maximum variance of the joint points, and the normalized values are used as the weight vector. Here N_{total} denotes the total frames for each action video. The weighted joint point loss is defined as,

$$L(\mathbf{y}_i, \mathbf{t}_i) = \mathbf{w}^T(\mathbf{y}_i - \mathbf{t}_i), \tag{3}$$

$$\mathbf{w} = \frac{(\alpha - 1)\boldsymbol{\sigma}}{\|\boldsymbol{\sigma}\|_\infty} + 1, \tag{4}$$

where $\mathbf{y}_i$ is the ground-truth pose and $\mathbf{t}_i$ is the estimated pose.

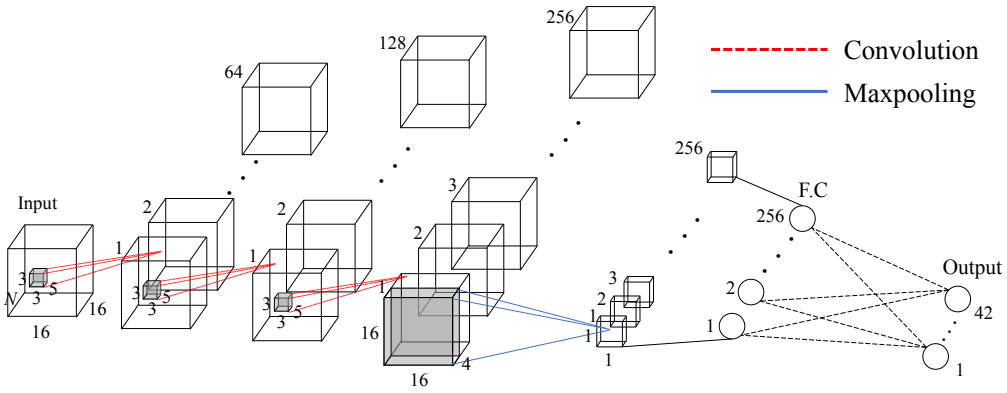


Fig. 4. Network structure of the proposed model.

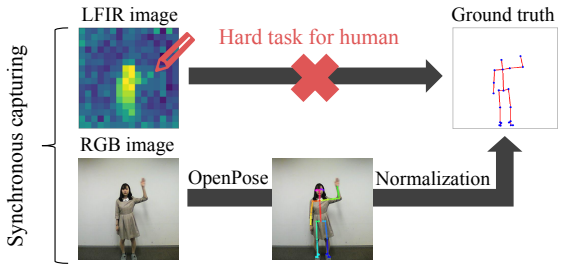


Fig. 5. Automatic annotation system.

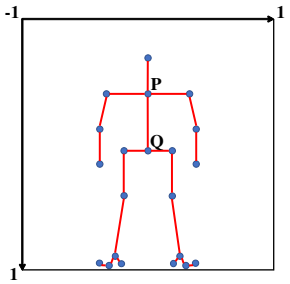


Fig. 6. Example of the pose space.

C. Automatic Annotation System

In order to estimate the human pose, the position of each joint point is required for training the CNN model. However, it is difficult to manually assign joint point positions to LFIR images. Therefore, we realize automatic annotation of the joint point positions utilizing high-resolution RGB images taken synchronously with the LFIR images (Fig. 5).

After capturing, the poses obtained by using OpenPose on the RGB images are used as the ground truth for the corresponding LFIR images. The positions of the detected joint points are scaled to be within the range of $[-1, 1]$ (Fig. 6) and the positions are aligned along with the x-coordinate of the joint point Q in Fig. 6 to the center of the space $(x = 0)$, and the human pose is offset so that the y-coordinate value of the bottom-most joint point is 1. Also, the length of body parts between P and Q in Fig. 6, which is the first frame where the person is standing, is normalized to 0.5.

TABLE I
ACTIONS IN THE DATASET

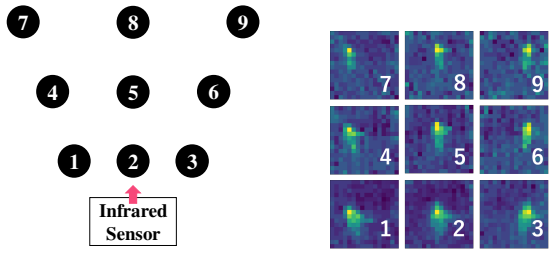
Action	Description
A	Waving the right and left hands alternately
B	Deep breathing
C	Picking up and putting down things on the floor
D	Standing and sitting

IV. EXPERIMENT

We prepared an original dataset, and experimented on the dataset to confirm the effectiveness of the proposed method.

A. Dataset

We captured data using an FIR sensor array (D6T-1616L manufactured by OMRON) and an RGB camera (BSW20KM11BK manufactured by Buffalo) synchronously. The data consists of four types of actions shown in Table I. We selected these actions because the movement of the human body are everyday actions and clear to understand in the spatial dimension. Each action was taken as a pair of RGB and LFIR video clips for 11 people and at 9 different positions relative to the sensor (Fig. 7). This results in 99 $(= 11 \times 9)$ videos (training: 55, testing: 44) for each action. The number of frames per video clip per position was about 170 frames. In the experiment, we set N to 16 so used approximately



(i) Positions of the sensor and a subject. (ii) Example at each location of the LFIR image.

Fig. 7. Positional relationship between the FIR sensor array and a subject.

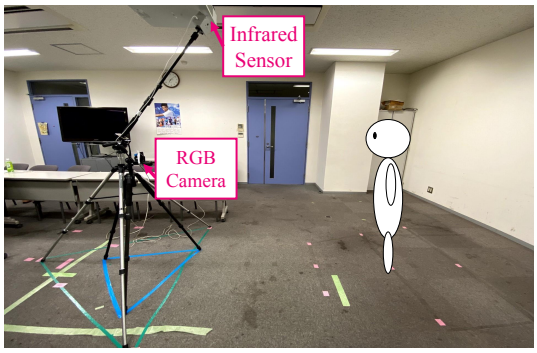
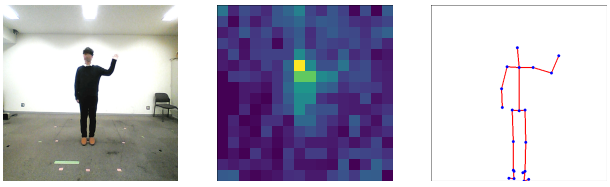


Fig. 8. Experiment environment.



(i) RGB image (ii) LFIR image (iii) Ground truth in the pose space

Fig. 9. Example of corresponding images and joint point positions.

$8,525\,(=(170-(16-1))\times 55)$ input data for training and $6,820\,(=(170-(16-1))\times 44)$ for testing.

Fig. 8 shows the experiment environment, and Fig. 9 shows examples of the captured RGB image, LFIR image, and the corresponding ground truth in the pose space. Note that since the RGB images taken for annotation are in high resolution, OpenPose works accurately. Rare mistakes were excluded by visual inspection.

B. Model Parameters

In the model, from N input LFIR images, we extract features using 3D CNN with three convolutional layers of 64, 128, and 256 channels with rectified linear unit layers. The kernel size of filters in each convolutional layer is $(frames, height, width) = (5, 3, 3)$ and the stride is 1. The padding size for the convolutional layers is

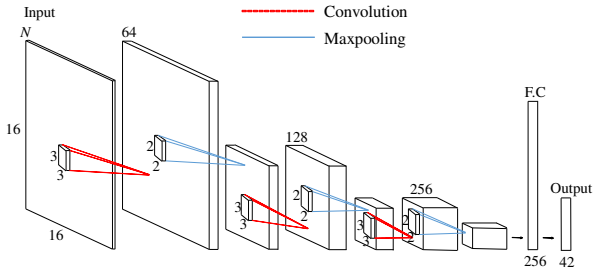


Fig. 10. Network structure of comparative model 1.

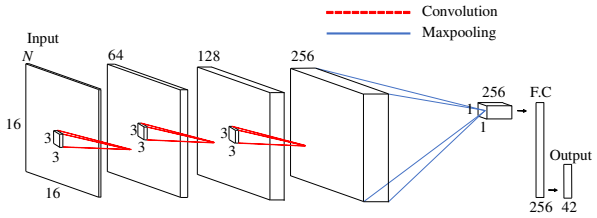


Fig. 11. Network structure of comparative model 2.

$(frames, height, width) = (0, 1, 1)$. After the global max pooling layer, the model outputs a $2J = 42$-dimensional vector $\mathbf{z} = f(X_i)$ using fully connected layers.

C. Loss Function Parameter

In this experiment, we set $\alpha = 50$ in (2). Thus, for each action, the weights of the joint points are scaled to the range of $[1, 50]$.

D. Methods and Evaluation Procedure

Since there is no study that addresses the same problem as ours, in the experiment, the following four methods made by ourselves were compared. Comparative method 1 uses the model shown in Fig. 10 which is a simple baseline model. It uses a simple three-step convolutional neural network. Comparative method 2 uses the same setting for convolutional layers as comparative method 1, while it utilizes a global max-pooling layer at the end of convolutional layers instead of the max-pooling layers in the model (Fig. 11). Therefore, the spatial size of the feature maps in this network is the same as that of the input image. In these models, the number of input LFIR images N is set to 1. Proposed methods 1 and 2 use the model shown in Fig. 4. In the model, the number of input LFIR images N is set to 16 for both methods, while the latter uses the weighted joint point loss for the training.

In order to compare the methods, we evaluated estimation accuracies on the test data which are captured at untrained subject positions. The models are trained by using the data captured at positions 1, 3, 5, 7, and 9 for each action (Fig. 7). Then the data of subject positions 2, 4, 6, and 8 are used for the evaluation.

TABLE II
RMSE OF THE GROUND TRUTH AND THE ESTIMATION RESULTS (10^{-2})

	Position	2	4	6	8	Average
Action A	Comparative method 1	4.33	**4.85**	5.35	5.11	4.91
	Comparative method 2	4.63	4.70	5.04	5.22	4.90
	Proposed method 1	2.88	3.46	3.45	**3.63**	3.36
	Proposed method 2	**2.65**	**3.12**	**3.32**	3.68	**3.19**
Action B	Comparative method 1	9.65	10.4	8.28	12.6	10.2
	Comparative method 2	7.10	5.62	5.52	7.19	6.36
	Proposed method 1	4.97	4.35	4.21	4.85	4.59
	Proposed method 2	**4.53**	**4.29**	**4.10**	**4.59**	**4.38**
Action C	Comparative method 1	11.2	10.0	16.8	10.9	12.2
	Comparative method 2	7.35	7.32	7.24	7.31	7.30
	Proposed method 1	5.13	5.12	5.07	4.99	5.07
	Proposed method 2	**4.91**	**4.96**	**5.04**	**4.96**	**4.97**
Action D	Comparative method 1	8.32	7.25	6.57	7.19	7.33
	Comparative method 2	5.53	5.67	5.01	6.34	5.64
	Proposed method 1	3.82	4.34	4.08	**5.29**	4.38
	Proposed method 2	**3.73**	**4.15**	**4.06**	5.40	**4.33**

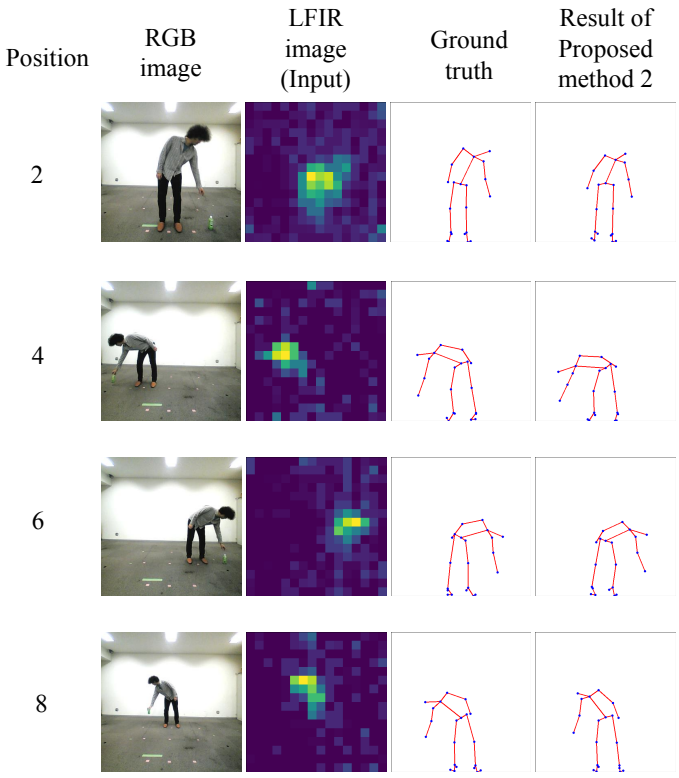
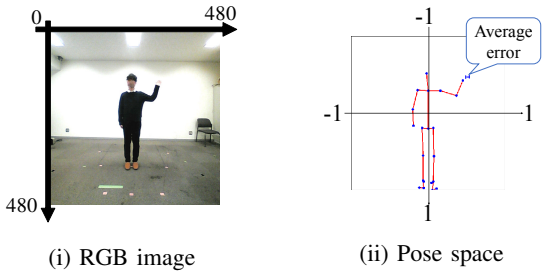


Fig. 12. Example of the pose estimation result (Action C).



(i) RGB image (ii) Pose space

Fig. 13. Scale relation between the RGB image and the pose space.

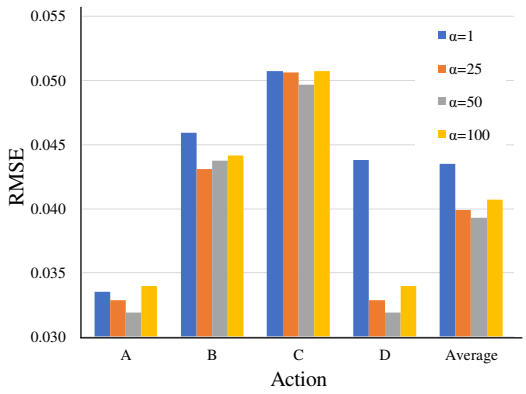


Fig. 14. Average RMSE of each action when the value of α is changed.

E. Result

Table II shows the Root Mean Square Error (RMSE) between the ground-truth joint points of each action and the estimation results, and Fig. 12 shows the example of the pose estimation results of Action C. On average, proposed method 2 achieves the highest accuracy over all actions. Comparative method 1 is quantitatively and qualitatively less accurate than the proposed method. This is because the insertion of a pooling layer after each convolutional layer fails to extract useful features from the data at various positions, resulting in inaccurate human pose estimation. We can see that increasing the number of input images and utilizing temporal information efficiently can help accurate estimation by comparing comparative method 2 and proposed method 1.

F. Estimation Error for each Joint Position based on RMSE

In the methods used in this experiment, coordinate normalization as shown in Fig. 13 was performed for training the CNN models. Since the size of the RGB image before normalization is 480×480 pixels and is normalized to the 2×2 pose space, the resolution of a pixel of the RGB image after the normalization is about $2/480 \fallingdotseq 0.004$. Based on this, the average error per joint point in proposed method 2 is as shown in Fig. 13 (ii) and is about 5 cm in actual size.

G. Effect of the Weighted Joint Point Loss

As shown in Table II, proposed method 2 is the most accurate, confirming the effectiveness of the weighted joint point loss in each action. In the experiment, $\alpha = 50$ was chosen for (2) while other values were also considered. Fig. 14 shows the average RMSE of each action for different α values. $\alpha = 1$ is equivalent to proposed method 1. From Fig. 14, we can see that the RMSE of the estimation results are lower on average compared to $\alpha = 1$, and $\alpha = 50$ is the most accurate.

V. Conclusion

In this paper, we proposed a method for estimating human pose from an LFIR image sequence.

In order to realize human pose estimation from LFIR images, we proposed a CNN model that fits low-resolution images and effectively utilizes temporal information. The weighted joint point loss, which is a loss function that promotes training a model by considering the movement of each joint according to the action, was introduced to estimate human pose more accurately. To train the CNN model, we proposed an automatic annotation system to automatically annotate ground-truth joint points to LFIR images.

Through an experiment, we confirmed that the human pose can be estimated approximately by utilizing temporal information and using a model suitable for low resolution. Also, the pose estimation accuracy was improved by introducing the proposed loss function.

Future tasks include development of a model that can extract appropriate features according to the motion, more accurate evaluation of pose estimation accuracy, and the estimation of untrained actions.

Acknowledgment

Part of this work was supported by Grant-in-Aid for Scientific Research (Grant Number 17H00745).

References

[1] Takayuki Fujita, Kentaro Masaki, and Kazusuke Maenaka. Human activity monitoring system using MEMS sensors and machine learning. *J. of Japan Society for Fuzzy Theory and Intelligent Informatics*, 20(1):3–8, 2008.

[2] Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Simone Orcioni, and Claudio Turchetti. Human activity monitoring system based on wearable SEMG and accelerometer wireless sensor nodes. *Biomedical Engineering Online*, 17(1):63–80, 2018.

[3] C. Suganthi Evangeline and Ashmiya Lenin. Human health monitoring using wearable sensor. *Sensor Review*, 39(3):364–376, 2019.

[4] Wan-Young Chung and Sung-Ju Oh. Remote monitoring system with wireless sensors module for room environment. *Sensors and Actuators B: Chemical*, 113(1):64–70, 2006.

[5] Ren C Luo, Ogst Chen, and Cheng Wei Lin. Indoor human monitoring system using wireless and pyroelectric sensory fusion system. In *Proc. 2010 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1507–1512, Oct 2010.

[6] Hyoga Fujita and Shingo Otsuka. Posture detection for elderly using infrared array sensor and fine tuning. In *Proc. 2018 IEEE Conf. on Visual Communications and Image Processing*, pages 1–4, Dec 2018.

[7] Takashi Hosono, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Tomoyoshi Aizawa, and Masato Kawade. Human tracking using a far-infrared sensor array and a thermo-spatial sensitive histogram. In *Proc. 12th Asian Conf. on Computer Vision, Part 2*, pages 262–274, Nov 2014.

[8] Chisato Toriyama, Yasutomo Kawanishi, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Tomoyoshi Aizawa, and Masato Kawade. Hand waving gesture detection using a far-infrared sensor array with thermo-spatial region of interest. In *Proc. 11th Int. Joint Conf. on Computer Vision and Computer Graphics Theory and Applications, vol.4*, pages 545–551, Feb 2016.

[9] Takayuki Kawashima, Yasutomo Kawanishi, Ichiro Ide, Hiroshi Murase, Daisuke Deguchi, Tomoyoshi Aizawa, and Masato Kawade. Action recognition from extremely low-resolution thermal image sequence. In *Proc. 17th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pages 1–6, Aug 2017.

[10] Yasutomo Kawanishi, Chisato Toriyama, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Tomoyoshi Aizawa, and Masato Kawade. Voting-based hand-waving gesture spotting from a low-resolution far-infrared image sequence. In *Proc. 2018 IEEE Conf. on Visual Communications and Image Processing*, pages 1–4, Dec 2018.

[11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 7291–7299, Apr 2017.

[12] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 7103–7112, June 2018.

[13] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proc. 15th European Conf. on Computer Vision, Part 6*, pages 466–481, Sept 2018.

[14] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields. In *Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4620–4628, June 2019.

[15] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proc. 15th European Conf. on Computer Vision, Part 14*, pages 269–286, Mar 2018.