IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 23, 2020, accepted January 9, 2021, date of publication January 13, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051375

Action Recognition From Thermal Videos Using Joint and Skeleton Information

GANBAYAR BATCHULUUN, JIN KYU KANG, DAT TIEN NGUYEN, TUYEN DANH PHAM, MUHAMMAD ARSALAN, AND KANG RYOUNG PARK, (Member, IEEE)
Division of Electronics and Electrical Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding author: Tuyen Danh Pham (phamdanhtuyen@gmail.com)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) through the Basic Science Research Program under Grant NRF-2019R1F1A1041123, in part by the MSIT, South Korea, under the Information Technology Research Center (ITRC) support program, supervised by the Institute for Information and Communications Technology Promotion (IITP), under Grant IITP-2020-2020-0-01789, and in part by the NRF funded by the MSIT through the Basic Science Research Program under Grant NRF-2019R1A2C1083813.

ABSTRACT Although various studies based on thermal images have been conducted, few studies have focused on the simultaneous extraction of joints and skeleton information of an object from a thermal image, and performed human action recognition using this information. Unlike in the case of visible light images, performing joint detection and skeleton generation on thermal images often leads to the complete disappearance of spatial information such as joints. In this case, it is extremely difficult to extract joints information from the object. Moreover, the accuracy of action recognition is significantly reduced owing to this issue. Therefore, a new method to extract joints and skeleton information is proposed in this study to address these issues. In the proposed method, an original 1-channel thermal image was converted into a 3-channel thermal image and then the images were combined to improve the extraction performance. A generative adversarial network (GAN) was used in the proposed method for extracting joints and skeleton information. In addition, research to recognize various human actions was conducted using the joints and skeleton information extracted by this method. The proposed human action recognition is performed by combining a convolutional neural network (CNN) and long short-term memory (LSTM). As a result of the experiments using self-collected and open data, it was found that the method proposed in this study shows good performance compared to other state-of-the-art methods.

INDEX TERMS Thermal image, skeleton generation, joint detection, action recognition, deep learning.

I. INTRODUCTION

Human action recognition using a camera-based surveillance system is a challenging task. In particular, performing action recognition using images captured in dark environments is especially difficult. To address this issue, near-infrared (NIR) cameras and thermal cameras (long-wavelength infrared (LWIR) camera) are used to visualize objects at near and far distances, respectively. NIR cameras cannot visualize objects at far distances without an additional illuminator whereas thermal cameras can visualize objects at near and far distances without any additional illuminator. However, thermal cameras have two major issues when performing object detection: thermal reflection and temperature similarity. Thermal reflection is caused by reflection of the heat radiated from a high temperature object on surrounding surfaces such as walls or the floor [2]. For example, a shadow-like figure is commonly detected below the body region of humans (red dashed region of Figure 1(a)) in thermal images. Temperature similarity means that the background and the object are indistinguishable from each other because they both have very similar temperatures (Figures 1 (e) and (f)). In addition, it is very difficult to perform human recognition, human identification, and joint detection when the temperature difference between the background and the object is extremely significant (Figures 1 (g) and (h)), because texture information such as patterns disappears from the body region of the object. In other words, the pixel value of the entire body region of a distant object is often 0 (black) or 255 (white) depending on the temperature of the environment in

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

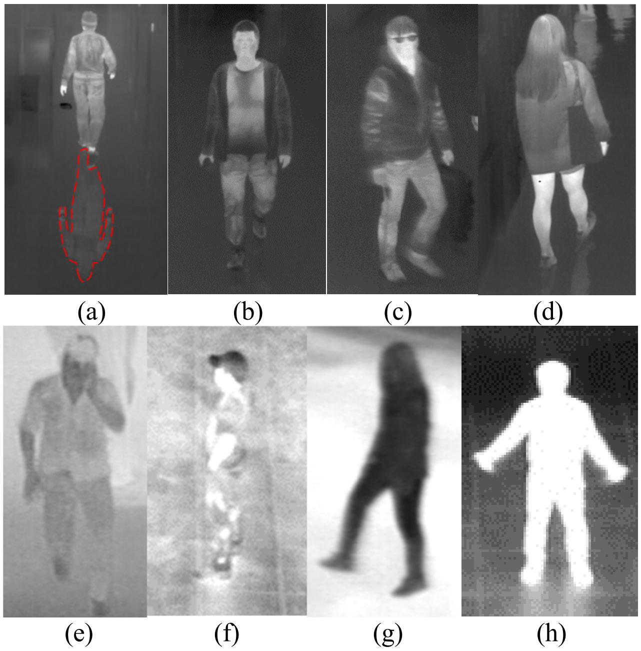


FIGURE 1. Examples of thermal images. (a)–(d) images with good quality; (e), (f) images where discrimination between object and background is difficult; (g) and (h) image where texture information inside object is not visible.

a thermal image. In this case, it is difficult to detect body joint features such as ankles, knees, hips, wrists, elbows, and shoulders easily. To address these issues, joints and skeleton information were extracted from thermal images acquired in various environments using a deep learning method in this study.

Furthermore, only a few studies have previously performed action recognition using thermal images, and studies that considered all these issues and used deep learning-based methods to recognize the action of distant objects in thermal images do not exist. Therefore, a human action recognition method that combines a generative adversarial network (GAN), a convolutional neural network (CNN), and long short-term memory (LSTM) is proposed in this paper to extract spatial and temporal features. This research is novel in the following four ways compared to the previous works:

- This is the first study that employs body skeleton and joint detection based on 1-channel and 3-channel thermal images. 1-channel thermal images have been used in previous studies. However, the 1-channel thermal images in the proposed method are converted into 3-channel thermal images, and then these images are combined to improve the accuracy of skeleton generation and joint detection.
- A joint-GAN is proposed in this study to simultaneously extract skeleton and joint information from the 1-channel and 3-channel thermal images.
- A new CNN-LSTM that uses the extracted skeleton and joint information as input is proposed, and the number of trainable parameters of the CNN-LSTM is reduced by decreasing the number of input features of the LSTM.
- The trained Joint-GAN and CNN-LSTM models proposed in this study are disclosed in [3] for fair performance evaluation by other researchers.

The rest of the paper is organized as follows. Previous studies related to skeletal generation, joint detection, and action recognition are described in Section II. A detailed description of the method proposed in this study is described in Section III. The experimental results and analysis are presented in Section IV, and lastly, the conclusion of the paper is presented in Section V.

II. RELATED WORKS

Conventional human action recognition methods can be mainly divided into two types: methods performed using thermal images and methods that do not use a thermal image.

A. PREVIOUS RESEARCH ACCORDING TO IMAGE ACQUISITION SENSORS

Most of the previous studies have used visible light images as input. In [4], authors proposed the method of human action recognition by using CNN and an optimized deep autoencoder (DAE) for a real-time challenge. The method showed the promising performance by using DAE and quadratic SVM on action recognition task. In [5], authors introduced the methods of action and activity recognition by using LSTM. In addition, they discussed the drawbacks of existing action and activity recognition methods based on RNN compared to LSTM. In [6], authors proposed multi-view action recognition (MVAR) method by using a CNN and a conflux long short-term memory (LSTMs) network. In the study, their proposed method achieved high action recognition accuracy by using parallel LSTMs. In [7], behavior classification was performed by extracting enhanced gait energy image (EGEI) handcrafted features. In [8], human action recognition was performed based on the convexity defect feature point; however, many inaccurate feature points were detected owing to noise in the environment, resulting in poor action recognition accuracy. In [9], [10], a method for simultaneous end-to-end training was proposed by connecting CNN and LSTM.

In addition, some studies have performed action recognition using features previously extracted from data acquired with a Kinect sensor or motion capture system, or with red, green, blue, and depth (RGB+D) data. In [11], [12], human action recognition was performed using CNN and joint maps information. In [13], action recognition was performed using skeleton information as input to a recurrent neural network (RNN). RNNs have two disadvantages: vanishing and exploding. That is, important information may disappear, or unimportant information may accumulate as the continuous input feature is lengthened. In [14]–[17], LSTM network-based action recognition methods using skeleton information are proposed to address this issue. LSTM-based methods used input, output, and forget gait functions to address vanishing and exploding issues. In [18], a method of performing action recognition was proposed by combining the scores obtained using skeleton joint information as input to LSTM networks and using joint distance maps as CNN inputs to extract both spatial and temporal information.

Although the above-described visible light image-based methods provide extensive texture information in an image, their performance is affected by low light or changes in ambient lighting, which is a disadvantage. In addition, a special device for obtaining depth information needs to be used and distance data is difficult to acquire at long distances when using a Kinect sensor or motion capture system, or features extracted in advance from RGB+D data. Because of these drawbacks, various studies have performed action recognition using thermal images. In [19], optical flow-based methods for accurately extracting motion features were proposed. However, feature extraction in optical flow-based methods is time-consuming and these features are sensitive to noise and illumination. In [20], human action recognition was performed by extracting gait energy image (GEI) handcrafted features. In [21], human action recognition was performed based on the projection-based distance (PbD) method. In [22], human action recognition was performed by extracting various features such as ratio of foreground and background and using fuzzy rules. In [23], human action recognition was performed by generating skeleton binary images from thermal images and using these images continuously as inputs to CNN-LSTM.

B. PREVIOUS RESEARCH ON EXTRACTION OF SKELETON AND JOINT INFORMATION

Previous studies that have extracted skeleton information can be divided into studies that did and did not use deep learning methods. In the latter studies, methods for extracting skeleton information from a binary image, grayscale image, and visible light image were proposed in [24]–[26], [27]–[30], and [31], respectively. In [26], a method for extracting skeleton information was proposed by performing a thinning algorithm based on mathematical morphology. In [27], the skeleton of the body parts was separately extracted from grayscale images using the Dijkstra's algorithm. In [28], the skeleton was extracted from grayscale images using a pseudo distance map (PDM). In [29], a grayscale image thinning method using PDM was proposed, and skeleton detection was performed using a binary-like thinning algorithm. In [30], skeleton extraction was performed directly from grayscale images based on anisotropic vector diffusion without using a segmentation algorithm. In [31], a co-skeletonization method was proposed to extract skeleton information from visible light images. Regarding skeleton detection based on deep learning algorithms, the following methods have been proposed. Methods to extract skeleton information from a binary image, a visible light image, and the thermal image were proposed in [32], [33], [34]–[37], and [23], respectively. In [32], a U-Net-based method was proposed. In [33], a skeleton was generated using CNN, and the generated skeleton image was partitioned into skeleton branches using Gaussian mixture models. In [34], a DeepFlux method was proposed to extract skeleton information from visible light images using CNN. In [35], a method based on a fully convolutional network (FCN) that extracts skeleton information from visible light images was proposed. In [36], a DeepSkeleton method was proposed to extract skeleton information from visible light images. In [37], a side-output residual network (SRN) was employed to extract skeleton information from visible light images, and in [23], a binary skeleton image was generated using a 1-channel thermal image.

Moreover, previous studies that extract joint information can be divided into studies that used and those that did not use deep learning methods. In the latter studies [38], [39], joints were detected by generating binary images from X-ray images. In [38], a method for extracting binary images from X-ray images using Otsu's binarization method and detecting joints from these binary images was proposed. In [39], a method to detect joint location and joint margin in X-ray images was proposed. In [40], a method to extract both skeleton and joint information from depth images was proposed using a thinning algorithm. In [41], skeleton and joint information was extracted from visible light images based on body parts dependent joint regressors. In this study, a more accurate method employed body part templates using two-layered random forests as joint regressors. In [42], although a study was conducted to extract skeleton and joint information using thermal images, this information was extracted from clothes with additional sensors installed, instead of generating a skeleton image directly from the thermal image.

In addition, the following joint detection methods are based on deep learning algorithms. In [43], a method using a hybrid architecture including deep CNN and a Markov random field was proposed. In this method, skeleton and joint information were extracted simultaneously from visible light images. In [44], a method to extract skeleton and joint information from visible light images was proposed using part affinity fields (PAF) and a two-branch multi-stage CNN. In [45], skeleton and joint information was extracted from visible light images using the convolutional pose machines (CPM) method. In [46], skeleton and joint information was extracted from visible light images using the CNN-based part detectors-based DeepCut method. In [47], skeleton and joint information was extracted from visible light images based on CPM. In this study, the cases of overlapped or truncated persons are also considered while extracting the skeleton and joint information of multiple persons. Although binary images and visible light images were used to extract skeleton and joint information in all of the above studies, only a few methods employ thermal images. This is because there is not enough texture information in the thermal images to detect skeleton and joint information. In addition, there is no existing research that extracts both skeleton and joint information from a thermal image using a GAN-based deep learning method. To address this issue, a method to detect skeleton and joints from thermal images using Joint-GAN and perform human action

TABLE 1. Summary of comparisons between the proposed and previous research on action recognition.

Category		Method	Advantage	Disadvantage
Without using thermal images	Without using deep learning	Using visible light images [7, 8]	Large data acquisition, processing, and training are not required	- Hand-craft features are unsuitable for various environments and camera settings - Performance is affected by shadows, illumination variations, and human clothing of various colors - Subject is not visible in a dark environment
	CNN-based	Using visible light images [4] and skeleton joint information [11, 12]	- Good at extracting spatial information - Feature extraction is suitable even in various environments and camera settings	- Temporal information loss is inevitable - Performance is affected by shadows, illumination variations, and human clothing of various colors - Subject is not visible in a dark environment
	RNN-based	Using skeleton joint information [13]	- Good at extracting temporal information - Feature extraction is suitable even in various environments and camera settings	- Spatial information loss is inevitable - Encounters vanishing and exploding issues - Special sensor is required to obtain skeleton joint information
	LSTM-based	Using visible light images [5] and skeleton joint information [14-18]	- Good at extracting temporal information - Overcomes vanishing and exploding issues - Feature extraction is suitable even in various environments and camera settings	- Spatial information loss is inevitable - Special sensor is required to obtain skeleton joint information
	CNN-LSTM	Using visible light images [6, 9, 10]	- Good at extracting spatial and temporal information - Overcomes vanishing and exploding issues - Feature extraction is suitable even in various environments and camera settings	- Memory consumption and training time are expensive - Performance is affected by shadows, illumination variations, and human clothing of various colors - Large data acquisition, processing, and training are required
Using thermal images	Without using deep learning	Using 1-channel thermal images [19–22]	- Large data acquisition, processing, and training are not required - Subject is visible in a dark environment	- Hand-craft features are unsuitable for various environments and camera settings - Performance is affected by temperature variations, halo effects, and thermal reflections
	CNN-LSTM	Using 1-channel thermal images and skeleton information [23]	- Good at extracting spatial and temporal information - Overcomes vanishing and exploding issues - Feature extraction is suitable even in various environments and camera settings - Subject is visible in a dark environment	- Memory consumption and training time are expensive - Big data acquisition, processing, and training are required - Performance is affected by temperature variations, halo effects and thermal reflections
		Joint-GAN with 1-channel and 3-channel thermal images, and skeleton and joint information (**proposed method**)	- Good at extracting spatial and temporal information, - 3-channel thermal image provides more information - Skeleton and joint detection by Joint-GAN	Memory consumption and large training time are inevitable

TABLE 2. Summary of comparisons between the proposed and previous research on skeleton and joint detection.

Category		Method	Advantage	Disadvantage
Without using thermal images	Without using deep learning	- Skeleton information using binary images [24–26], grayscale images [27–30], and visible light images [31] - Joint information using X-ray images [38, 39] - Skeleton and joint information using depth images [40] and visible light images [41]	Large data acquisition, processing, and training are not required	- Hand-craft features are unsuitable for various environments and camera settings - Performance is affected by shadows, illumination variations, and human clothing of various colors in visible light image - Depth or X-ray sensor s required
	CNN-based	- Skeleton information using binary images [32, 33] and visible light images [34–37] - Skeleton and joint information using visible light images [43–47]	Feature extraction is suitable even in various environments and camera settings	- Performance is affected by shadows, illumination variations, and human clothing of various colors in visible light image - Subject is not visible in a dark environment
Using thermal images	Without using deep learning	Skeleton and joint information using 1-channel thermal images and a suit with sensors [42]	- Large data acquisition, processing, and training are not required - Subject is visible in a dark environment	- Hand-craft features are unsuitable for various environments and camera settings - Performance is affected by temperature variations, halo effects, and thermal reflections
	CNN-based	Skeleton information using 1-channel thermal images [23]	- Feature extraction is suitable even in various environments and camera settings - Subject is visible in a dark environment	- Performance is affected by temperature variations, halo effects, and thermal reflections - Acquisition of correct skeleton is difficult
	GAN-based	Skeleton and joint information using 1-channel and 3-channel thermal images (**proposed method**)	- Feature extraction is suitable even in various environments and camera settings - 3-channel thermal image provides more information - Skeleton and joint detection by Joint-GAN	Memory consumption and large training time are inevitable

recognition using outputs of Joint-GAN as inputs of CNN-LSTM is proposed in this study. The comparison of the advantages and disadvantages of the methods proposed in this study and the aforementioned previous studies is as shown in Tables 1 and 2.

III. PROPOSED METHOD

A. OVERALL PROCEDURE OF THE PROPOSED METHOD

In this section, the methods proposed in this study are described in detail. The overall flowchart of the proposed methods is shown in Figure 2. In the proposed methods,

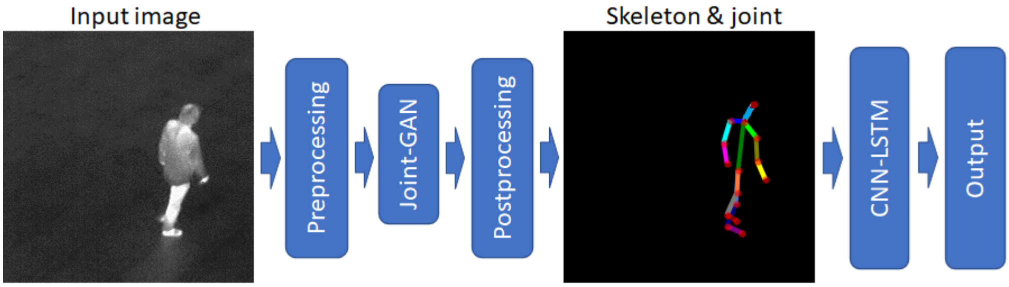


FIGURE 2. Overall flowchart of the proposed method.

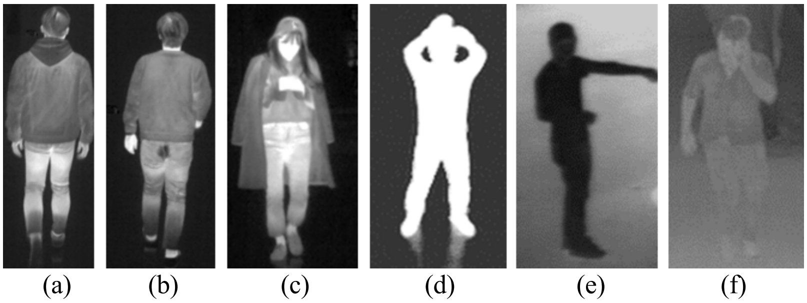


FIGURE 3. Example of thermal images. (a)–(c) images with high spatial texture information; (d), (f) images with low spatial texture information; (e) images of similar pixel values between object and background.

skeleton and joint extraction is performed based on 1-channel and 3-channel thermal images by using preprocessing, Joint-GAN, and postprocessing, and action recognition is performed based on joint and skeleton information using CNN-LSTM. In Section III.*B*, the preprocessing, Joint-GAN, and postprocessing for joint and skeleton extraction are explained in detail. In Section III.*C*, the action recognition method using the CNN-LSTM that uses the extracted joint and skeleton information as input is described.

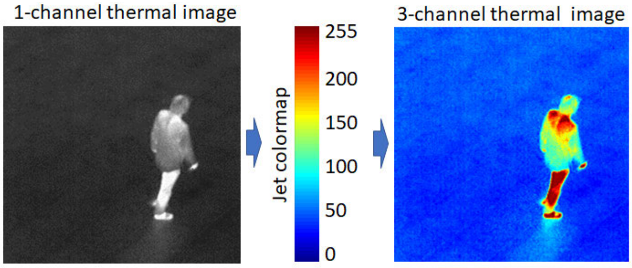


FIGURE 4. Example of preprocessing. The color conversion from a 1-channel grayscale image to a 3-channel RGB image.

B. SKELETON AND JOINT EXTRACTION

1) PREPROCESSING

In this section, the proposed skeleton and joint extraction method, together with its Joint-GAN, is described in detail. There are three common cases when extracting skeleton and joint information from images acquired with a thermal camera. The first case corresponds to thermal images with clear spatial texture information, such as that shown in Figure 3(a–c). It is easy to extract joint information from these images. In the second case, the body region of the object does not appear to have spatial texture information owing to the extreme difference in temperature between the body and the environment, such as the binary images shown in Figures 3(d) and 3(e). Although skeleton information can be obtained from these images, it is difficult to extract joint information. In the third case, pixel values corresponding to the object and the background may appear similar to each other owing to temperature similarity, as shown in Figure 3(f). In this case, it is difficult to extract the skeleton and joint information since the human body is hardly distinguishable from the background. When the temperature of the surrounding background is lower, higher, and similar to that of a person, images are acquired as shown in Figure 3(d), Figure 3(e), and Figure 3(f), respectively. Considering these issues, a GAN-based method that simultaneously extracts joint and skeleton information is proposed for the first time in this study.

The preprocessing of the image that is used as an input to the Joint-GAN method is as follows. A 1-channel thermal image was converted into a 3-channel thermal image using a colormap function. The jet colormap array [48] was used to perform color conversion. Jet colormap array was selected as the most appropriate mapping function for the present study. Jet colormap array maps 256 pixel values from 0 to 255 from 1-channel image to 3-channel image. For example, the pixel value of the region with the highest temperature is

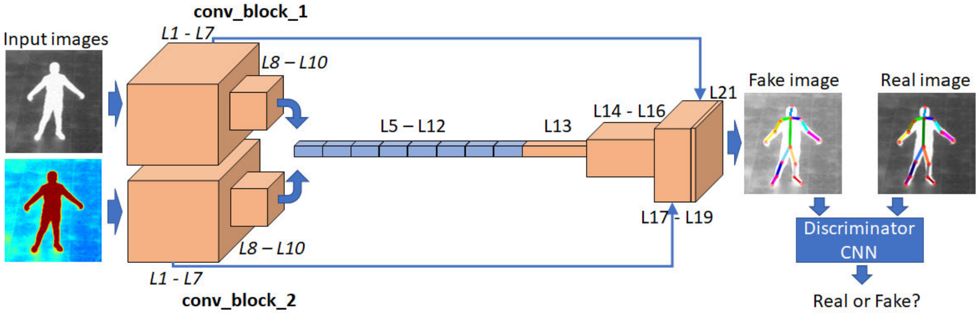


FIGURE 5. Architecture of the proposed Joint-GAN method.

TABLE 3. Description of the generator network of the proposed Joint-GAN.

Layer number	Layer type	Number of filters	Number of parameters	Layer connection (connected to)
0	input_layer_1	0	0	input_1
1	input_layer_2	0	0	input_2
2	conv_block_1	64	484,928	input_layer_1
3	conv_block_2	64	495,296	input_layer_2
4	add_1			conv_block_1 & conv_block_2
5	res_block_1	64	73,920	add_1
6	res_block_2	64	73,920	res_block_1
7	res_block_3	64	73,920	res_block_2
8	res_block_4	64	73,920	res_block_3
9	res_block_5	64	73,920	res_block_4
10	res_block_6	64	73,920	res_block_5
11	res_block_7	64	73,920	res_block_6
12	res_block_8	64	73,920	res_block_7
13	conv2d_1	256	147,712	res_block_8
14	up2d_1			conv2d_1
15	lrelu_1			up2d_1
16	conv2d_2	256	590,080	lrelu_1
17	up2d_2			conv2d_2
18	lrelu_2			up2d_2
19	conv2d_3	64	147,520	lrelu_2
20	add_2			conv2d_3 & conv2d_8 (conv_block_1) & conv2d_8 (conv_block_2)
21	conv2d_4	3	1,731	add_2
22	tanh		0	conv2d_4
Total number of trainable parameters: 2,458,627				

255 (white color), and the pixel value of the region with the lowest temperature is 0 (black color) in a 1-channel image. In contrast, the pixel value of the region with the highest temperature is [255,0,0] (red color), and the pixel value of the area with the lowest temperature is [0,0,255] (blue color) in the 3-channel ([Red, Green, Blue]) image. A color conversion example is shown in figure 4. In the proposed method, the 1-channel thermal image was changed to a 3-channel thermal image. The reason is that, in various previous studies, object detection, recognition, and classification using a color visible light image showed higher performance than using a grayscale visible light image [49].

TABLE 4. Description of the convolution block of the generator network.

Layer number	Layer type	Number of filters	Layer connection (connected to)
1	conv2d_5	64	input
2	prelu_1		conv2d_5
3	conv2d_6	64	prelu_1
4	prelu_2		conv2d_6
5	conv2d_7	64	prelu_2
6	conv2d_8	64	conv2d_7
7	maxpool_1		conv2d_8
8	conv2d_9	64	maxpool_1
9	conv2d_10	64	conv2d_9
10	maxpool_2		conv2d_10

2) JOINT-GAN

The 3-channel image obtained through preprocessing, such as that shown in Figure 4, was used as the input for the Joint-GAN. The structure of the Joint-GAN is shown in Figure 5. L5–L21 and *L1–L10* represent the layer numbers and layer numbers of convolution blocks, respectively. In addition, the specific contents of the structure are shown in Table 3–7. In Table 3–7, the filter size, stride, and padding are (3 × 3), (1 × 1), and (1 × 1), respectively. Prelu, lrelu, tanh, res_block, conv2d, add, conv_block, up2d, dense, and sigmoid represent parametric rectified linear unit (relu), leaky relu, hyperbolic tangent activation function, residual block, 2-dimensional convolution layer, addition operation, convolution block, upsampling, fully connected layer, and sigmoid activation function, respectively. Two images of a 3-channel thermal image (224 × 224 × 3) and a 1-channel thermal image (224 × 224 × 1) are used as inputs and the output image size is (224 × 224 × 3), as shown in Table 3. As shown in Table 6, the strides of conv_block_1, conv_block_3, and conv_block_6 were 1, and the padding of conv_block_1 was (1 × 1). The stride and padding of the other convolution blocks were 2 and (0 × 0), respectively. The filter size of all convolution blocks, input image size, and output size were (3 × 3), (224 × 224 × 3), and (1 × 1), respectively.

3) POSTPROCESSING

In postprocessing, skeleton and joint information are extracted from the RGB output image obtained by the

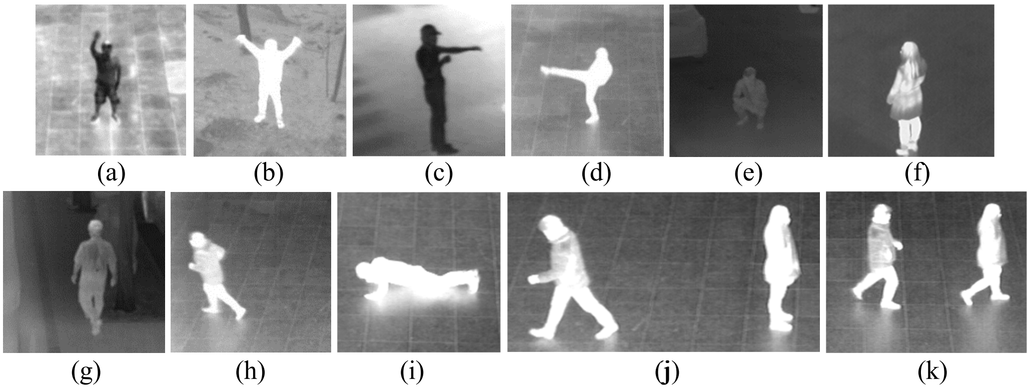


FIGURE 6. Example images of human actions in this study. (a) one hand waving; (b) two hands waving; (c) punching; (d) kicking; (e) sitting; (f) standing; (g) walking; (h) running; (i) lying down; (j) leaving; (k) approaching.

TABLE 5. Description of the residual block.

Layer number	Layer type	Number of filters	Layer connection (connected to)
1	conv2d_11	64	input
2	prelu_3		conv2d_11
3	conv2d_12	64	prelu_3
4	add_3		conv2d_12 & input

TABLE 6. Description of the discriminator network of the Joint-GAN.

Layer number	Layer type	Number of filters	Number of parameters	Layer connection (connected to)
0	input layer	0	0	input_3
1	conv_block_1	32	896	input layer
2	conv_block_2	64	18,496	conv_block_1
3	conv_block_3	64	36,928	conv_block_2
4	conv_block_4	128	73,856	conv_block_3
5	conv_block_5	128	147,584	conv_block_4
6	conv_block_6	256	295,168	conv_block_5
	conv_block_7	256	590,080	conv_block_6
7	lrelu_1		0	conv_block_7
8	dense		36,865	lrelu_1
9	sigmoid		0	dense
Total number of trainable parameters: 1,199,873				

TABLE 7. Description of the convolution block of the discriminator network.

Layer number	Layer type	Layer connection (connected to)
1	conv2d_1	input
2	lrelu_2	conv2d_1

TABLE 8. Algorithm for extracting skeleton and joint information.

```
in = (224 × 224 × 3), out = (224 × 224 × 3), joint_img = (224 × 224 × 3)
for i in range(224):
    for j in range(224):
        val = abs(in(i,j,0) – out(i,j,0))
              + abs (in(i,j,1) – out(i,j,1))
              + abs(in(i,j,2) – out(i,j,2))
        if (val > threshold):
            joint_img(i,j) = out(i,j)
        else  joint_img(i,j) = 0
```

Joint-GAN using the following algorithm. As shown in Table 8, in and out represent the input grayscale image (converted to 3-channel grayscale image) and the output color image of the generator, respectively. Moreover, in(i,j,0), in(i,j,1), and in(i,j,2) represent the red, green, and blue pixels corresponding to the (i,j) coordinates of the in image, respectively. The skeleton and joint information extracted from the out image are then converted into a joint_img image. The obtained joint_img is used as the input for the CNN-LSTM described in Section III.*C* to obtain the action recognition results.

C. ACTION RECOGNITION

The action recognition method is described in detail in this section. The purpose of this study is to recognize 11 actions: "one hand waving", "two hands waving", "punching", "kicking", "sitting", "standing", "walking", "running", "lying down", "leaving," and "approaching." The duration of each action was assumed to be different, as shown in Figure 6. In addition, the length of the input image sequence was set to 30 since the number of consecutive images of each action was different. A new CNN-LSTM is proposed in this study to extract spatial and temporal features from the skeleton and joint images described in Section III.*B.3*. CNN and LSTM are suitable networks to extract spatial and temporal information, respectively. The LSTM-based method addresses the vanishing and exploding issues that occur in the conventional RNN-based method. As shown in Figure 7, the proposed CNN-LSTM is interconnected and continuously trained using the input image. Various layer configurations of the proposed CNN-LSTM were constructed prior to the experiments to choose the most suitable network. The selected network is described in detail in Table 9. In Table 9, conv2d and pool represent the 2D convolutional layer and max pooling layer, respectively.

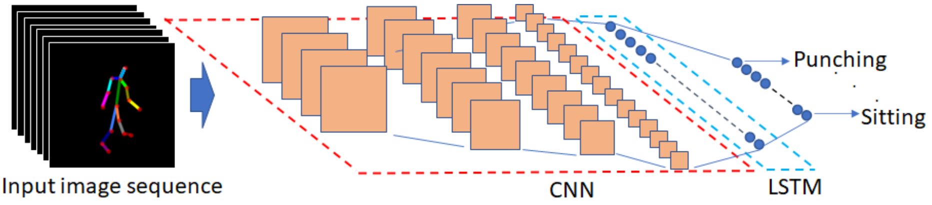


FIGURE 7. Example of our CNN-LSTM architecture.

TABLE 9. Detailed description of CNN-LSTM.

Iteration	Layer number	Layer type	Size of feature map (height × width × channel)	Number of filters	Filter size	Stride	Padding	Number of Parameter
30 times	0	input layer	224 × 224 × 3					0
	1	conv2d_1_1	222 × 222 × 64	64	3 × 3	1 × 1	0 × 0	1,792
	2	relu1_1	222 × 222 × 64					
	3	conv2d_1_2	220 × 220 × 64	64	3 × 3	1 × 1	0 × 0	36,928
	4	relu1_2	220 × 220 × 64					
	5	pool_1	110 × 110 × 64			2 × 2	0 × 0	
	6	conv2d_2_1	108 × 108 × 128	128	3 × 3	1 × 1	0 × 0	73,856
	7	relu2_1	108 × 108 × 128					
	8	conv2d_2_2	106 × 106 × 128	128	3 × 3	1 × 1	0 × 0	147,584
	9	relu2_2	106 × 106 × 128					
	10	pool_2	53 × 53 × 128			2 × 2	0 × 0	
	11	conv2d_3_1	51 × 51 × 128	128	3 × 3	1 × 1	0 × 0	147,584
	12	relu3_1	51 × 51 × 128					
	13	conv2d_3_2	49 × 49 × 128	128	3 × 3	1 × 1	0 × 0	147,584
	14	relu3_2	49 × 49 × 128					
	15	pool_3	24 × 24 × 128			2 × 2	0 × 0	
	16	conv2d_4_1	22 × 22 × 128	128	3 × 3	1 × 1	0 × 0	147,584
	17	relu4_1	22 × 22 × 128					
	18	conv2d_4_2	20 × 20 × 128	128	3 × 3	1 × 1	0 × 0	147,584
	19	relu4_2	20 × 20 × 128					
	20	pool_4	10 × 10 × 128			2 × 2	0 × 0	
	21	conv2d_5_1	8 × 8 × 64	64	3 × 3	1 × 1	0 × 0	73,792
	22	relu5_1	8 × 8 × 64					
	23	conv2d_5_2	6 × 6 × 64	64	3 × 3	1 × 1	0 × 0	36,928
	24	relu5_2	6 × 6 × 64					
	25	pool_5	3 × 3 × 64			2 × 2	0 × 0	
	26	fc6_1	100 × 1					57,700
	27	relu6_1	100 × 1					
	28	dropout6_1	100 × 1					
	29	LSTM	100 × 1					80,400
1 time	30	fc7_1	Number of classes × 1					909
	31	Softmax layer7_1	Number of classes × 1					
Total trainable parameters: **1,100,225**								

In addition, relu and fc represent rectified linear unit and fully connected layers, respectively. As shown in Table 9, the output feature with the size of 100 × 1 extracted from layer #28 is used as an input to the LSTM layer (layer #29). Then, the output of the LSTM layer is passed through the layers #30 and #31 to get the final result. Moreover, this concept is presented in Figure 7.

The input to the CNN-LSTM model is a 3-channel color image (skeleton and joint image obtained by our Joint-GAN) with the size of 224 × 224 × 3. The model is fed by 30 sequential images, iteratively. In addition, we do not use 3D CNN but 2D CNN in this method. As shown in Table 9, temporal patterns are captured by iterating layers #0 $\sim$ #29 by 30 times using 30 sequential images. The number of classes in the selected configuration was nine. In addition, the number of trainable parameters was decreased by reducing the number of input features of the LSTM, as shown in Table 9. The "approaching" and "leaving" actions were recognized in this study based on the coordinates obtained by a region of interest (ROI) detection method instead of using the CNN-LSTM [50]. In other words, the distance between two objects was measured using coordinates, and it was recognized as "leaving" when the distance value increased and as "approaching" when the distance value decreased.

IV. EXPERIMENTAL RESULTS

A. DESCRIPTION OF EXPERIMENTAL SETUP AND DATABASES

Experiments were performed using the self-collected data DTh-DB and DI&V-DB [51]. These databases consist of thermal images of near-field objects acquired in dark or bright

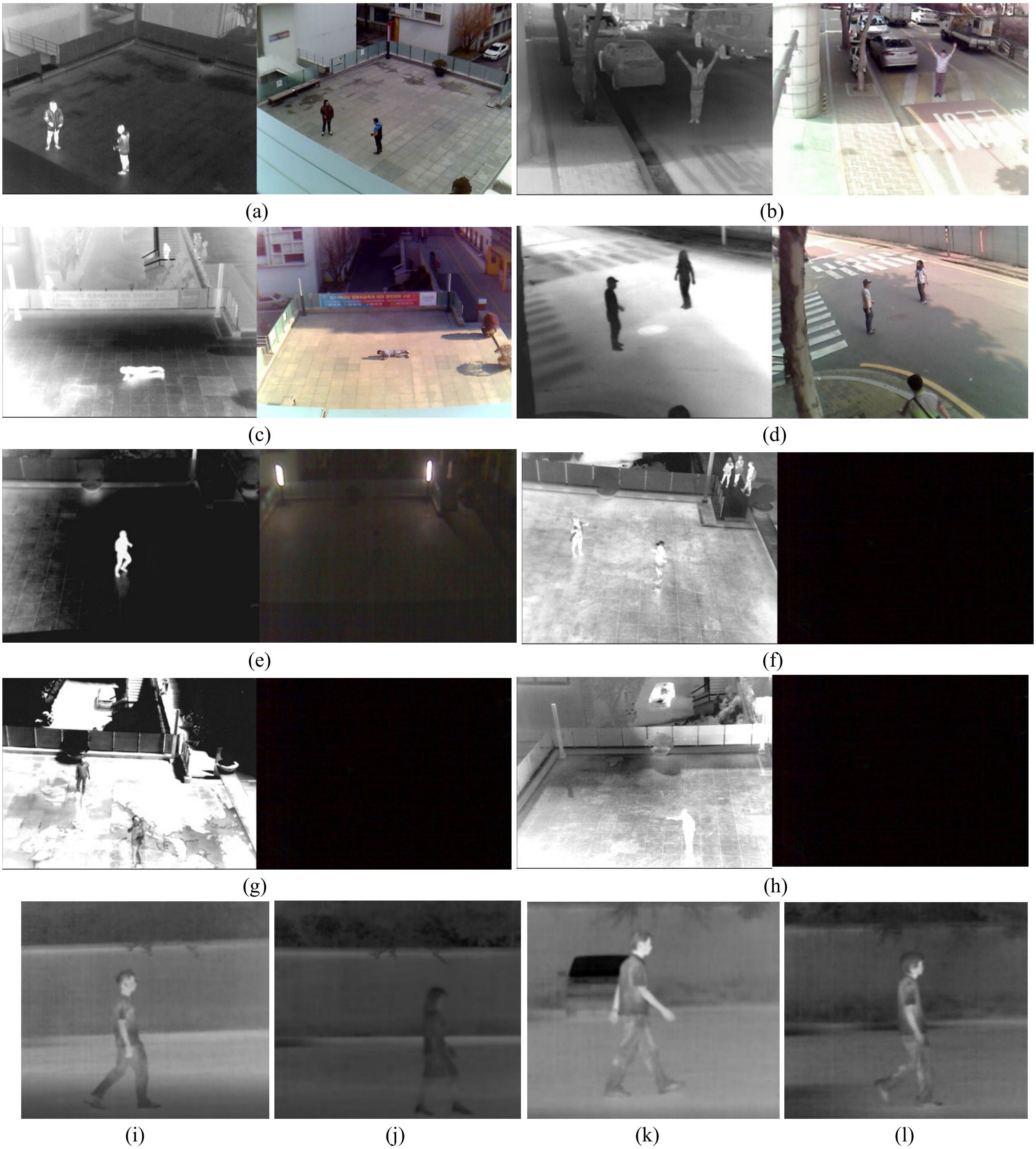


FIGURE 8. Example images from the databases. (a–h) images were taken from the self-collected DTh-DB and DI&V-DB (thermal images on the left and corresponding visible light images on the right); (i–l) images were taken from the CASIA C dataset. (a–d) images were captured in daytime; (e–h) images were captured in nighttime.

indoor environments (including dawn, daytime, and night) and distant objects acquired in dark or bright outdoor environments. Images including distant objects acquired outdoors were used in this study. Databases were built using different camera settings in various weathers and seasons at nine locations. Although the database contains visible light images and thermal images, only thermal images were used in this study. The frame rate of the thermal camera was 30 frames per second (fps) [52]. The depth of the image was 14 bits and the size was 640 × 480 pixels. [21], [22] can be referred

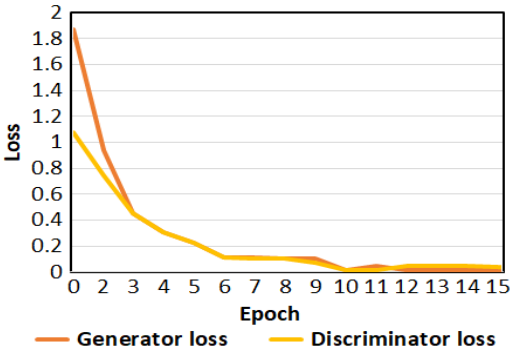


FIGURE 9. Training loss curves of Joint-GAN.

to for more information on the camera settings used when building the database. In addition, comparative experiments were performed using the CASIA C open dataset [53]. Example images of all databases used in this study are shown in Figure 8. The experiment was conducted in two-fold cross validation. In other words, half of the total data were used for training, the other half for testing, and the average value of the two testing accuracies obtained by repeating the same process after swapping the training and testing data was set as the final accuracy. The training and testing of the algorithm proposed in this study were performed using a desktop computer. The desktop computer ran under an Intel core i7-6700 CPU @ 3.40 GHz, an Nvidia GeForce GTX TITAN X graphic processing unit (GPU) card [54], and a random-access memory (RAM) of 32 GB. The model and algorithm proposed in this study were implemented using the OpenCV library (version 4.3.0) [55], Python (version 3.5.4), and the Keras application programming interface (API) (version 2.1.6-tf) with Tensorflow backend engine (version 1.9.0) [56].

B. TRAINING AND TESTING PROCEDURES OF THE MODELS

The Joint-GAN and CNN-LSTM models used in this study were trained as follows. In the CNN-LSTM, the length of the input series, training epoch, learning rate, momentum, batch-size, optimizer, and loss functions were 30, 65, 0.0001, 0.9, 1, adaptive moment estimation (Adam) [57], and categorical cross-entropy loss [58], respectively. In the Joint-GAN, the batch-size, training epoch, generator loss, discriminator loss, learning rate, and optimizer were set to 1, 100, a loss calculated on features by visual geometry group (VGG)-19 [59] with binary cross-entropy loss, binary cross-entropy loss, 0.0001, and Adam, respectively. Images of 224 × 224 pixels were used for training and testing. The training loss curves of the Joint-GAN method are shown in Figure 9. Moreover, the loss and accuracy curves obtained in the training stage of the CNN-LSTM method for each epoch are shown in Figure 10. Furthermore, the accuracy and loss curves of action recognition methods using the original 1-channel grayscale thermal image (Method 1), converted 3-channel color thermal image (Method 2), and the skeleton and joint image obtained from both the 1-channel and 3-channel images by our method (Proposed method) are shown in Figure 10. Although the classification was performed using the softmax function in CNN-LSTM during the training stage, it was performed by extracting the 100 × 1 feature from layer number 29 in Table 9 during the testing stage. In other words, similarity between the extracted feature and the reference features corresponding to each action class is calculated by using the Euclidean distance, and the class with the lowest distance value is determined as the final result.

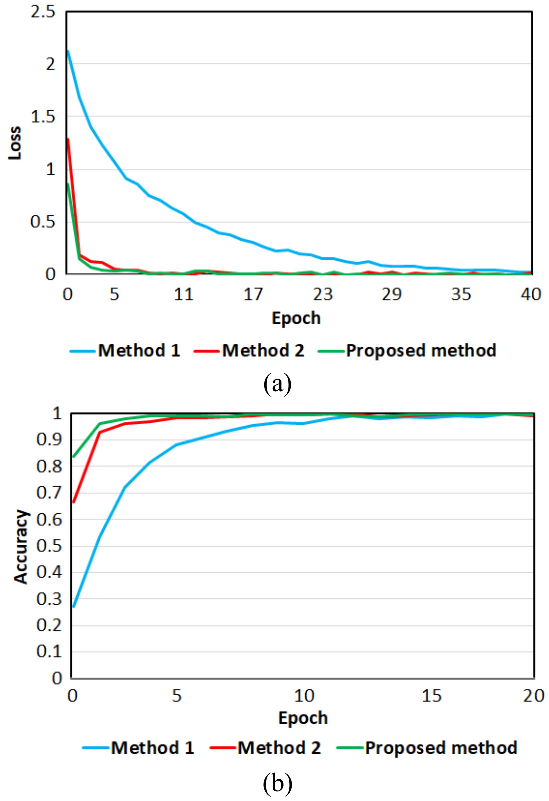


FIGURE 10. Training loss and accuracy curves of CNN-LSTM. (a) Loss curves of the three methods; (b) accuracy curves of the three methods.

C. TESTING RESULTS

1) ABLATION STUDY ON JOINT-GAN

In this section, comparative experiments were performed in order to compare the accuracies of joint and skeleton generation by using the original 1-channel grayscale thermal image (Method 1), the converted 3-channel color image (Method 2), and the skeleton and joint image obtained from both the 1-channel and 3-channel images by our method (Proposed Method). The same Joint-GAN and 5 sub-datasets of DTh-DB and DI&V-DB were used for all methods. For the accuracy measurement, the resulting image was compared with the ground-truth joint and skeleton image based on similarity. Accuracy was measured using three metrics as in Equations (1)–(3).

$$\mathrm{MSE} = \frac{\left(\sqrt{\sum_{\mathrm{y}=1}^{H}\sum_{\mathrm{x}=1}^{W}\left(T\left(\mathrm{x},\mathrm{y}\right)-O\left(\mathrm{x},\mathrm{y}\right)\right)^{2}}\right)^{2}}{\mathrm{MN}} \quad (1)$$

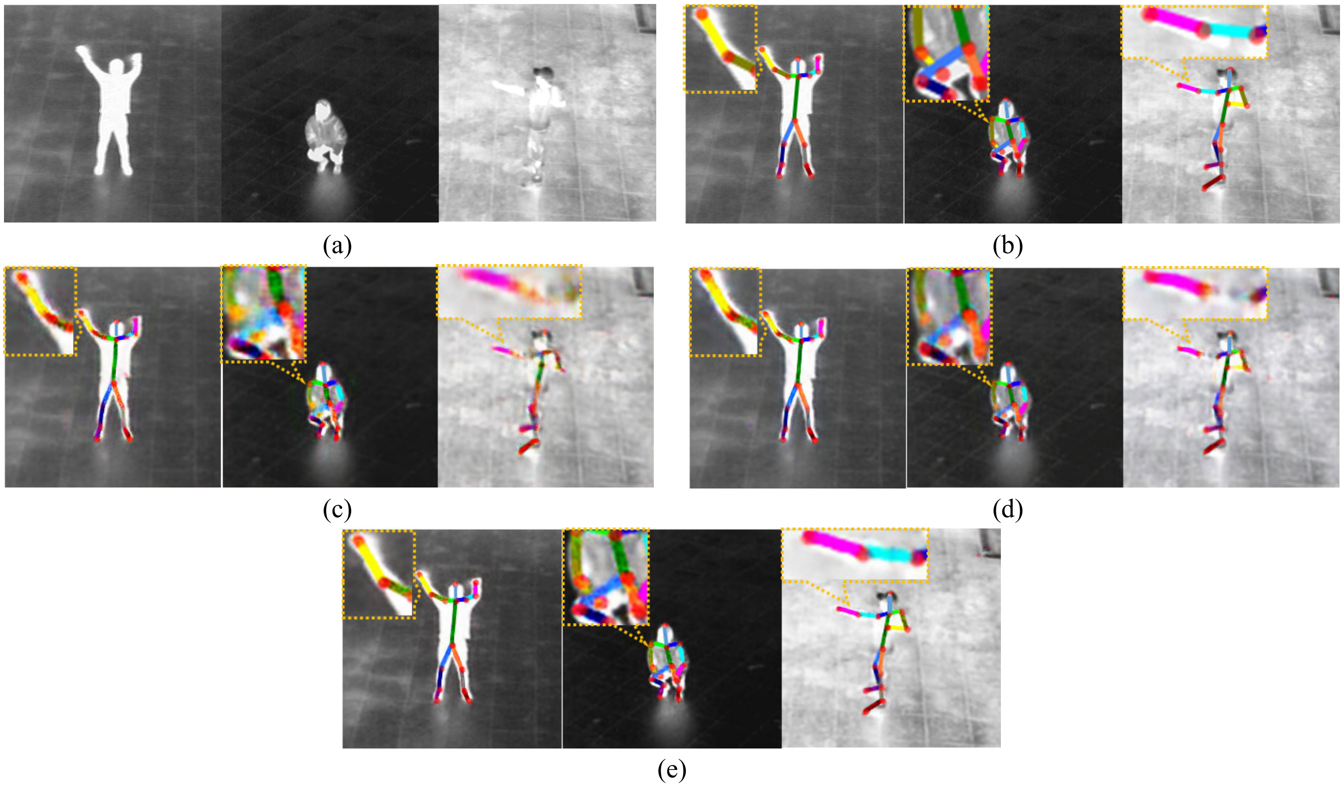


FIGURE 11. Results obtained in the ablation study. (a) Original images; (b) ground-truth images; (c) results by the Method 1; (d) results by the Method 2; (e) results by the proposed method.

TABLE 10. Accuracies obtained in the ablation study using five databases.

Sub-dataset	Method 1		Method 2		Proposed method	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1	22.581	0.9494	22.592	0.9490	22.777	0.9513
2	23.141	0.9532	23.342	0.9541	23.733	0.9642
3	21.326	0.9412	21.357	0.9458	21.561	0.9595
4	24.972	0.9597	24.981	0.9598	24.992	0.9661
5	22.755	0.9483	22.756	0.9485	22.781	0.9517
Average	22.955	0.95036	23.0056	0.95144	**23.1688**	**0.95856**

$$\mathrm{PSNR} = 10log_{10}\left(\frac{255^2}{\mathrm{MSE}}\right) \quad (2)$$

$$\mathrm{SSIM} = \frac{(2\mu_O\mu_T + R1)(2\sigma_{OT} + R2)}{(\mu_O^2 + \mu_T^2 + R1)(\sigma_O^2 + \sigma_T^2 + R2)} \quad (3)$$

In Equation (1), W and H represent image width and height, respectively. In Equations (1) and (3), O and T denote the output image and the target image respectively. In Equation (2), PSNR is the peak signal-to-noise ratio [60]. In the structural similarity index measure (SSIM) [61] equation, μ_T and σ_T represent the mean and standard deviation of the pixel values of a target image, respectively; μ_O and σ_O represent the mean and standard deviation of the pixel values of the output image, respectively. σ_{OT} is the covariance of the two images. $R1$ and $R2$ are positive constants that prevent the denominator from being zero. The higher values of PSNR and SSIM represent that two images are similar whereas the lower values represent that two images are less similar. The accuracy of the 3-channel thermal image-based method was higher than that of the 1-channel thermal image-based method. In addition, the proposed method, which combines both images, had the highest accuracy, as shown in Table 10 and Figure 11.

2) ABLATION STUDY ON CNN-LSTM

In this section, the accuracy of action recognition was compared by using the original 1-channel grayscale thermal image (Method 1), the converted 3-channel color image (Method 2), and the skeleton and joint image obtained from both the 1-channel and 3-channel images by our Joint-GAN (Proposed Method). The same CNN-LSTM network and five sub-datasets of DTh-DB and DI&V-DB were used for all methods, and the results are shown in Tables 11–14. In this

TABLE 11. **Confusion matrix of the human action recognition experiment results by Method 1 (unit: %).**

Recognized Actual	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting	Leaving	Approaching
Waving with two hands	97.47	0	1.69	0.84	0	0	0	0	0	0	0
Waving with one hand	0	100	0	0	0	0	0	0	0	0	0
Punching	1.05	3.13	95.82	0	0	0	0	0	0	0	0
Kicking	7.24	0	0	92.76	0	0	0	0	0	0	0
Lying	3.22	0	0	4.09	92.69	0	0	0	0	0	0
Walking	0.45	0	0	0	0.22	82.59	11.38	4.47	0.89	0	0
Running	0	0	3.92	3.43	0	11.28	79.41	0.98	0.98	0	0
Standing	0	0	0	0	0	0	0	100	0	0	0
Sitting	0	0	0	0	7.29	0	0	1.46	91.25	0	0
Leaving	0	0	0	0	0	0	0	0	0	100	0
Approaching	0	0	0	0	0	0	0	0	0	0	100

TABLE 12. **Confusion matrix of the human action recognition experiment results by Method 2 (unit: %).**

Recognized Actual	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting	Leaving	Approaching
Waving with two hands	100	0	0	0	0	0	0	0	0	0	0
Waving with one hand	0	97.14	2.86	0	0	0	0	0	0	0	0
Punching	10.14	1.94	87.92	0	0	0	0	0	0	0	0
Kicking	0	0	0	100	0s	0	0	0	0	0	0
Lying	0	0	0	0	95.48	0	0	4.52	0	0	0
Walking	2	0	0	0.23	0	85.05	7.15	1.56	4.01	0	0
Running	0	0	0.49	0.49	0	2.94	89.7	0	6.38	0	0
Standing	0	0	0	0	0	0	2.56	96.35	1.09	0	0
Sitting	0	0	0	0	7.88	3.5	0	0	88.62	0	0
Leaving	0	0	0	0	0	0	0	0	0	100	0
Approaching	0	0	0	0	0	0	0	0	0	0	100

experiment, the resulting image obtained by the Joint-GAN was not used in Method 1 and Method 2 but the original 1-channel grayscale image and converted 3-channel color image were used as inputs to the CNN-LSTM in Method 1 and Method 2, respectively. Metrics for comparing action recognition accuracy are true positive rate (TPR), positive predictive value (PPV), accuracy (ACC), and F1 score, which are shown in Equations (4) ∼ (7). In these equations, #TP, #FN, #FP, and #TN mean the numbers of true positive, false negative, false positive, and true negative, respectively.

$$\mathrm{TPR} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}} \tag{4}$$

$$\mathrm{PPV} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP}} \tag{5}$$

$$\mathrm{ACC} = \frac{\#\mathrm{TP} + \#\mathrm{TN}}{\#\mathrm{TP} + \#\mathrm{TN} + \#\mathrm{FP} + \#\mathrm{FN}} \tag{6}$$

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} \tag{7}$$

It was found that the accuracy of the action recognition of the proposed method was higher than that of Methods 1 and 2, as shown in Tables 11–14.

As a result of the experiment, higher classification results were shown when extracting features from the LSTM layer and classifying them by Euclidean distance than when using the softmax function during the testing stage as shown in Table 15.

3) ABLATION STUDY USING OPEN DATABASES

Additional experiments were performed using the CASIA C open dataset [53] to examine the applicability of the proposed methods in other environments. Methods 1 and 2 are the results of skeleton and joint extraction performed by using 1-channel thermal images and 3-channel thermal images, respectively, as shown in Table 16. In addition, Methods 1 and 2 are the results of action recognition performed by using 1-channel thermal images and 3-channel thermal images, as shown in Table 17. Moreover, the comparison of results of the skeleton and joint extraction methods is shown in Figure 12. It was found that the method proposed in this study showed the highest results of skeleton and joint extraction and action recognition, as shown in Tables 16, 17, and Figure 12.

4) COMPARISONS BETWEEN THE PROPOSED METHOD AND OTHER STATE-OF-THE-ART METHODS

The results of comparing the proposed method with other state-of-the-art methods are shown in this section. Conventional skeleton generation methods were compared with the

TABLE 13. Confusion matrix of the human action recognition experiment results by the proposed method (unit: %).

Recognized Actual	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting	Leaving	Approaching
Waving with two hands	96.63	0	3.37	0	0	0	0	0	0	0	0
Waving with one hand	0	100	0	0	0	0	0	0	0	0	0
Punching	0	0	100	0	0	0	0	0	0	0	0
Kicking	0	5.88	0	94.12	0	0	0	0	0	0	0
Lying	0	0	0	0	95.06	2.36	0	0	2.58	0	0
Walking	0	0	0.66	0.22	0	89.74	8.26	0.23	0.89	0	0
Running	1.96	0	4.41	1.47	0	1.47	85.79	0	4.9	0	0
Standing	0	0	0	0	0	0	0	100	0	0	0
Sitting	0	0	0	0	0	0	0	0	100	0	0
Leaving	0	0	0	0	0	0	0	0	0	100	0
Approaching	0	0	0	0	0	0	0	0	0	0	100

TABLE 14. Accuracies of the human action recognition experiment for the three methods (unit: %).

Human actions	Method 1				Method 2				Proposed method			
	TPR	PPV	ACC	F1	TPR	PPV	ACC	F1	TPR	PPV	ACC	F1
Waving with two hands	97.46	85.24	98.65	90.94	100	75.48	97.75	86.02	96.62	98.28	99.65	97.44
Waving with one hand	100	93.00	99.39	96.37	97.13	95.42	99.39	96.26	100	95.55	99.62	97.72
Punching	95.82	98.17	98.83	96.97	87.91	98.49	97.36	92.90	100	97.10	99.41	98.52
Kicking	92.76	87.98	98.71	90.30	100	99.10	99.94	99.54	94.12	98.11	99.50	96.07
Lying	92.69	94.31	98.24	93.49	95.48	94.27	98.59	94.87	95.05	100	99.33	97.46
Walking	82.59	94.15	97.04	87.99	85.04	95.49	97.51	89.96	89.73	96.63	98.24	93.05
Running	79.41	76.06	97.28	77.69	89.71	79.91	98.04	84.52	85.78	82.55	98.07	84.13
Standing	100	95.30	99.21	97.59	96.35	94.96	98.59	95.65	100	99.82	99.97	99.90
Sitting	91.25	98.12	98.95	94.56	88.63	89.15	97.77	88.88	100	92.95	99.24	96.34
Leaving	100	100	100	100	100	100	100	100	100	100	100	100
Approaching	100	100	100	100	100	100	100	100	100	100	100	100
Average	93.82	92.94	98.75	93.26	94.57	92.93	98.63	93.51	**96.48**	**96.45**	**99.37**	**96.42**

TABLE 15. Comparison of accuracies obtained by using Euclidean distance and Softmax function using 3-channel (color) images.

Methods	TPR	PPV	ACC	F1
Softmax function	95.21	95.87	98.72	95.53
Euclidean distance	**96.48**	**96.45**	**99.37**	**96.42**

TABLE 16. Comparison of the skeleton and joint extraction methods (unit: %).

Methods	PSNR	SSIM
Method 1	19.1421	0.88159
Method 2	20.5143	0.90035
Proposed method	**23.0069**	**0.95014**

TABLE 17. Comparison of the human action recognition methods (unit: %).

Methods	TPR	PPV	ACC	F1
Method 1	92.61	91.56	96.93	92.08
Method 2	93.41	91.82	97.22	92.60
Proposed method	**96.01**	**96.21**	**99.12**	**96.10**

Joint-GAN-based skeleton and joint extraction method proposed in this study, as shown in Table 18. The Perceptual loss function-based network (PLN) method [62], the cycle consistent adversarial network (CycleGAN)-based method [63], and the fully convolutional network (FCN) method [64] were selected as the conventional methods for comparative experiments. In the additional experiments, we compared the proposed method with the joint detection methods [65], [66] by measuring TPR, PPV and average precision (AP) between ground truth joint points and detected ones. Additionally, conventional action recognition methods were compared with the CNN-LSTM-based method proposed in this study,

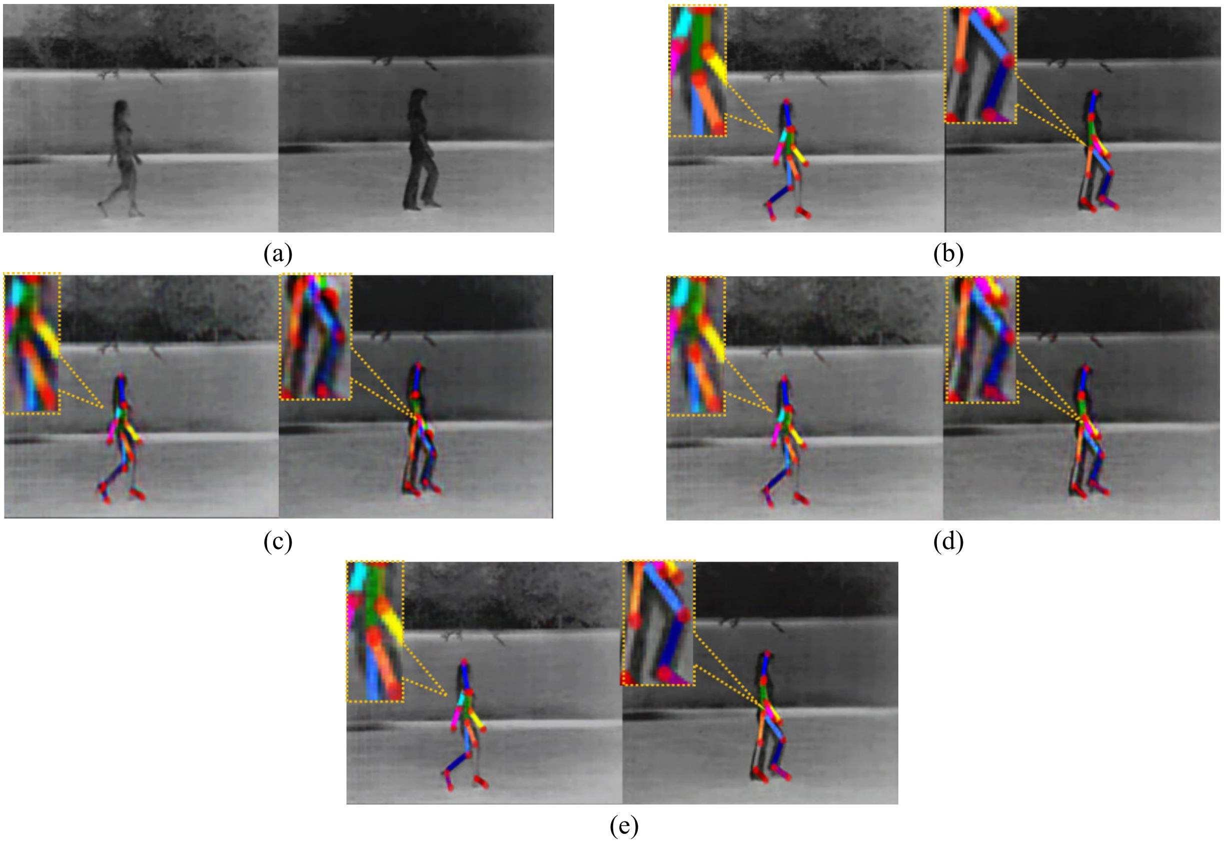


FIGURE 12. Results obtained in the ablation study using the CASIA C dataset. (a) Original images; (b) ground-truth images; (c) results by the Method 1; (d) results by the Method 2; (e) results by the proposed method.

TABLE 18. Comparison of the skeleton generation methods (unit: %).

Methods	PSNR	SSIM
PLN [62]	20.543	0.89519
CycleGAN [63]	22.132	0.93513
FCN [64]	20.991	0.90121
Proposed method	**23.1688**	**0.95856**

TABLE 19. Comparison of the skeleton generation methods (unit: %).

Methods	TPR	PPV	AP
Jaouedi et al. [65]	0.8389	0.8619	84.71
Pham et al. [66]	0.8531	0.8814	86.63
Proposed method	**0.8962**	**0.9121**	**89.35**

as shown in Table 20. The AP [1] represents the area under the precision-recall curve as shown in Equation (8).

$$\mathbf{AP} = \int_0^1 \boldsymbol{p}(\boldsymbol{r})\boldsymbol{dr} \tag{8}$$

where $\boldsymbol{p}(\boldsymbol{r})$ represents the graph of positive predictive value (PPV) according to true positive rate (TPR) ($\boldsymbol{r}$), where both PPV and TPR are shown between 0 and 1.

TABLE 20. Comparison of the human action recognition methods (unit: %).

	Methods	TPR	PPV	ACC	F1
Traditional	Fourier descriptor-based [67]	72.13	68.22	83.12	70.12
	GEI-based [7]	79.59	86.21	85.94	82.76
	Convexity defect-based [8]	78.45	82.12	84.29	80.24
Deep	Jaouedi et al. [65]	93.53	95.91	97.54	94.70
	Pham et al. [66]	92.84	93.47	97.02	93.15
	Proposed method	**96.48**	**96.45**	**99.37**	**96.42**

For human action recognition, comparison experiments were performed using traditional algorithm-based methods such as Fourier descriptor-based method [67], the gait energy image (GEI)-based method [7], and the convexity defect-based method [8], and deep learning algorithm-based methods such as Jaouedi *et al.* [65], Pham *et al.* [66].

In addition, the features used in the traditional algorithm-based methods were used as inputs to the CNN-LSTM proposed in this study to perform additional comparative experiments, as shown in Table 21. Moreover, the results

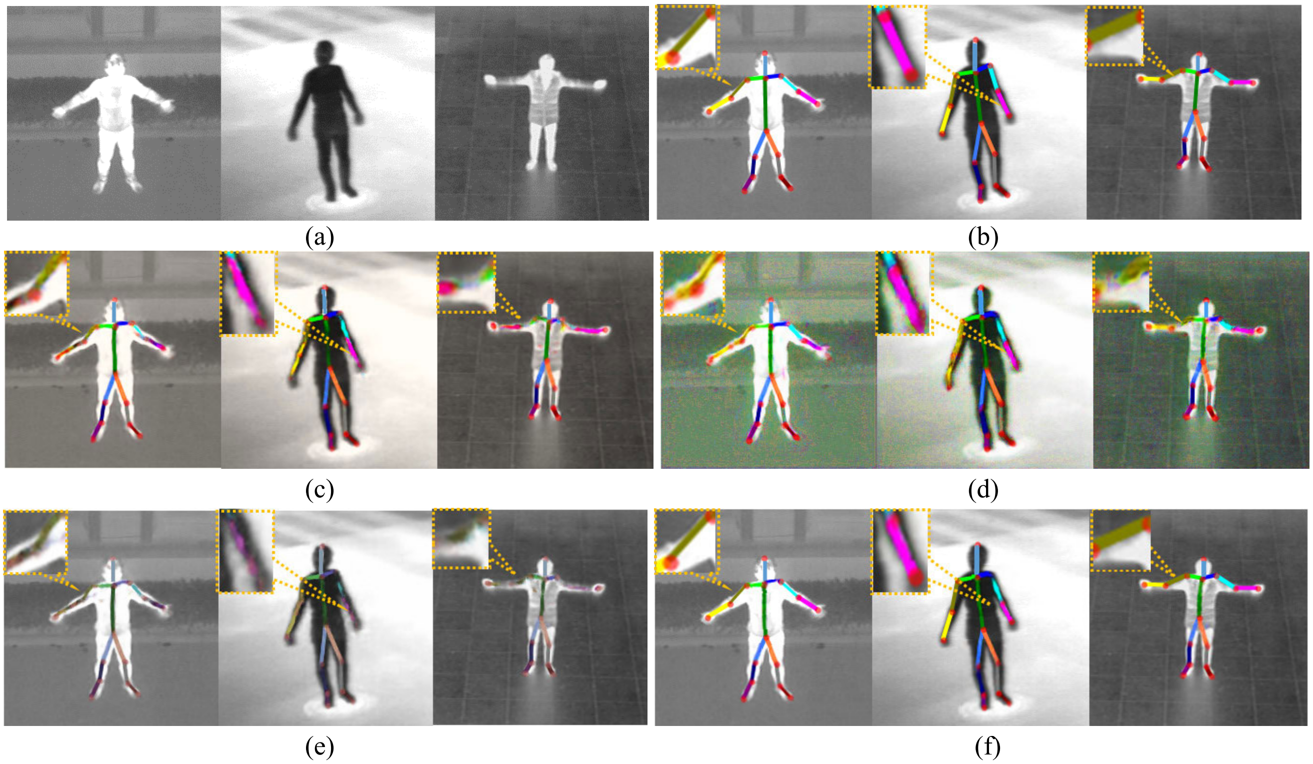


FIGURE 13. Comparison of results between the Joint-GAN and the previous methods. (a) Original images; (b) ground-truth images; (c) results using the PLN method [62]; (d) results using the CycleGAN method [63]; (e) results using the FCN method [64]; (f) results using the proposed Joint-GAN method.

TABLE 21. Comparison of the human action recognition methods. The results were obtained using the features extracted by previous methods and the proposed CNN-LSTM (unit: %).

Methods	TPR	PPV	ACC	F1
Fourier descriptor-based [67] + CNN-LSTM	89.41	87.21	89.18	88.29
GEI-based [7] + CNN-LSTM	91.41	92.21	95.18	91.80
Convexity defect-based [8] + CNN-LSTM	90.92	91.54	94.02	91.22
Proposed method	**96.48**	**96.45**	**99.37**	**96.42**

TABLE 22. Processing time of the proposed method (unit: ms).

Sub-part	Processing time
Preprocessing	0.96
Joint-GAN	15.88
Postprocessing	2.07
CNN-LSTM	87.73
Total	**106.64**

obtained by the conventional skeleton generation methods were compared with the results from the images obtained by the Joint-GAN method of this study, as shown in Figure 13. The method proposed in this study showed better results than the state-of-the-art methods, as shown in Tables 17∼19 and Figure 13.

5) PROCESSING TIME

In Table 22, the processing time of each sub-part of the proposed method (Figure 2) is presented. As shown in Table 22, the processing time of the CNN-LSTM is higher than other sub-parts because the CNN-LSTM is iterated by 30 times to produce final result as shown in Table 9. The total frame rate of the proposed method is 9.38 frames per second (1000/106.64). Thus, we can confirm that the proposed method can run fast enough to perform both skeleton generation and action recognition.

V. CONCLUSION

In this study, a method to extract joints and skeleton information was proposed by converting the original 1-channel thermal image into a 3-channel thermal image, combining these images, and using them as an input for the proposed Joint-GAN. In addition, a method for recognizing various human actions based on CNN-LSTM was proposed using the extracted joints and skeleton information. Comparative experiments were performed using original 1-channel thermal images and converted 3-channel thermal images to evaluate the performance of the proposed method. According to the experimental results using the self-collected DTh-DB and DI&V-DB databases, together with the CASIA C open database, the Joint-GAN and CNN-LSTM methods proposed in this study showed higher accuracy than other state-of-the-art methods. In the proposed Joint-GAN, we assigned different colors to skeleton parts to distinguish human body parts. By doing so, we can provide more information for action recognition. However, we encountered the error cases caused by the assigned colors. For example, as shown in Figures 11 (e) and 12 (e), both light and dark green colors show lower detection accuracies whereas blue, pink, and yellow colors show higher accuracies compared to other colors. This reveals that the different colors play different roles in the image-to-image translation method. Furthermore, this error affects the performance of the proposed action recognition method.

In future work, enhanced image-to-image translation method would be researched irrespective of the colors of skeleton. In addition, the performance of the Joint-GAN proposed in this study would be evaluated by applying it to the visible and near-infrared light images.

REFERENCES

[1] P. Zhang and W. Su, "Statistical inference on recall, precision and average precision under random selection," in *Proc. 9th Int. Conf. Fuzzy Syst. Knowl. Discovery*, Sichuan, China, May 2012, pp. 1348–1352.

[2] G. Batchuluun, H. S. Yoon, D. T. Nguyen, T. D. Pham, and K. R. Park, "A study on the elimination of thermal reflections," *IEEE Access*, vol. 7, pp. 174597–174611, 2019.

[3] Digital Media Lab. (2020). *Dongguk Joint-GAN and CNN-LSTM for Action Recognition*. [Online]. Available: http://dm.dgu.edu/link.html

[4] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments," *Future Gener. Comput. Syst.*, vol. 96, pp. 386–397, Jul. 2019.

[5] A. Ullah, K. Muhammad, T. Hussain, M. Lee, and S. W. Baik, "Deep LSTM-based sequence learning approaches for action and activity recognition," in *Deep Learning in Computer Vision*, 1st ed. Boca Raton, FL, USA: CRC Press, 2020, pp. 127–150. [Online]. Available: https://www.taylorfrancis.com/chapters/deep-lstm-based-sequence-learning-approaches-action-activity-recognition-amin-ullah-khan-muhammad-tanveer-hussain-miyoung-lee-sung-wook-baik/e/10.1201/9781351003827-5

[6] A. Ullah, K. Muhammad, T. Hussain, and S. W. Baik, "Conflux LSTMs network: A novel approach for multi-view action recognition," *Neurocomputing*, Dec. 2020, doi: 10.1016/j.neucom.2019.12.151.

[7] L. Chunli and W. Kejun, "A behavior classification based on enhanced gait energy image," in *Proc. Int. Conf. Netw. Digit. Soc.*, Wenzhou, China, May 2010, pp. 589–592.

[8] M. M. Youssef, "Hull convexity defect features for human action recognition," Ph.D. dissertation, Univ. Dayton, Dayton, OH, USA, Aug. 2011.

[9] B. Brattoli, U. Buchler, A.-S. Wahl, M. E. Schwab, and B. Ommer, "LSTM self-supervision for detailed behavior analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3747–3756.

[10] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[11] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 807–811, Mar. 2018.

[12] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017.

[13] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1110–1118.

[14] R. Cui, A. Zhu, S. Zhang, and G. Hua, "Multi-source learning for skeleton -based action recognition using deep LSTM networks," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Phoenix, AZ, USA, Aug. 2018, pp. 3697–3703.

[15] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer LSTM networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Santa Rosa, CA, USA, Mar. 2017, pp. 148–157.

[16] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 816–833.

[17] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," Nov. 2016, *arXiv:1611.06067*. [Online]. Available: http://arxiv.org/abs/1611.06067

[18] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based action recognition using LSTM and CNN," Jul. 2017, *arXiv:1707.02356*. [Online]. Available: http://arxiv.org/abs/1707.02356

[19] H. Eum, J. Lee, C. Yoon, and M. Park, "Human action recognition for night vision using temporal templates with infrared thermal camera," in *Proc. 10th Int. Conf. Ubiquitous Robots Ambient Intell. (URAI)*, Jeju Island, South Korea, Oct./Nov. 2013, pp. 617–621.

[20] J. Han and B. Bhanu, "Human activity recognition in thermal infrared imagery," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)-Workshops*, San Diego, CA, USA, Jun. 2005, pp. 17–24.

[21] G. Batchuluun, Y. G. Kim, J. H. Kim, H. G. Hong, and K. R. Park, "Robust behavior recognition in intelligent surveillance environments," *Sensors*, vol. 16, no. 7, pp. 1–23, 2016.

[22] G. Batchuluun, J. H. Kim, H. G. Hong, J. K. Kang, and K. R. Park, "Fuzzy system based human behavior recognition by combining behavior prediction and recognition," *Expert Syst. Appl.*, vol. 81, pp. 108–133, Sep. 2017.

[23] G. Batchuluun, D. T. Nguyen, T. D. Pham, C. Park, and K. R. Park, "Action recognition from thermal videos," *IEEE Access*, vol. 7, pp. 103893–103917, 2019.

[24] C. W. Niblack, D. W. Capson, and P. B. Gibbons, "Generating skeletons and centerlines from the medial axis transform," in *Proc. 10th Int. Conf. Pattern Recognit.*, Atlantic City, NJ, USA, Aug. 1990, pp. 881–885.

[25] W. Niblack, P. B. Gibbons, and D. Capson, "Generating connected skeletons for exact and approximate reconstruction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Champaign, IL, USA, Jun. 1992, pp. 826–828.

[26] C. K. Lee and Y. W. Pang, "One-pixel width image skeleton generation using mathematical morphology," in *Proc. IEEE Winter Workshop Nonlinear Digit. Signal Process.*, Tampere, Finland, Jan. 1993, pp. 6.1-7.1–6.1-7.5.

[27] T.-L. Liu, D. Geiger, and A. L. Yuille, "Segmenting by seeking the symmetry axis," in *Proc. 14th Int. Conf. Pattern Recognit.*, Brisbane, QLD, Australia, Aug. 1998, pp. 994–998.

[28] J.-H. Jang and K.-S. Hong, "A pseudo-distance map for the segmentation-free skeletonization of gray-scale images," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, Vancouver, BC, Canada, Jul. 2001, pp. 18–23.

[29] A. Nedzved, S. Ablameyko, and S. Uchida, "Gray-scale thinning by using a pseudo-distance map," in *Proc. 18th Int. Conf. Pattern Recognit.*, Hong Kong, Aug. 2006, pp. 239–242.

[30] Z. Yu and C. Bajaj, "A segmentation-free approach for skeletonization of gray-scale images via anisotropic vector diffusion," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, USA, Jun./Jul. 2004, p. 1.

[31] K. R. Jerripothula, J. Cai, J. Lu, and J. Yuan, "Object co-skeletonization with co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3881–3889.

[32] O. Panichev and A. Voloshyna, "U-Net based convolutional neural network for skeleton extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 1186–1189.

[33] C. Liu, D. Luo, Y. Zhang, W. Ke, F. Wan, and Q. Ye, "Parametric skeleton generation via Gaussian mixture models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Long Beach, CA, USA, Jun. 2019, pp. 1167–1171.

[34] Y. Wang, Y. Xu, S. Tsogkas, X. Bai, S. Dickinson, and K. Siddiqi, "DeepFlux for skeletons in the wild," Nov. 2018, *arXiv:1811.12608v1*. [Online]. Available: https://arxiv.org/abs/1811.12608v1

[35] W. Shen, K. Zhao, Y. Jiang, Y. Wang, Z. Zhang, and X. Bai, "Object skeleton extraction in natural images by fusing scale-associated deep side outputs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 222–230.

[36] W. Shen, K. Zhao, Y. Jiang, Y. Wang, X. Bai, and A. Yuille, "DeepSkeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5298–5311, Nov. 2017.

[37] W. Ke, J. Chen, J. Jiao, G. Zhao, and Q. Ye, "SRN: Side-output residual network for object symmetry detection in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 302–310.

[38] S. A. Bhisikar and S. N. Kale, "Automatic joint detection and measurement of joint space width in arthritis," in *Proc. IEEE Int. Conf. Adv. Electron., Commun. Comput. Technol. (ICAECCT)*, Pune, India, Dec. 2016, pp. 429–432.

[39] Y. Huo, K. L. Vincken, M. A. Viergever, and F. P. Lafeber, "Automatic joint detection in rheumatoid arthritis hand radiographs," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, San Francisco, CA, USA, Apr. 2013, pp. 125–128.

[40] Y. Zhao, J. He, H. Cheng, and Z. Liu, "A 2.5D thinning algorithm for human skeleton extraction from a single depth image," in *Proc. Chin. Automat. Congr. (CAC)*, Hangzhou, China, Nov. 2019, pp. 3330–3335.

[41] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3041–3048.

[42] S. Transue, P. Nguyen, T. Vu, and M.-H. Choi, "Thermal-depth fusion for occluded body skeletal posture estimation," in *Proc. IEEE/ACM Int. Conf. Connected Health, Appl., Syst. Eng. Technol. (CHASE)*, Philadelphia, PA, USA, Jul. 2017, pp. 167–176.

[43] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," Sep. 2014, *arXiv:1406.2984*. [Online]. Available: http://arxiv.org/abs/1406.2984

[44] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," Nov. 2016, *arXiv:1611.08050*. [Online]. Available: http://arxiv.org/abs/1611.08050

[45] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," Jan. 2016, *arXiv:1602.00134*. [Online]. Available: http://arxiv.org/abs/1602.00134

[46] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," Apr. 2015, *arXiv:1511.06645*. [Online]. Available: http://arxiv.org/abs/1511.06645

[47] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," in *Proc. Eur. Conf. Comput. Vis. Workshops*, Amsterdam, The Netherlands, Oct. 2016, pp. 627–642.

[48] Mathworks. (2020). *Colormap*. [Online]. Available: https://www.mathworks.com/help/matlab/ref/jet.html

[49] G. Batchuluun, Y. W. Lee, D. T. Nguyen, T. D. Pham, and K. R. Park, "Thermal image reconstruction using deep learning," *IEEE Access*, vol. 8, pp. 126839–126858, 2020.

[50] G. Batchuluun, H. S. Yoon, J. K. Kang, and K. R. Park, "Gait-based human identification by combining shallow convolutional neural network-stacked long short-term memory and deep convolutional neural network," *IEEE Access*, vol. 6, pp. 63164–63186, 2018.

[51] Digital Media Lab. (2019). *Dongguk Thermal Image Database (DTh-DB) and Dongguk Items & Vehicles Database (DI&V-DB)*. [Online]. Available: http://dm.dgu.edu/link.html

[52] G. Batchuluun, N. R. Baek, D. T. Nguyen, T. D. Pham, and K. R. Park, "Region-based removal of thermal reflection using pruned fully convolutional network," *IEEE Access*, vol. 8, pp. 75741–75760, 2020.

[53] D. Tan, K. Huang, S. Yu, and T. Tan, "Efficient night gait recognition based on template matching," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Hong Kong, Aug. 2006, pp. 1000–1003.

[54] NVIDIA Corporation. (2019). *NVIDIA Titan X*. [Online]. Available: https://www.nvidia.com/en-us/geforce/products/10series/titan-x-pascal/

[55] OpenCV. (2019). *OpenCV: Open Source Computer Vision*. [Online]. Available: http://opencv.org/

[56] Keras. (2019). *Keras: The Python Deep Learning Library*. [Online]. Available: https://keras.io/

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[58] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. London, U.K.: MIT Press, 2012.

[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–14.

[60] D. Salomon, *Data Compression: The Complete Reference*, 4th ed. New York, NY, USA: Springer-Verlag, 2006.

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[62] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," Mar. 2016, *arXiv:1603.08155*. [Online]. Available: http://arxiv.org/abs/1603.08155

[63] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," Mar. 2017, *arXiv:1703.10593*. [Online]. Available: http://arxiv.org/abs/1703.10593

[64] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," Mar. 2014, *arXiv:1411.4038*. [Online]. Available: http://arxiv.org/abs/1411.4038

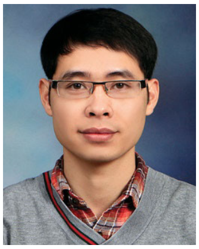
[65] N. Jaouedi, F. J. Perales, J. M. Buades, N. Boujnah, and M. S. Bouhlel, "Prediction of human activities based on a new structure of skeleton features and deep learning model," *Sensors*, vol. 20, no. 17, pp. 1–15, 2020.

[66] H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, S. A. Velastin, and P. Zegers, "A unified deep framework for joint 3D pose estimation and action recognition from a single RGB camera," *Sensors*, vol. 20, no. 7, pp. 1–15, 2020.

[67] N. M. Tahir, A. Hussain, S. A. Samad, H. Husain, and R. A. Rahman, "Human shape recognition using Fourier descriptor," *J. Elect. Electron. Syst. Res.*, vol. 2, pp. 19–25, Jun. 2009.

GANBAYAR BATCHULUUN received the B.S. degree in electronic engineering from Huree University, Ulaanbaatar, Mongolia, in 2011, the M.S. degree in electronic engineering from Pai Chai University, Daejeon, South Korea, in 2014, and the Ph.D. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2019. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University, since March 2019. He designed the entire system and wrote the original draft of article. His research interests include biometrics and pattern recognition.

JIN KYU KANG received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2016, where he is currently pursuing the combined course of M.S. and Ph.D. degrees in electronics and electrical engineering. He helped to perform the experiments and analysis. His research interests include biometrics and deep learning.



DAT TIEN NGUYEN received the B.S. degree in electronics and telecommunications from HUST, Hanoi, Vietnam, in 2009, and the Ph.D. degree in electronics and electrical engineering from Dongguk University, in 2015. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University, since March 2015. He has helped in the experiments and analysis of the study. His research interests include image processing, biometrics, and deep learning.

TUYEN DANH PHAM received the B.S. degree in electronics and telecommunications from HUST, Hanoi, Vietnam, in 2010, and the M.S. and Ph.D. degrees in electronics and electrical engineering from Dongguk University, in 2013 and 2017, respectively. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University, since March 2017. He supervised this research and revised the original article. His research interests include image processing, biometrics, and deep learning.

MUHAMMAD ARSALAN received the B.S. degree in computer engineering from COMSATS University Islamabad, Pakistan, in 2012, the M.S. degree in computer science from the National College of Business Administration & Economics (NCBAE), Lahore, Pakistan, in 2016, and the Ph.D. degree in electronics and electrical engineering from Dongguk University, in 2020. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University, since September 2020. He helped experiments and analysis. His research interests include computer vision and deep learning.

KANG RYOUNG PARK (Member, IEEE) received the B.S. and M.S. degrees in electronic engineering and the Ph.D. degree in electrical and computer engineering from Yonsei University, Seoul, South Korea, in 1994, 1996, and 2000, respectively. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University, since March 2013. He has assisted the study in experiments and analysis. His research interests include image processing and biometrics.

• • •