IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 19, 2019, accepted July 26, 2019, date of publication July 29, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2931804

Action Recognition From Thermal Videos

GANBAYAR BATCHULUUN, DAT TIEN NGUYEN, TUYEN DANH PHAM, CHANHUM PARK, AND KANG RYOUNG PARK

Division of Electronics and Electrical Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding author: Kang Ryoung Park (parkgr@dongguk.edu)

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT through the Basic Science Research Program under Grant NRF-2019R1F1A1041123, in part by the NRF funded by the Ministry of Education through the Basic Science Research Program under Grant NRF-2018R1D1A1B07041921, and in part by the NRF funded by the Korea Government, MSIT, under Grant NRF-2017R1C1B5074062.

ABSTRACT Human action recognition using a camera-based surveillance system remains a challenging task. In particular, action recognition is difficult when a human is not visible in an image captured in a dark environment. The existing studies have utilized near-infrared (NIR) and thermal cameras to solve this problem. Compared to NIR cameras, thermal cameras enable long- and short-distance objects to be visible without an additional illuminator. However, thermal cameras have two major disadvantages: a halo effect and a temperature similarity. A halo effect occurs around an object with a high temperature. In a human object, such a halo effect is similar to a shadow under the body area. It is more difficult to segment a human area from an image with a halo effect. Moreover, if the background and human object have similar temperatures, it becomes more difficult to segment the human area. These disadvantages influence not only the accuracy of the segmentation of the human area but also the performance of human action recognition. Unfortunately, no studies have considered these issues. To address these problems, this study proposes the cycle-consistent generative adversarial network (CycleGAN)-based methods for removing halo effects from thermal images and restoring the areas of the human bodies. In addition, this study also considered a method for creating a skeleton image from a thermal image to analyze body movements. To extract more spatial and temporal features from skeleton image sequences thus created, a method for human action recognition that combines a convolutional neural network (CNN) and long short-term memory (LSTM) was proposed. In an experiment using an open database (Dongguk activities & actions database (DA&A-DB2)), the proposed method demonstrated a better performance than the existing methods.

INDEX TERMS Human action recognition, halo effect, image restoration and skeleton generation, thermal camera, CNN stacked LSTM, and CycleGAN.

I. INTRODUCTION

Human action recognition using a camera-based surveillance system remains a challenging task. In particular, action recognition is difficult when a human is not visible in an image captured in a dark environment. Existing studies have utilized near-infrared (NIR) and long wavelength infrared (LWIR) cameras to solve this problem. LWIR cameras (thermal cameras) enable long- and short-distance objects to be visible without an additional illuminator, whereas NIR cameras need an additional illuminator to make only short-distance objects visible in a dark environment. A thermal camera makes an object visible in either a dark or bright environment by measuring temperatures ranging from −40 °C to +80 °C. This study utilized a thermal camera to acquire data on long-distance objects in either a dark or bright environment.

A halo effect and temperature similarity are two challenges in images obtained using a thermal camera. A halo effect occurs around an object with a high temperature. The higher temperature of the object, the larger the halo effect. Such a halo effect is similar to a shadow under the area of a human body in a thermal image. However, in a thermal image, the pixel value of the area of the object is quite similar to that of the area with a halo effect. By contrast, in a visible image, the pixel value of the shadow is lower than that of the area of the object. For this reason, segmenting a human area from an image with a halo effect is more difficult than segmenting a human area from an image with a shadow. The size and pixel

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

TABLE 1. Summary of comparisons between the proposed method and previous studies.

Category		Method	Advantage	Disadvantage
Without deep learning methods		Using visible light images [3–7, 12, 16, 18, 37, 39]	Large data acquisition, processing, and training are not required	- Performance is affected by shadows, variations in illumination, and human clothing of various colors - Objects are not visible in a dark environment
		Using thermal images [9, 13, 15]	- Large data acquisition, processing, and training are not required - Objects are visible in a dark environment	- Performance is affected by temperature variations
With deep learning methods	CNN-based	Using visible light images [19, 25, 26] Using depth images [20, 21] Using skeleton joint information [22–24]	- Good at extracting spatial information - Appropriate features are extracted in various environments and camera settings.	- Inevitable to higher loss of temporal information - Performance is affected by shadows, variations in illumination, and human clothing of various colors
	RNN-based	Using skeleton joint information [27]	- Good at extracting temporal information - Appropriate features are extracted in various environments and camera settings.	- Inevitable higher loss of spatial information, encounters vanishing and exploding problems - Performance is affected by shadows, variations in illumination, and human clothing of various colors
	LSTM	Using skeleton joint information [28–31]	- Good at extracting temporal information, while overcoming vanishing and exploding problem - Appropriate features are extracted in various environments and camera settings.	- Inevitable higher loss of spatial information - Performance is affected by shadows, variations in illumination, and human clothing of various colors
	CNN-LSTM	Using visible light images [32–34] Using skeleton joint information [35]	- Good at extracting spatial and temporal information while overcoming vanishing and exploding problem - Appropriate features are extracted in various environments and camera settings.	- Memory consumption and training time are expensive - Performance is affected by shadows, variations in illumination, and human clothing of various colors
		Using thermal images and skeleton information (**Proposed method**)	- Good at extracting spatial and temporal information while overcoming vanishing and exploding problem - Appropriate features are extracted in various environments and camera settings. - Extraction of robust features in various environments through CycleGAN-based image restoration and halo effect removal	Memory consumption and training time are expensive

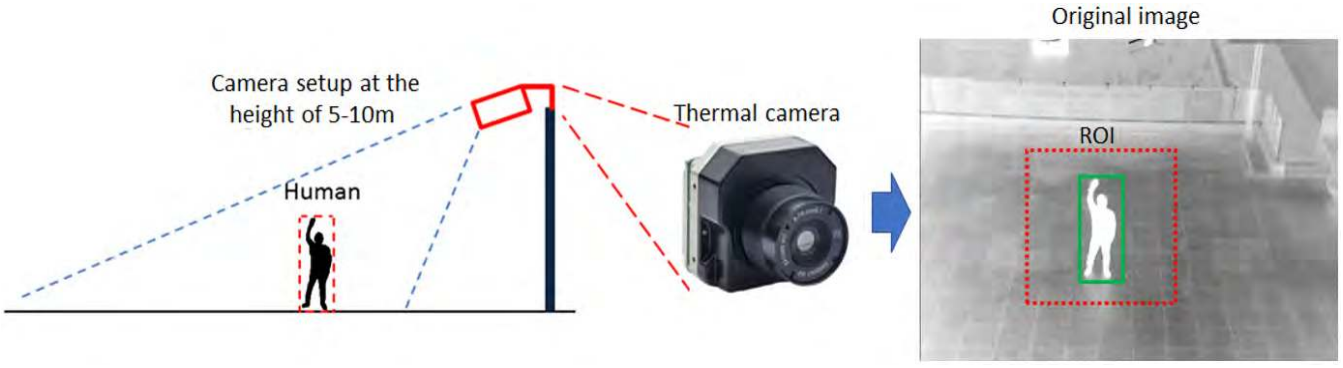

FIGURE 1. Example of camera setup and experiment environment.

value of a halo effect depend on the material of the ground upon which the object stands. In addition, because the area of a halo effect is usually connected with that of an object in a thermal image, it is difficult to segment the area of the human. When the background and human have similar temperatures, segmenting the area of the human becomes more difficult. To date, no studies have dealt with these issues.

This study proposes a new method to address these issues. The proposed method utilizes convolutional neural network (CNN)- and cycle-consistent generative adversarial network (CycleGAN)-based methods for removing halo effects from thermal images and restoring the areas of the human bodies. This study also examines a new skeleton generation method to analyze the body movement of a long-distance object for action recognition in a thermal image. Depending on the temperature of the environment, the pixel value of the entire body area of a long-distance object in a thermal image is frequently either 0 (black) or 255 (white). In either case, body joint features such as the ankles, knees, hips, wrists, elbows, and shoulders are barely visible and difficult to detect. To solve this problem, a deep learning-based method for converting original thermal images into skeleton images is examined.

No existing studies have utilized a deep learning-based method for recognizing the action of a long-distance object in a thermal image to address the above issues. This study attempts to extract more spatial and temporal features, and proposes a method for human action recognition that combines a CNN with long short-term memory (LSTM).

The remaining sections of this paper are organized as follows. Section II reviews the existing studies dealing with action recognition, skeleton generation, and deep learning. Section III examines the contributions of the proposed method. Section IV describes the proposed method in detail. Section V presents the experiment results and those of a

FIGURE 2. Overall flowchart of the proposed method.

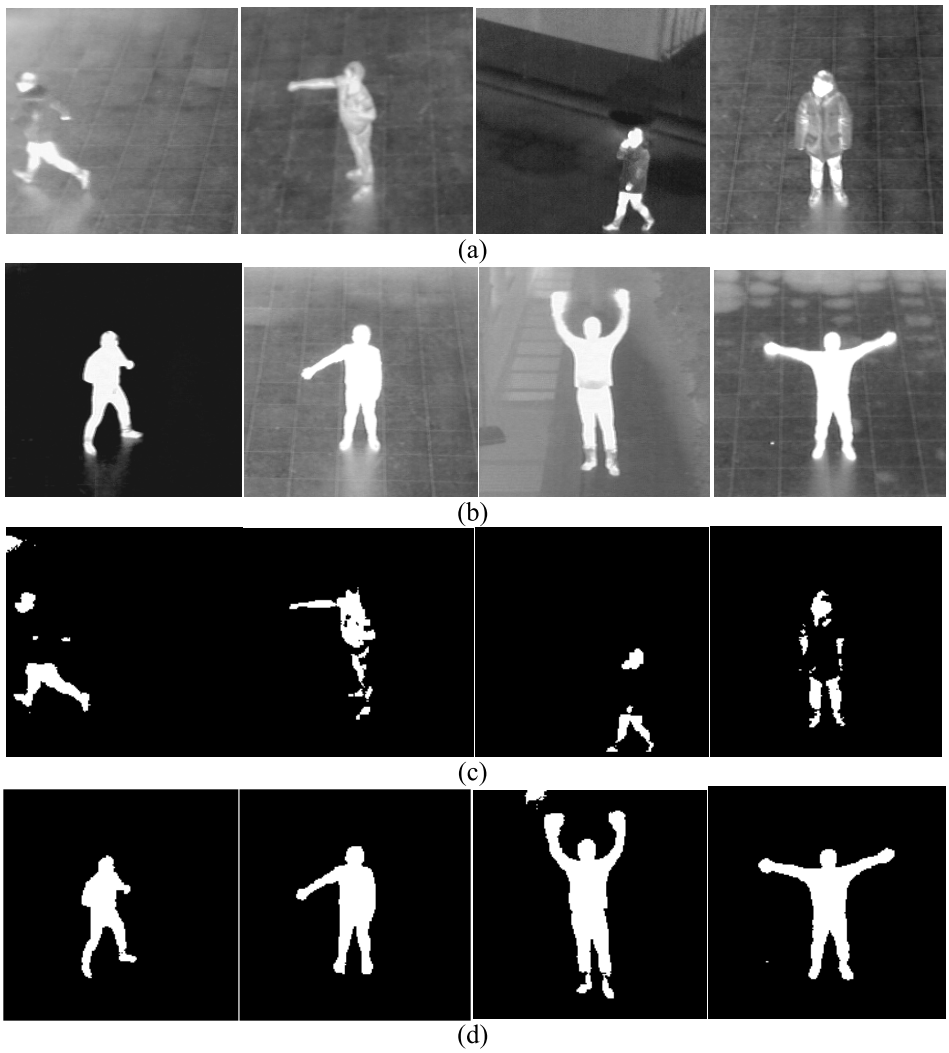

FIGURE 3. Example of captured thermal images and results of background subtraction method: (a), (b) thermal images, (c), (d) results of background subtraction using images in (a) and (b), respectively.

comparative experiment. Finally, Section VI provides some concluding remarks.

II. RELATED WORKS

Current methods of human action recognition can be classified mainly into two groups: deep learning-based methods and handcrafted feature-based methods without deep learning.

Among the existing studies on the latter group, the methods in [1] and [2] extract invariant Fourier descriptors for the scale and rotation from silhouette images and utilize those features for human action recognition through a support vector machine (SVM) and neural network (NN). However, although they are useful in expressing the shapes of objects, Fourier descriptors have difficulty expressing different actions with the same shape. A person standing has

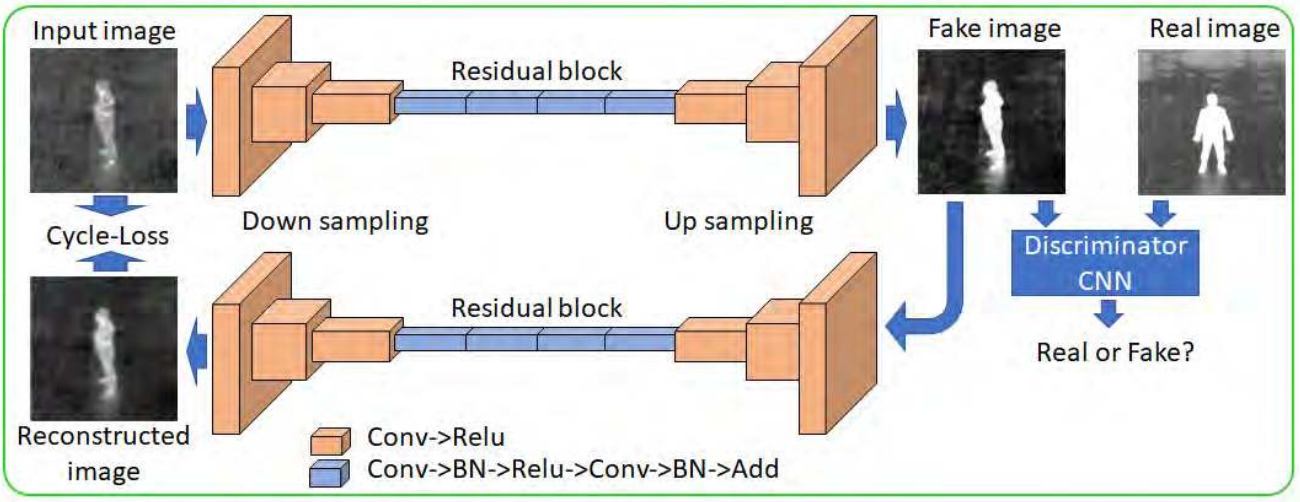

FIGURE 4. Image restoration using CycleGAN.

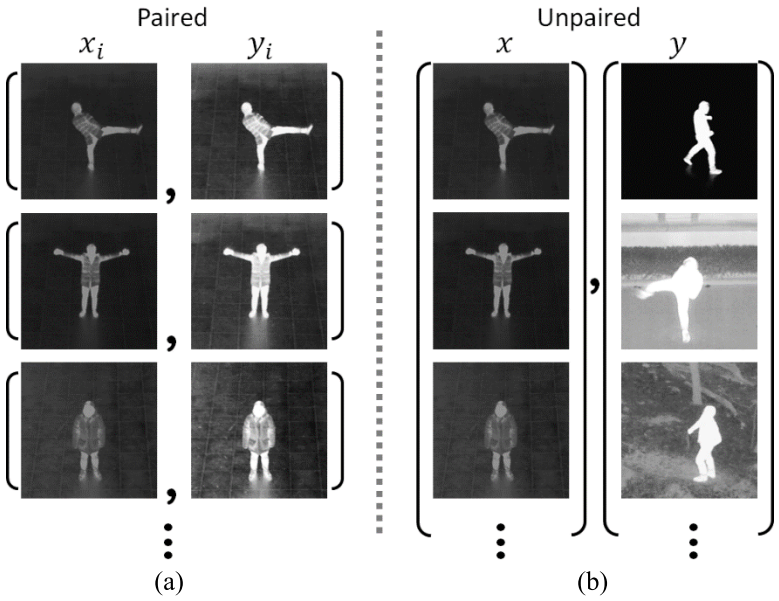

FIGURE 5. Example of paired and unpaired training data. Examples of (a) paired and (b) unpaired images.

the same shape as another person lying down but they are both conducting a different type of action. In [3], human action recognition is applied using a local descriptor-based scale invariant feature transform (SIFT) and Zernike moment features. However, it takes a long time to extract such features during the test phase. In [4]–[6], local spatiotemporal features are created by applying a corner detection method and attempting an action recognition using an SVM. Unfortunately, the background is not clear, and when the background includes different objects, the number of detected corners increases, thereby decreasing the accuracy of the human action recognition. In [7] and [8], motionlets and motion saliency methods are proposed, which demonstrate high accuracy only when an object has been clearly segmented from its background. In [4], [5], and [9], methods for extracting motion features using space-time interest points and a histogram of oriented gradients (HoG) are proposed. However, these methods require a long time to extract the features and are sensitive to noise and illumination. In [9]–[14], various handcrafted features such as a gait flow image (GFI), gait history image (GHI), motion history image (MHI), and accumulated motion image (AMI) are extracted to achieve human action recognition. When these features are extracted, body areas are detected in continuous images to obtain the gravity center points of each area. Based on the gravity center points, all areas were combined into a single image. However, the accuracy of the gravity center points was lowered based on the detection accuracy of the body areas and the background

TABLE 2. Detailed description of generator structure (Conv, ReLU, BN, and ResBlock indicate the convolutional layer, rectified linear unit, batch normalization layer, and residual block, respectively).

Group name	Layer type	Size of feature map (height × width × channel)	Number of filters	Filter size	Stride	Number of parameters
Input	Input layer	224 × 224 × 1				
Down sampling	Conv1	216 × 216 × 32	32	9 × 9	1 × 1	2,624
	ReLU1	216 × 216 × 32				
	Conv2	104 × 104 × 64	64	9 × 9	2 × 2	165,952
	ReLU2	104 × 104 × 64				
	Conv3	51 × 51 × 128	128	3 × 3	2 × 2	73,856
	ReLU3	51 × 51 × 128				
Residual block	ResBlock1	51 × 51 × 128				296,192
	ResBlock2	51 × 51 × 128				296,192
	ResBlock3	51 × 51 × 128				296,192
	ResBlock4	51 × 51 × 128				296,192
	ResBlock5	51 × 51 × 128				296,192
	ResBlock6	51 × 51 × 128				296,192
Up sampling	DeConv4	103 × 103 × 64	64	3 × 3	2 × 2	73,792
	ReLU4	103 × 103 × 64				
	DeConv5	208 × 208 × 32	32	4 × 4	2 × 2	32,800
	ReLU5	208 × 208 × 32				
	DeConv6	216 × 216 × 3	3	9 × 9	1 × 1	7,779
	ReLU6	216 × 216 × 3				
Output	DeConv7	224 × 224 × 1	1	9 × 9	1 × 1	244
	ReLU7	224 × 224 × 1				
	Output layer	224 × 224 × 1				
Total params: 2,134,199						

TABLE 3. Detailed description of residual block (Conv, ReLU, BN, and Add indicate the convolutional layer, rectified linear unit, batch normalization layer, and addition function, respectively).

Block	Layer type	Size of feature map (height × width × channel)	Number of filters	Filter size	Stride	Padding	Number of parameters
ResBlock	Conv	51 × 51 × 128	128	3 × 3	1 × 1	1 × 1	147,584
	BN	51 × 51 × 128					512
	ReLU	51 × 51 × 128					
	Conv	51 × 51 × 128	128	3 × 3	1 × 1	1 × 1	147,584
	BN	51 × 51 × 128					512
	Add	51 × 51 × 128					
Total params: 296,192							

noise, which resulted in a decrease in extraction accuracy for handcrafted features. In [15], a method for achieving faint action recognition is considered, which utilizes information of the width and height of the human areas detected in thermal images. In [16], human action recognition is conducted based on a convexity defect feature point. However, the accuracy of the action recognition was not satisfactory because many inaccurate feature points were detected owing to the background noise. In addition, it takes a long time to calculate the contours, polygons, convex hulls, and convexity defects in each frame. In [17], the gait energy image (GEI) based ethnicity determination method is examined. In [18], action recognition is conducted by extracting point-cloud features from silhouette sequences. All studies mentioned thus far were conducted based on handcrafted features, and deep learning-based human action recognition has been attempted in the following ways.

In [19], two CNNs are used, namely, one utilizing optical flow features and the other utilizing the original images as inputs for action recognition. In [20] and [21]; [22]–[24]; and [19], [25], and [26] depth images, joint map data, and visible light images are used, respectively, for action recognition. Such input data contain many spatial features but lack temporal features. To solve this problem, the study in [27] utilized skeleton information as the input of a recurrent neural network (RNN) for action recognition. However, an RNN causes a vanishing and exploding problem. For example, as the length of the sequential input features increases, important information may disappear, or trivial information may be accumulated. In [28]–[30], attempts were made to address this problem by proposing LSTM network-based action recognition methods using skeleton information. The LSTM-based methods use input, output, and forget gait functions to solve the vanishing and exploding problem. For action recognition, the study in [31] used joint distance maps as the input of CNN and skeleton joint information as the input of LSTM to extract spatial and temporal information, and combined their output scores. In [32]–[34],

TABLE 4. Detailed description of structure of discriminator CNN (Conv, LReLU, and InsNorm indicate the convolutional layer, leaky rectified linear unit, and instance normalization layers, respectively).

Layer number	Layer type	Size of feature map (height × width × channel)	Number of filters	Filter size	Stride	Padding	Number of parameters
0	Input layer	224 × 224 × 1					0
1	Conv1_1	112 × 112 × 32	32	4 × 4	2 × 2	1 × 1	544
2	LReLU1_1	112 × 112 × 32					
3	Conv1_2	56 × 56 × 64	64	4 × 4	2 × 2	1 × 1	32,832
4	LReLU1_2	56 × 56 × 64					
5	InsNorm_1	56 × 56 × 64					2
6	Conv2_1	28 × 28 × 128	128	4 × 4	2 × 2	1 × 1	131,200
7	ReLU2_1	28 × 28 × 128					
8	InsNorm_2	28 × 28 × 128					2
9	Conv3_1	14 × 14 × 256	256	4 × 4	2 × 2	1 × 1	524,544
10	ReLU3_1	14 × 14 × 256					
11	InsNorm_3	14 × 14 × 256					2
12	Conv4_1	7 × 7 × 384	384	4 × 4	2 × 2	1 × 1	1,573,248
13	ReLU4_1	7 × 7 × 384					
14	InsNorm_4	7 × 7 × 384					2
15	Conv5_1	7 × 7 × 1	1	4 × 4	1 × 1	Unknown	6,145
16	Output layer	7 × 7 × 1					
Total number of parameters: 2,268,521							

a simultaneous learning method connecting a CNN and an LSTM is proposed.

However, there have been no recognition methods using a CNN and an LSTM for the various actions of long-distance objects. This study proposes a method for recognizing various actions including waving with one hand, waving with two hands, punching, kicking, sitting, standing, walking, running, lying down, leaving, and approaching. The proposed CNN-LSTM structure is interconnected, and sequential learning is conducted using input images. There are only a few existing studies on human action recognition using a thermal camera. Some of the existing studies [9], [13], [15] did not utilize a deep learning algorithm such as CNN-LSTM. The present study proposes a thermal camera-based method for recognizing various actions of a long-distance object under both dark and bright environments using a deep learning algorithm.

In addition, this study also proposes a human action recognition method that generates and utilizes skeleton images from thermal images. The existing action recognition methods of the studies in [22]–[24], [27], [29], [31], and [35], which are based on skeleton information, utilized skeletons generated beforehand. There are already existing methods for extracting skeleton information from depth images [36]–[39], visible light images [40], [41], and thermal images [42]. However, no methods that can generate skeleton images directly from thermal images have been proposed. There are also no methods for generating a skeleton image of a long-distance object from thermal images obtained under various environments. To improve the performance of the proposed human action recognition method, this study examined CycleGAN-based methods of image restoration and the removal of a halo effect using thermal images obtained under various environments.

Table 1 shows a brief summary of related studies, which are compared with the proposed method.

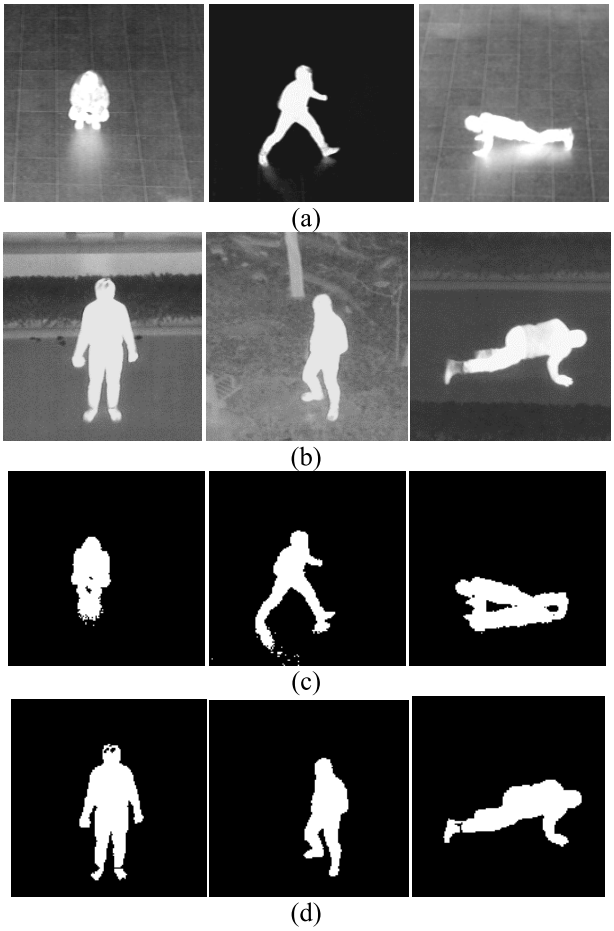

FIGURE 6. Example of captured thermal images and results of background subtraction method: (a), (b) thermal images and (c), (d) results of background subtraction using images in (a) and (b), respectively.

III. CONTRIBUTIONS

Our method is a novel approach compared to previous studies in the following ways:

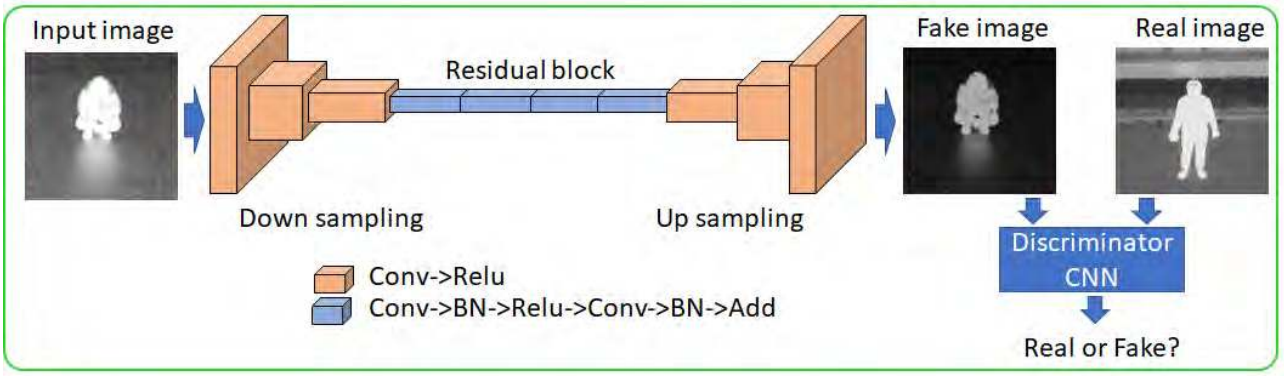

FIGURE 7. Halo effect removal using CycleGAN.

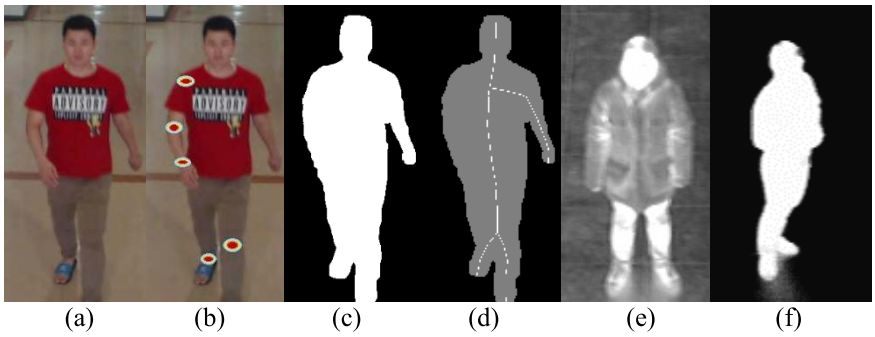

FIGURE 8. Examples of previous skeleton generation methods and captured thermal images. (a) A visible light image, (b) an example of joint detection, (c) a binary image of (a), (d) an example of skeleton generation using the image in (c), and (e), (f) captured thermal images.

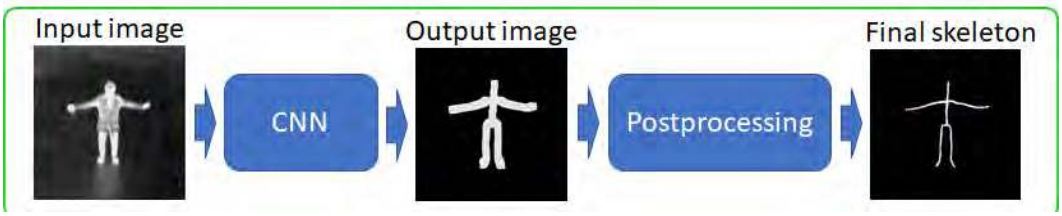

FIGURE 9. Skeleton generation using the proposed method.

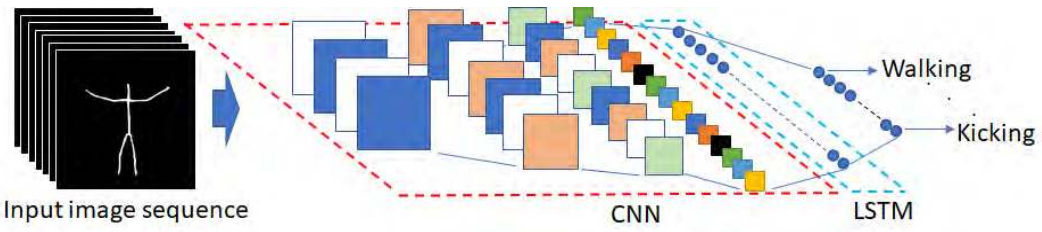

FIGURE 10. Human action recognition using CNN-LSTM.

- To date, no CNN-LSTM based methods have been proposed to recognize various actions of a long-distance object in thermal images, such as waving with one hand, waving with two hands, punching, kicking, sitting, standing, walking, running, lying down, leaving, and approaching. Accordingly, this study proposes a CNN-LSTM-based method for recognizing various actions of a long-distance object in thermal images.

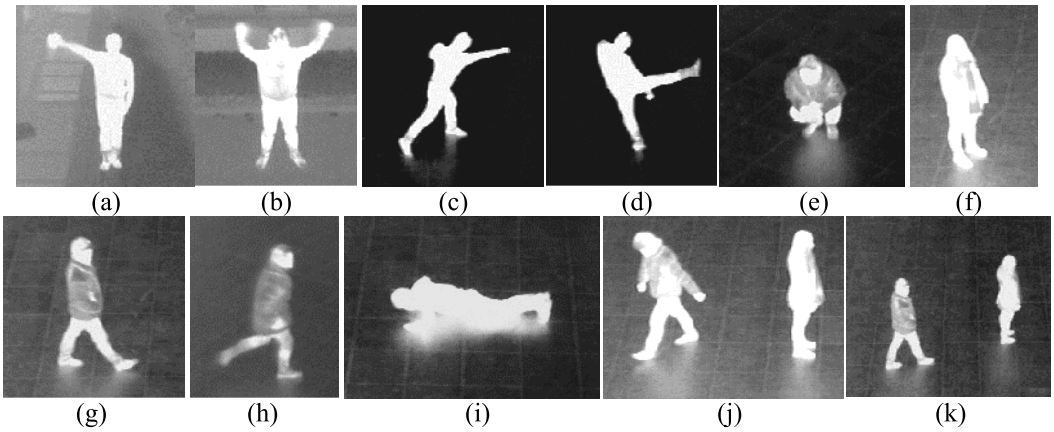

FIGURE 11. Example of images of human actions: (a) one hand waving, (b) two hands waving, (c) punching, (d) kicking, (e) sitting, (f) standing, (g) walking, (h) running, (i) lying down, (j) leaving, and (k) approaching.

TABLE 5. Detailed description of structure of the CNN-LSTM (Conv, Pool, ReLU, and Fc indicate the convolutional layer, max pooling, rectified linear unit, and fully connected layers, respectively. Nine classes are used).

Layer number	Layer type	Size of feature map (length × height × width × channel)	Number of filters	Filter size	Stride	Padding	Number of parameters
0	Input layer	5 × 224 × 224 × 3					0
1	Conv1_1	5 × 222 × 222 × 64	64	3 × 3	1 × 1	0 × 0	1,792
2	ReLU1_1	5 × 222 × 222 × 64					
3	Conv1_2	5 × 220 × 220 × 64	64	3 × 3	1 × 1	0 × 0	36,928
4	ReLU1_2	5 × 220 × 220 × 64					
5	Pool_1	5 × 110 × 110 × 64	1	2 × 2	2 × 2	0 × 0	
6	Conv2_1	5 × 108 × 108 × 128	128	3 × 3	1 × 1	0 × 0	73,856
7	ReLU2_1	5 × 108 × 108 × 128					
8	Conv2_2	5 × 106 × 106 × 128	128	3 × 3	1 × 1	0 × 0	147,584
9	ReLU2_2	5 × 106 × 106 × 128					
10	Pool_2	5 × 53 × 53 × 128	1	2 × 2	2 × 2	0 × 0	
11	Conv3_1	5 × 51 × 51 × 256	256	3 × 3	1 × 1	0 × 0	295,168
12	ReLU3_1	5 × 51 × 51 × 256					
13	Conv3_2	5 × 49 × 49 × 256	256	3 × 3	1 × 1	0 × 0	590,080
14	ReLU3_2	5 × 49 × 49 × 256					
15	Pool_3	5 × 24 × 24 × 256	1	2 × 2	2 × 2	0 × 0	
16	Fc4	5 × 1000 × 1					147,457,000
17	ReLU4	5 × 1000 × 1					
18	Dropout4	5 × 1000 × 1					
19	LSTM	1000 × 1					8,004,000
20	Fc5	50 × 1					50,050
21	Softmax layer	50 × 1					
22	Output layer	Number of classes × 1					
		Total number of trainable parameters: 156,656,458					

- There are no methods for changing low-quality thermal images obtained under various environments into high-definition (HD) thermal images. Accordingly, this study proposes a method for generating HD thermal images from low-quality thermal images of long-distance objects through CycleGAN. To develop the proposed method, hyper-parameters, the number of filters, the numbers of convolution layers, and the sizes of the filters in the existing CycleGAN were modified according to the experiments.
- There are no methods for analyzing and removing halo effects of objects in various long-distance environments. Accordingly, this study proposes a halo effect removal method using the modified CycleGAN considering its high ability of image transformation.
- Some methods for matching a short-distance object with a skeleton or extracting skeleton information in a thermal image have been developed. However, no method has been proposed for generating a skeleton image from a thermal image. In addition, only a few studies on applying a thermal image-based skeleton have been conducted. Accordingly, this study proposes a method for generating a skeleton image directly from an original thermal image through a deep learning algorithm.
- The developed CNN model, the data generated, and the Dongguk activities and actions database (DA&A-DB2)

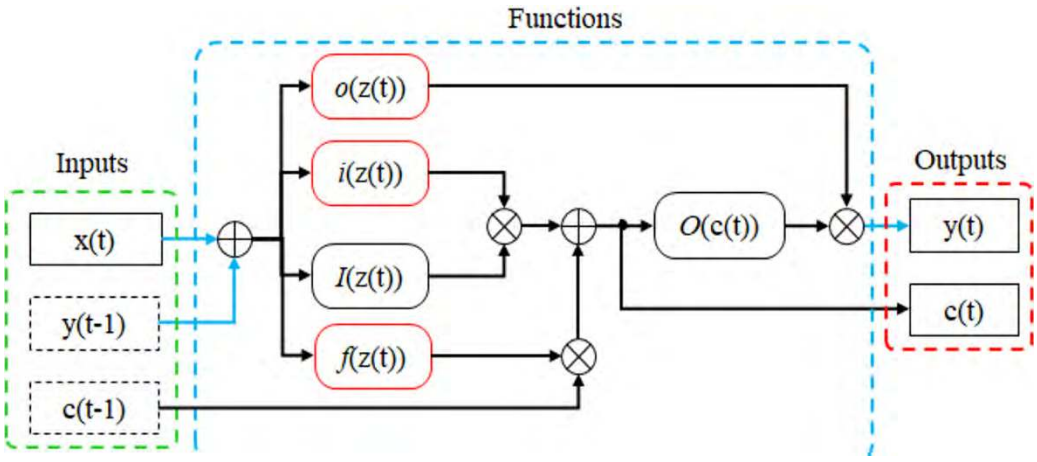

FIGURE 12. Example of conventional LSTM architecture.

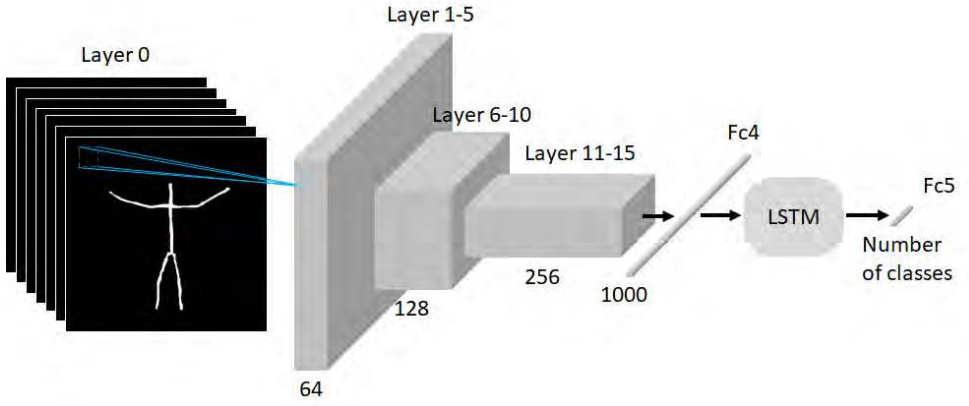

FIGURE 13. Our CNN-LSTM architecture.

were released [43] for a fair performance evaluation by other researchers.

IV. PROPOSED METHOD

A. CAMERA SETTINGS AND HUMAN DETECTION

This section provides a simple description regarding the camera setup and human detection method. Figure 1 shows the camera setup and experimental environment. A thermal camera can acquire images (depth of 14 bits and size of 640 pixels × 480 pixels) at a rate of 30 fps [44]. During the experiment, a thermal camera was placed at various heights (5–10 m) under various environments to acquire the images. Settings similar to those of a conventional CCTV camera were applied.

Because it is difficult to detect an object under visible light in images obtained under various environments (a dark environment, variations in illumination, and severe shadows), this study utilized a thermal camera. The details of the object detection method are shown in [45]. As indicated in the thermal image on the right side of Figure 1, a larger ROI (red dashed box) than the detected area (green) being captured was applied.

B. OVERALL PROCEDURE OF PROPOSED METHOD

Figure 2 shows an overall flowchart of the proposed method. As indicated in Figure 1, the original cropped image was used as the input. During the halo effect removal phase, the halo effect is removed from the input image using a CycleGAN network. In the skeleton generation phase, a skeleton is generated from a thermal image using a CNN. In the action recognition phase, human action recognition was conducted using a CNN-LSTM and sequential skeleton images. The details of each phase are described step by step in the following section.

C. IMAGE RESTORATION

1) OVERALL PROCEDURE OF IMAGE RESTORATION

This section describes the method for restoring the thermal camera images. When thermal images are obtained in various environments, if the background and object have similar temperatures, the images acquired are similar to those shown in Figure 3(a). By contrast, if the background and object have different temperatures, the images acquired are similar to those shown in Figure 3(b). When the object was detected in the images of Figure 3(a), a portion of the human body

FIGURE 14. Examples from each of the 16 sub-datasets: (a)–(p) sub-datasets I–XVI.

TABLE 6. Description of experimental database.

Sub-dataset	Condition	Detail Description
I (shown in Figure 14(a))	Humidity of 62.6%, wind speed of 1.3 m/s, 21.9 °C, afternoon, cloudy, autumn	- Object is clear at the current position but is difficult to see in the left-upper position where the area is brighter in the thermal image
II (shown in Figure 14(b))	6.0 °C, afternoon, cloudy, humidity of 39.6%, wind speed of 1.9 m/s	- Object is clear at the current position but is difficult to see in the upper position where the area is brighter in the thermal image
III (shown in Figure 14(c))	14.0 °C, afternoon, sunny, humidity of 43.4%, wind speed of 3.1 m/s	- The temperature outside the building is increased by the air heating system of the building in the thermal image
IV (shown in Figure 14(d))	1.2 °C, morning, humidity of 73.0%, wind speed of 1.6 m/s	- The temperature of the window of the building is changed over time, making it difficult to visualize the objects in the thermal image - The reflection of the object in the window makes it difficult to visualize the object
V (shown in Figure 14(e))	1.0 °C, afternoon, humidity of 50.6%, wind speed of 1.7 m/s	- The intensity of trees and leaves increases when it is sunny in a cold environment
VI (shown in Figure 14(f))	31.3 °C, noon, humidity of 43.4%, wind speed of 3.1 m/s	- The temperatures of the human body and background are similar, making it difficult to segment the human area from the background
VII (shown in Figure 14(g))	10.2 °C, afternoon, cloudy and rainy, humidity of 60.6%, wind speed of 1.7 m/s	- A halo effect is shown below the human area in the thermal image - The top part of the object is not visible in the thermal image owing to a temperature similarity
VIII (shown in Figure 14(h))	38.5 °C, noon, humidity of 35.3%, wind speed of 1.2 m/s	- The temperature of the background is higher than that of the human body
IX (shown in Figure 14(i))	18.9 °C, night, humidity of 62.6%, wind speed of 1.3 m/s	- A halo effect is shown below the human area in the thermal image - The temperatures of the human body and background are similar - The object is not seen in the visible light image
X (shown in Figure 14(j))	10.9 °C, dark night, humidity of 48.3%, wind speed of 2.0 m/s	- The dataset was collected at night and the halo effect is shown below the human area - The object is not seen in the visible light image
XI (shown in Figure 14(k))	10.9 °C, dark night, humidity of 48.3%, wind speed of 2.0 m/s	- The object is not shown in the visible light image - The reflection of the object in the window makes it difficult to visualize the object
XII (shown in Figure 14(l))	20.2 °C, dark night, humidity of 58.6%, wind speed of 1.2 m/s	- The temperature outside the building is increased by the air heating system of the building in the thermal image - The object is not seen in the visible light image
XIII (shown in Figure 14(m))	-2.0 °C, dark night, humidity of 50.6%, wind speed of speed of 1.8 m/s	- The intensity of the trees and leaves is high owing to sunlight during the daytime
XIV (shown in Figure 14(n))	12.0 °C, dark night, humidity of 63.1%, wind speed of 1.5 m/s	- A halo effect is shown below the human area in the thermal image - The object is not seen in the visible light image
XV (shown in Figure 14(o))	28.0 °C, night, humidity of 45.1%, wind speed of 1.6 m/s	- The temperature of the background is higher than that of the human body owing to sunlight during the daytime - The object is not seen in the visible light image
XVI (shown in Figure 14(p))	10.0 °C, dark night, humidity of 63.1%, wind speed of 1.5 m/s	- A halo effect is shown below the human area in the thermal image - The object is not seen in the visible light image

area disappeared or was cut out, as shown in Figure 3(c), which decreased the detection accuracy for the object. However, when the object was detected in the images shown in Figure 3(b), the results were good, as indicated in Figure 3(d). Accordingly, this study utilized the CycleGAN network [46] and conducted image restoration, as illustrated in Figure 4, to convert the thermal images of Figure 3(a) into those of 3(b). As shown in Figure 4, the cropped thermal image mentioned in Section IV.A was used as the input.

In Figure 4, Conv, BN, Relu, and Add denote the convolutional layer, batch normalization layer, rectified linear unit, and addition function, respectively. As illustrated in Figure 5(b), unpaired training data were used to train CycleGAN.

2) DESCRIPTION OF CycleGAN AND DISCRIMINATOR CNN STRUCTURES

This section describes the CycleGAN network structures used for the image restoration methods in detail. CycleGAN shows the high-transformation results obtained from a model trained using unpaired training data. Because the proposed image restoration method used unpaired training data, it applied CycleGAN. In addition, before being used, the original CycleGAN model was fitted to our database by modifying the

hyper-parameters, number of filters, number of convolution layers, and size of the filters, as indicated in Tables 2–4. Tables 2–4 describe the detailed structures of the generator, residual block, and discriminator used by the CycleGAN, respectively.

D. REMOVAL OF HALO EFFECT

This section describes the method of the halo effect removal.

As shown in Figure 6(a), a halo effect appears like a shadow under the area of the human body. This study examined methods for removing halo effects, as indicated in Figure 6(b), from the images shown in Figure 6(a). When an object is detected in the images of Figure 6(a), the body area is connected to the area of the halo effect, as shown in the images of Figure 6(c). Thus, the accuracy of the detection method for the object is degraded. However, if the object is detected in the images of Figure 6(b), the detection results are satisfactory, as shown in Figure 6(d). For the halo effect removal method, the existing CycleGAN structure was fitted to our database by modifying the hyper-parameters, number of filters, number of convolution layers, and size of the filters, as presented in Tables 2–4. The cropped thermal image mentioned in Section 4.1 was used as the input of the CycleGAN network, as illustrated in Figure 7. Here, Conv, BN, Relu, and Add denote the convolutional layer, batch normalization layer, rectified linear unit, and addition function, respectively, as shown in Figure 7. As illustrated in Figure 5(b), unpaired training data were used to train the CycleGAN.

E. SKELETON GENERATION

This section describes the method of skeleton generation. It is difficult to detect an object or extract skeleton information from an image obtained using a visible light camera in a dark environment, where a person is barely visible. Some methods for detecting a human body in a dark environment using a thermal camera have been developed. However, no method has been proposed to extract the skeleton information of a detected object.

Accordingly, this study proposes a method for extracting a skeleton image from a thermal image obtained in a dark environment. The existing methods for extracting skeleton information from images obtained in a bright environment were implemented, as shown in Figure 8(a)–(d). Because the image in Figure 8(a) has much more spatial information of the joints than the image in Figure 8(c), a skeleton was made by detecting the locations of the joints, as shown in Figure 8(b) [40], [41]. In the case of Figure 8(c), where little spatial information of the joint is given, a thinning method was applied to make a skeleton, as shown in Figure 8(d) [47]–[50].

However, Figures 8(e) and 8(f) may have spatial information, as shown in Figure 8(a), or no such information, as shown in Figure 8(c), depending on the environment where the thermal image is acquired. Moreover, if the method of Figure 8(d) is applied to the image of Figure 8(e), or the method of Figure 8(b) is applied to that of Figure 8(f), the desired skeleton has difficulty being generated. Thus, this study proposes a method for generating a skeleton from thermal images, as shown in Figures 8(e) and 8(f).

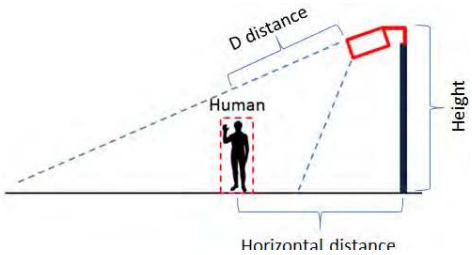

FIGURE 15. Example of camera setup.

TABLE 7. Average distance of camera setup used to collect the 16 sub-datasets (unit: meters).

Datasets	Height	Horizontal distance	D distance
Sub-datasets I, II, VII, X, IX, XIV, XVI	10	15	18
Sub-datasets III, VI, XII	5	15	15.8
Sub-datasets IV, XI	8	10	12.8
Sub-datasets V, XIII	7.7	11	13.4
Sub-datasets VIII, XV	6	11	12.5

TABLE 8. Numbers of frames and the types of human actions in our database.

	#Frame	
Behavior	Day	Night
Walking	11,984	11,439
Running	3,174	4,336
Standing	5,556	916
Sitting	4,312	3,448
Approaching	4,526	5,490
Leaving	1,016	1,116
Waving with two hands	29,584	14,090
Waving with one hand	27,532	18,058
Punching	32,048	17,452
Lying down	7,762	5,488
Kicking	33,693	23,241
Total	266,261	

We generate skeleton image by using the open source of CNN proposed in [51]. The network in [51] was originally proposed for style transfer and super-resolution reconstruction based on perceptual loss, and we adopted this network for skeleton generation. The detailed explanations for this CNN can be referred to [51]. The structure of this CNN was the same as that of generator of CycleGAN. The structure was fitted to our database by modifying the hyper-parameters, number of filters, number of convolution layers, and size of the filters, as shown in Table 2.

As illustrated in Figure 9, the original thermal image was used as the input image of the CNN, and the skeleton image was used as the output image. A skeleton in an image for training is created, and was set to be thicker than the conventional skeleton. In an additional experiment, the output image extracted by a CNN was postprocessed (size filtering and morphological operations) to generate a narrow skeleton in an image, as illustrated in Figure 9.

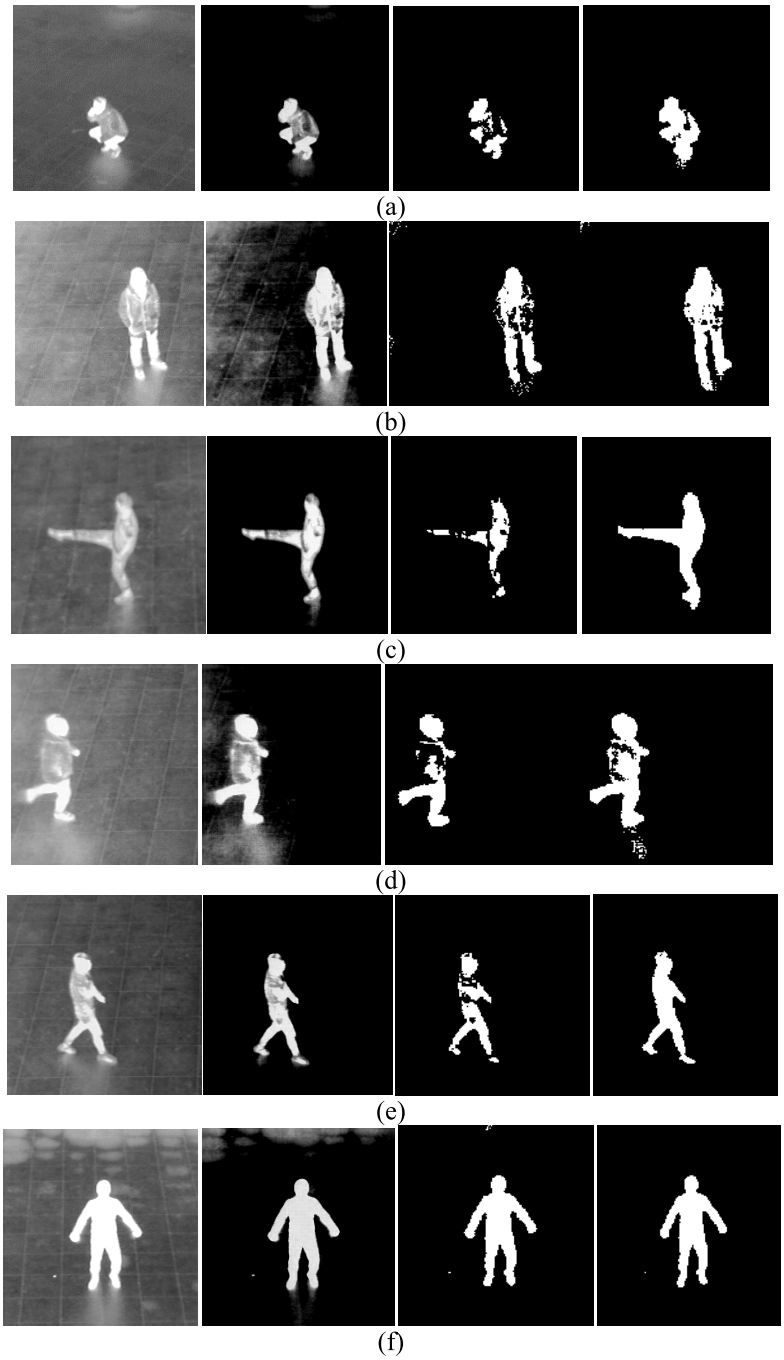

FIGURE 16. Examples of image restoration. (a)–(f) examples 1–6, respectively, where the first, second, third, and fourth images are the original image, restored image, binarized version of the original image, and binarized version of the restored image, respectively.

F. ACTION RECOGNITION

1) OVERALL PROCEDURE OF ACTION RECOGNITION

We propose an action recognition method that extracts the action information of a long-distance object from thermal images obtained in a dark or bright environment by using a CNN stacked LSTM (CNN-LSTM). As illustrated in Figure 10, the human action recognition was conducted by adopting a sequence of skeleton images as the input. As shown in Figure 11, this study attempted to recognize 11 actions such as waving with one hand, waving with two hands, punching, kicking, sitting, standing, walking, running, lying down, leaving, and approaching.

2) DESCRIPTION OF CNN-LSTM STRUCTURE

To solve the problem of long-term dependencies, an LSTM was applied to various research areas such as action

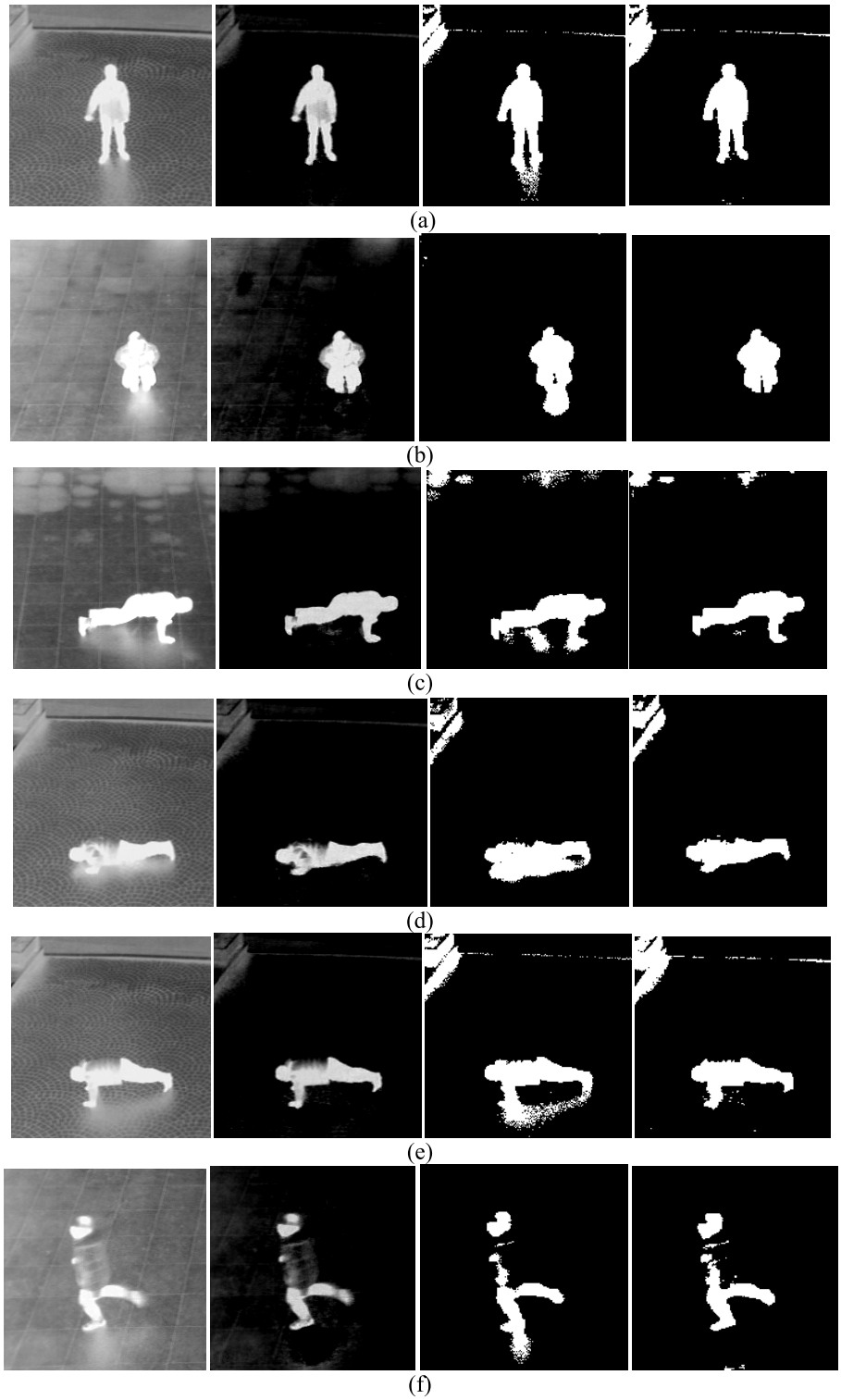

FIGURE 17. Examples of halo effect removal. (a)–(f) show examples 1–6, respectively, where the first, second, third, and fourth images are the original image, halo effect removed image, binarized version of the original image, and binarized version of the halo effect removed image, respectively.

recognition [32], text recognition [52], gait recognition [53], caption generator [54], speech recognition [55], person re-identification [56], and gait diagnosis [57]. Regarding the long-term memory and temporal information, LSTM-based methods have turned out to be most effective in solving the vanishing and exploding gradient problem. Accordingly, this study utilized an LSTM to extract temporal information from sequential images. In addition, this study also connected a

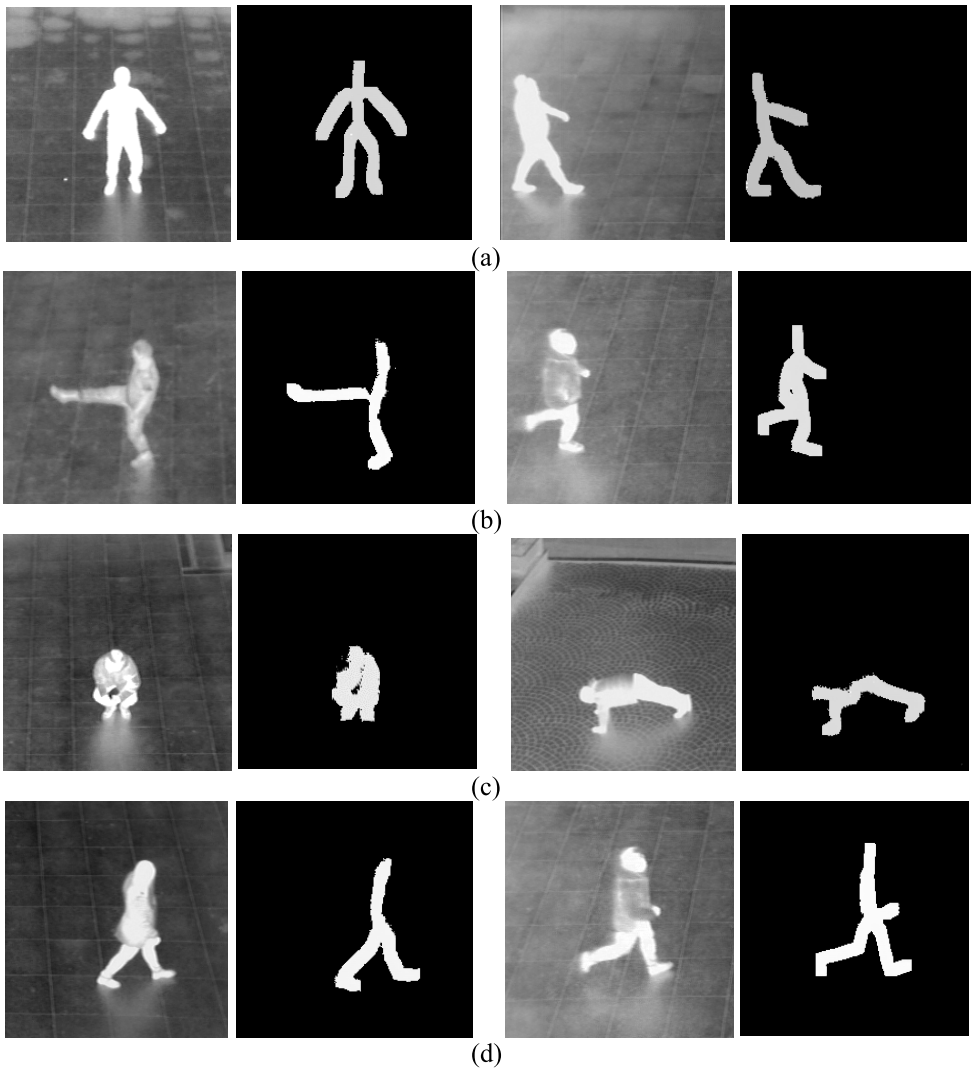

FIGURE 18. Examples of skeleton generation. Left and right image pairs in (a)–(d) show examples 1–8, respectively, where the first and third images are the original images whereas the second and fourth images are the generated skeleton images.

CNN to an LSTM to extract the spatial features. To enhance the accuracy of human action recognition, various CNN structures were redesigned and tested. Table 5 shows the most appropriate structure as demonstrated through the test results. The optimal frame numbers (5 frames) for CNN-LSTM of Table 5 were experimentally determined with training data, which showed the highest accuracy of human action recognition.

In Figure 12, x(t), y(t-1), and c(t-1) denote the current input, previous output, and previous cell values; y(t) and c(t) indicate the current output and cell values; $i(.)$, $f(.)$, and $o(.)$ indicate the input, forget, and output gait functions (sigmoid function); and $I(.)$ and $O(.)$ are the input and output activation functions (tanh function), respectively. Finally, blue arrows, red box and dashed black boxes indicate weighted connections, gait functions and previous information, respectively.

V. EXPERIMENTAL RESULTS

A. DESCRIPTION OF EXPERIMENT SETUP AND DATABASE

Open databases for action recognition acquired from visible light camera environments have already been developed [58]–[60]. However, no databases have been acquired from a thermal camera environment.

This study conducted an experiment using the DA&A-DB2 database [43], which contains thermal images of long-distance objects obtained in various environments (time zones, weather, seasons, and camera settings) and locations. Although the database consists of both visible light images and thermal images, this study used only thermal images. The database consists of 16 sub-datasets including a total of 266,261 images. Figure 14 and Table 6 provide the detailed information of the database. A desktop computer was used for training and testing. The specifications of the desktop computer include an Nvidia graphics card (Nvidia GeForce

TABLE 9. Various methods for action recognition.

Methods	Explanations
Method 1	Original image → Action recognition
Method 2	Original image → Image restoration → Action recognition
Method 3	Original image → Skeleton generation → Action recognition
Method 4	Original image → Halo effect removal → Action recognition
Method 5	Original image → Image restoration → Halo effect removal → Action recognition
Method 6	Original image → Image restoration & Halo effect removal → Action recognition
Method 7	Original image → Image restoration → Halo effect removal → Skeleton generation → Action recognition
Method 8	Original image → Image restoration → Halo effect removal → Skeleton generation → Thinning → Action recognition
Method 9	Original image → Skeleton generation → Thinning → Action recognition
Method 10	Original image → Image restoration → Skeleton generation → Action recognition
Method 11	Original image → Halo effect removal → Skeleton generation → Action recognition

GTX TITAN X [61]), Intel CPU (core i7-6700 CPU @ 3.40 GHz (with eight CPUs)), and 32 GB of RAM. The proposed method was implemented using a Python-based Keras application programming interface (API) with a Tensorflow backend engine [62] and the OpenCV library [63].

Figure 15 and Table 7 show the height of the camera, the distance between an object and the location of the camera (horizontal distance), and another distance between an object and a camera (D distance). Table 8 shows the types of human actions and the number of images for each action.

B. TRAINING OF CycleGAN AND CNN-LSTM MODELS

This section describes the training phase of the proposed methods in detail. All methods used images with size of 224 × 224 pixels for training and testing. When the generator of CycleGAN was trained using the image restoration method, the cycle-consistency loss, identity loss, training epoch, learning rate, mini-batch, loss function, and optimizer were set to 10.0, 1.0, 6,000, 0.00001, 1, mean squared error [64], and adaptive moment estimation methods (Adam) [65], respectively. When the generator was trained using the halo effect removal method, the learning rate and training epoch were set to 0.00001 and 5,000 respectively. In the case of the skeleton image generation method, the learning rate and training epoch were set to 0.00001 and 2,000, respectively. In the case of the human action recognition method, the training epoch, learning rate, momentum, mini-batch, optimizer, and loss function were set to 5, 0.0001, 0.9, 10, Adam, and categorical cross entropy, respectively. The epoch number was determined based on outcome images obtained during training. The learning rate was determined based on the epoch number.

C. TESTING

1) TESTING OF IMAGE RESTORATION

This section describes the testing results of the image restoration method. Figure 16 shows the image restoration results. The corresponding binarized images are also shown to see

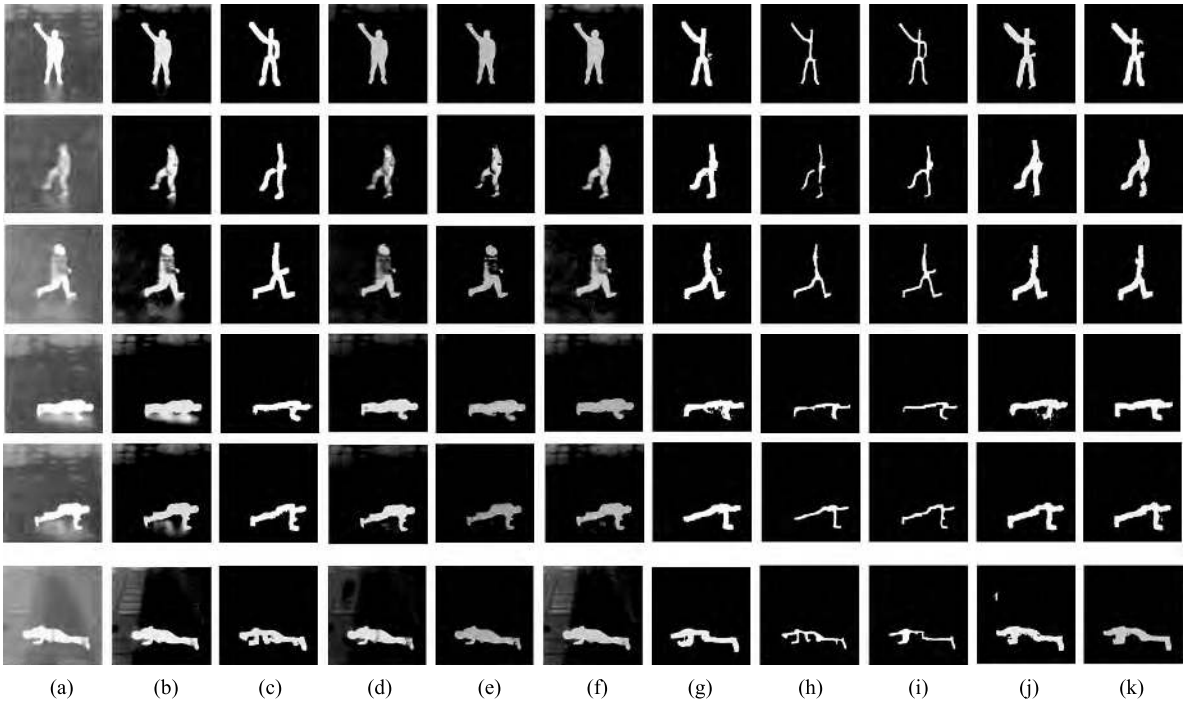

FIGURE 19. Methods for action recognition: (a)–(k) methods 1–11.

TABLE 10. Confusion matrix of the results of human action recognition using method 1 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	0	0	0	0	0	0
Waving with one hand	0	99.7	0	0	0	0	0	0	0
Punching	0	0.3	100	3	0	0	0	0	0
Kicking	0	0	0	97	5.2	0	0	0	0
Lying	0	0	0	0	94.8	0	0	0	0
Walking	0	0	0	0	0	93.5	7.5	0	0
Running	0	0	0	0	0	6.3	90.7	0	0
Standing	0	0	0	0	0	0	0	100	0
Sitting	0	0	0	0	0	0.2	1.8	0	100

TABLE 11. Confusion matrix of the results of human action recognition using method 2 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	0	0	0	0	0	0
Waving with one hand	0	100	0	0.7	0	0	0	0	0
Punching	0	0	100	0	0	0	0	0	0
Kicking	0	0	0	99.3	0	0	0	0	0
Lying	0	0	0	0	96.7	0	0	0	0
Walking	0	0	0	0	0	92.5	7.2	0	0
Running	0	0	0	0	0.3	7.5	88.5	0	0
Standing	0	0	0	0	3	0	4.3	100	0
Sitting	0	0	0	0	0	0	0	0	100

TABLE 12. Confusion matrix of the results of human action recognition using method 3 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	0	0	0	0	0	0
Waving with one hand	0	100	0.8	0	0	0	0	0	0
Punching	0	0	99.2	0	0	0	0	0	0
Kicking	0	0	0	100	0	0	0	0	0
Lying	0	0	0	0	94.8	0	0	0	0
Walking	0	0	0	0	0	90.4	8.6	0	0
Running	0	0	0	0	0	7.6	89.2	0	0
Standing	0	0	0	0	0	2	2.2	100	0
Sitting	0	0	0	0	5.2	0	0	0	100

how the results differ. As illustrated in Figure 16, the binarized version of the restored image expresses the human body area more accurately than the binarized version of the original image.

2) TESTING OF HALO EFFECT REMOVAL

This section presents the testing results of the halo effect removal method. Figure 17 shows the results of the halo effect removal method. The corresponding binarized images

TABLE 13. Confusion matrix of the results of human action recognition using method 4 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	4.4	0	0	0	0	0
Waving with one hand	0	100	0	0	0	0	0	0	0
Punching	0	0	100	0	0	0	0	0	0
Kicking	0	0	0	95.6	2.2	0	0	0	0
Lying	0	0	0	0	94.3	0	0	0	0
Walking	0	0	0	0	0	88.5	11.1	0	0
Running	0	0	0	0	0	11.1	88.9	0	0
Standing	0	0	0	0	0	0.4	0	100	0
Sitting	0	0	0	0	3.5	0	0	0	100

TABLE 14. Confusion matrix of the results of human action recognition using method 5 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	3.4	0	0	0	0	0
Waving with one hand	0	100	0	1.3	0	0	0	0	0
Punching	0	0	100	0	0	0	0	0	0
Kicking	0	0	0	95.3	5.2	0	0	0	0
Lying	0	0	0	0	94.8	0	0	0	0
Walking	0	0	0	0	0	93.7	7.2	0	0
Running	0	0	0	0	0	6.3	92.8	0	0
Standing	0	0	0	0	0	0	0	100	0
Sitting	0	0	0	0	0	0	0	0	100

TABLE 15. Confusion matrix of the results of human action recognition using method 6 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	1.3	0	0	0	0	0
Waving with one hand	0	100	0	0.7	0	0	0	0	0
Punching	0	0	100	4.4	0	0	0	0	0
Kicking	0	0	0	93.6	1	0	0	0	0
Lying	0	0	0	0	93.8	0	0	0	0
Walking	0	0	0	0	0	96.6	4	0	0
Running	0	0	0	0	0	3.4	92.5	0	0
Standing	0	0	0	0	0	0	2.1	100	0
Sitting	0	0	0	0	5.2	0	1.4	0	100

are also shown to see how the results differ. As illustrated in Figure 17, the binarized images with the halo effects removed express the human body area more accurately than those with the halo effects.

D. TESTING OF SKELETON GENERATION

This section describes the testing results of the skeleton generation method, the images of which are shown in Figure 18.

TABLE 16. Confusion matrix of the results of human action recognition using method 7 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	0	0	0	0	0	0
Waving with one hand	0	100	0.3	3.7	0	0	0	0	0
Punching	0	0	99.7	0	0	0	0	0	0
Kicking	0	0	0	96.3	0	0	0	0	0
Lying	0	0	0	0	100	0	0	0	0
Walking	0	0	0	0	0	91.4	8.2	0	0
Running	0	0	0	0	0	8.6	83.9	0	0
Standing	0	0	0	0	0	0	2.5	100	0
Sitting	0	0	0	0	0	0	5.4	0	100

TABLE 17. Confusion matrix of the results of human action recognition using method 8 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	0	0	0	0	0	0
Waving with one hand	0	100	0	3.4	0	0	0	0	0
Punching	0	0	100	0	0	0	0	0	0
Kicking	0	0	0	96.6	4.6	0	0	0	0
Lying	0	0	0	0	95.4	0	0	0	0
Walking	0	0	0	0	0	85	7.5	0	0
Running	0	0	0	0	0	11.5	84.6	0	0
Standing	0	0	0	0	0	3.5	2.9	100	0
Sitting	0	0	0	0	0	0	5	0	100

TABLE 18. Confusion matrix of the results of human action recognition using method 9 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	1	0	0	0	0	0
Waving with one hand	0	100	0	1	0	0	0	0	0
Punching	0	0	100	0	0	0	0	0	0
Kicking	0	0	0	98	0	0	0	0	0
Lying	0	0	0	0	95.2	0	0	0	0
Walking	0	0	0	0	0	87.4	7.2	0	0
Running	0	0	0	0	0	6.9	92.8	0	0
Standing	0	0	0	0	0	5.7	0	98	0
Sitting	0	0	0	0	4.8	0	0	2	100

1) TESTING OF HUMAN ACTION RECOGNITION

This section presents the testing results of human action recognition. The 11 methods of Table 9 were applied to the test. Figure 19 shows the input images for each method. Tables 10–25 provides the accuracies of the 11 methods for the action recognition. Table 26 lists the details of the processing time. The processing time was measured in the experiment environment described in Section V.A.

As shown in Table 25, method 10 using image restoration and skeleton generation, and method 11 (the proposed method) using halo effect removal and skeleton generation, produced the highest accuracies. In other words, the image

TABLE 19. Confusion matrix of the results of human action recognition using method 10 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	0	0	0	0	0	0
Waving with one hand	0	100	2.8	1.7	0	0	0	0	0
Punching	0	0	97.2	0	0	0	0	0	0
Kicking	0	0	0	98.3	0	0	0	0	0
Lying	0	0	0	0	100	0	0	0	0
Walking	0	0	0	0	0	94	8.3	0	0
Running	0	0	0	0	0	6	89.6	0	0
Standing	0	0	0	0	0	0	0.7	100	0
Sitting	0	0	0	0	0	0	1.4	0	100

TABLE 20. Confusion matrix of the results of human action recognition using method 11 (unit: %).

Actual / Recognized	Waving with two hands	Waving with one hand	Punching	Kicking	Lying	Walking	Running	Standing	Sitting
Waving with two hands	100	0	0	1	0	0	0	0	0
Waving with one hand	0	100	3.4	0	0	0	0	0	0
Punching	0	0	96.6	0	0	0	0	0	0
Kicking	0	0	0	99	0	0	0	0	0
Lying	0	0	0	0	99.3	0	0	0	0
Walking	0	0	0	0	0	93.7	6.5	0	0
Running	0	0	0	0	0	6.3	92.8	0	0
Standing	0	0	0	0	0	0	0	100	0
Sitting	0	0	0	0	0.7	0	0.7	0	100

TABLE 21. Accuracies of human action recognition methods (unit: %).

	Method 1			Method 2			Method 3		
Human actions	**TPR**	**PPV**	**ACC**	**TPR**	**PPV**	**ACC**	**TPR**	**PPV**	**ACC**
Waving with two hands	100	100	100	100	100	100	100	100	100
Waving with one hand	99.7	100	99.98	100	99.44	99.95	100	98.33	99.85
Punching	100	98.68	99.76	100	100	100	99.2	100	99.85
Kicking	97	91.11	99.1	99.3	100	99.95	100	100	100
Lying	94.8	100	99.32	96.7	100	99.56	94.8	100	99.32
Walking	93.5	95.88	98.66	92.5	96.03	98.56	90.4	95.17	98.19
Running	90.7	88.46	98.56	88.5	85.76	98.22	89.2	86.16	98.29
Standing	100	100	100	100	95.7	99.32	100	97.5	99.61
Sitting	100	98.58	99.85	100	100	100	100	93.72	99.32
Leaving	100	100	100	100	100	100	100	100	100
Approaching	100	100	100	100	100	100	100	100	100
Average	**97.79**	**97.52**	**99.57**	**97.91**	**97.9**	**99.6**	**97.61**	**97.35**	**99.49**

restoration, halo effect removal, and skeleton generation were effective at improving the accuracy of human action recognition.

For examples, as shown in Figures 3(the 1st and 3rd images of (a)), 6(the 1st and 3rd images of (a)), 11(i), 14(j), and 17 (the 1st images of (b) and (d)), there are many cases that the

TABLE 22. Accuracies of human action recognition methods (unit: %).

	Method 4			Method 5			Method 6		
Human actions	**TPR**	**PPV**	**ACC**	**TPR**	**PPV**	**ACC**	**TPR**	**PPV**	**ACC**
Waving with two hands	100	96	99.68	100	96.89	99.76	100	98.73	99.9
Waving with one hand	100	100	100	100	98.88	99.9	100	99.44	99.95
Punching	100	100	100	100	100	100	100	98.28	99.68
Kicking	95.6	95.93	99.39	95.3	90.97	98.97	93.6	98.23	99.41
Lying	94.3	100	99.24	94.8	100	99.32	93.8	100	99.19
Walking	88.5	93.72	97.78	93.7	96.08	98.7	96.6	97.87	99.29
Running	88.9	81.05	97.82	92.8	88.7	98.7	92.5	93.48	99.05
Standing	100	99.68	99.95	100	100	100	100	99.05	99.85
Sitting	100	95.65	99.54	100	100	100	100	92.89	99.22
Leaving	100	100	100	100	100	100	100	100	100
Approaching	100	100	100	100	100	100	100	100	100
Average	**97.03**	**96.55**	**99.4**	**97.87**	**97.41**	**99.58**	**97.86**	**98**	**99.6**

TABLE 23. Accuracies of human action recognition methods (unit: %).

	Method 7			Method 8			Method 9		
Human actions	**TPR**	**PPV**	**ACC**	**TPR**	**PPV**	**ACC**	**TPR**	**PPV**	**ACC**
Waving with two hands	100	100	100	100	100	100	100	99.05	99.93
Waving with one hand	100	96.46	99.68	100	97.25	99.76	100	99.16	99.93
Punching	99.7	100	99.95	100	100	100	100	100	100
Kicking	96.3	100	99.73	96.6	91.96	99.14	98	100	99.85
Lying	100	100	100	95.4	100	99.39	95.2	100	99.36
Walking	91.4	95.41	98.34	85	95.49	97.58	87.4	95.81	97.90
Running	83.9	83.87	97.8	84.6	79.73	97.48	92.8	87.8	98.63
Standing	100	98.89	99.83	100	95.99	99.36	98	95.31	98.95
Sitting	100	96.54	99.63	100	96.76	99.66	100	91.47	99.05
Leaving	100	100	100	100	100	100	100	100	100
Approaching	100	100	100	100	100	100	100	100	100
Average	**97.39**	**97.38**	**99.54**	**96.51**	**96.11**	**99.31**	**97.4**	**97.15**	**99.42**

temperature of human body is similar to that of background or severe halo effects happen, which produces incorrect segmentation of body area and consequent error of human action recognition. Therefore, we use CycleGAN in order to make the body area more distinctive from background and remove halo effects.

As shown in Table 9, the methods 1 and 2 respectively show the cases before and after CycleGAN. In addition, the methods 3 and 10 (or 11) show the cases before and after CycleGAN, respectively. As shown in Table 25, the accuracy by the method 2 is higher than that by the method 1. In addition, the accuracies by the methods 10 and 11 are higher than that by the method 3. From these results, we confirm that CycleGAN can improve the overall accuracy of human action recognition.

We use the skeleton image for CNN-LSTM instead of the full frame because the motion information based on the skeleton image can be more distinctive than that of the full frame. As shown in Table 9, the methods 2 and 10 show the cases of action recognition without and with skeleton generation, respectively. In addition, the methods 4 and 11 show the cases of action recognition without and with skeleton generation, respectively. As shown in Table 25, the accuracies of action recognition by the methods 10 and 11

TABLE 24. Overall accuracies of human action recognition methods (unit: %).

	Method 10			Method 11		
Human actions	**TPR**	**PPV**	**ACC**	**TPR**	**PPV**	**ACC**
Waving with two hands	100	100	100	100	99.05	99.93
Waving with one hand	100	93.16	99.36	100	93.4	99.39
Punching	97.2	100	99.49	96.6	100	99.39
Kicking	98.3	100	99.88	99	100	99.93
Lying	100	100	100	99.3	100	99.9
Walking	94	95.53	98.68	93.7	96.46	98.75
Running	89.6	88.97	98.53	92.8	88.7	98.7
Standing	100	99.68	99.95	100	100	100
Sitting	100	99.05	99.9	100	98.58	99.85
Leaving	100	100	100	100	100	100
Approaching	100	100	100	100	100	100
Average	**98.1**	**97.85**	**99.62**	**98.31**	**97.84**	**99.62**

TABLE 25. Overall accuracies of human action recognition methods (unit: %).

	Method 1	**Method 2**	**Method 3**	**Method 4**	**Method 5**	**Method 6**	**Method 7**	**Method 8**	**Method 9**	**Method 10**	**Method 11**
TPR	97.79	97.91	97.61	97.03	97.87	97.86	97.39	96.51	97.4	98.1	98.31
PPV	97.52	97.9	97.35	96.55	97.41	98	97.38	96.11	97.15	97.85	97.84
ACC	99.57	99.6	99.49	99.4	99.58	99.6	99.54	99.31	99.42	99.62	99.62

TABLE 26. Overall processing time of human action recognition methods (unit: ms).

	Image restoration by using CycleGAN	Skeleton generation by using CNN	Halo effect removal by using CycleGAN	Image restoration & Halo effect removal	Thining	Action recognition by using CNN-LSTM	Total	Frames per second
Method 1	n/a	n/a	n/a	n/a	n/a	62.1	62.1	16.1
Method 2	7.66	n/a	n/a	n/a	n/a	60.342	68	14.7
Method 3	7.854	n/a	n/a	n/a	n/a	60.536	68.4	14.62
Method 4	n/a	n/a	7.92	n/a	n/a	69.123	77	12.99
Method 5	7.66	n/a	7.902	n/a	n/a	67.056	82.6	12.1
Method 6	n/a	n/a	n/a	7.94	n/a	67.953	75.89	13.17
Method 7	7.66	7.671	7.92	n/a	n/a	68.07	91.3	10.95
Method 8	7.66	7.671	7.92	n/a	0.313	67.498	91	10.99
Method 9	n/a	7.854	n/a	n/a	0.286	65.689	73.8	13.55
Method 10	7.66	8.049	n/a	n/a	n/a	78.47	94.2	10.62
Method 11	n/a	7.731	7.92	n/a	n/a	80.754	96.4	10.37

n/a means not available

are higher than those by the methods 2 and 4, respectively, which confirms that the skeleton image is more effective for human action recognition than the full frame data.

According to Table 26, the processing time of both methods 10 and 11 was approximately 10 fps. The following section describes a comparative experiment with the existing methods.

TABLE 27. Comparison of previous methods with the proposed method (unit: %).

Method	TPR	PPV	ACC
Fourier descriptor-based [1]	46.6	51	50.4
GEI-based [10]	55.5	46.3	65.1
Convexity defect-based [16]	47.7	77.4	71.8
Projection-based distance [66]	94.8	99.1	97.6
Fuzzy system-based [67]	99.9	96.3	98.8
Proposed method	**98.3**	**97.8**	**99.6**

E. COMPARISONS

This section compares the proposed method with the existing methods. Table 27 shows the accuracies of five existing methods of human action recognition. As is clear from Table 27, the proposed method shows higher recognition rates than the existing methods.

VI. CONCLUSIONS

This study proposed human action recognition methods using thermal image restoration, halo effect removal, and skeleton generation. These approaches were combined in various ways to produce different results. Various techniques including CycleGAN, CNN, and CNN-LSTM were adopted for the proposed methods. In addition, an experiment was conducted using the DA&A-DB2 open database, which was built solely for the present study. There are many databases acquired by thermal camera [68]–[75]. However, most of them are for pedestrian or object detection, and there is no existing database including the images of human action with halo effects. Therefore, we collected our DA&A-DB2 database for experiments. For fair performance evaluation, this database is released to other researchers as shown in [43]. The proposed methods were compared with five existing methods. In a comparative experiment, the proposed methods achieved the highest accuracy. Moreover, the proposed methods using image restoration, halo effect removal, and skeleton generation were effective and efficient for human action recognition.

Because the existing state-of-the art methods used the images where the front or back side of body area is captured by camera as shown in Figure 8 (a), joint positions can be easily detected from the skeleton image. However, our database frequently includes the cases where the joint positions are difficult to be detected as shown in Figure 8(f) and the right people of Figures 11(j) and (k). Therefore, we use the skeleton image for the input to CNN-LSTM instead of joint positions.

In further studies, we will focus on the removal or intensity reduction of halo effects on thermal images, which are caused by more diverse objects and machines in various environments. We will also develop a method for improving the processing time using a lighter model with fewer parameters and CycleGAN and CNN-LSTM layers.

REFERENCES

[1] 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.

[2] D. Toth and T. Aach, "Detection and recognition of moving objects using statistical motion detection and Fourier descriptors," in *Proc. IEEE Int. Conf. Image Anal. Process.*, Mantova, Italy, Sep. 2003, pp. 430–435.

[3] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Miami, FL, USA, Jun. 2009, pp. 58–65.

[4] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th IEEE Int. Conf. Pattern Recognit.*, Cambridge, U.K., Aug. 2004, pp. 32–36.

[5] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nice, France, Oct. 2003, pp. 432–439.

[6] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.

[7] L. Wang, Y. Qiao, and X. Tang, "Motionlets: Mid-level 3D parts for human motion recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Porland, OR, USA, Jun. 2013, pp. 2674–2681.

[8] O. Arandjelović, "Contextually learnt detection of unusual motion-based behaviour in crowded public spaces," in *Proc. 26th Int. Symp. Comput. Inf. Sci.*, London, U.K., Sep. 2011, pp. 403–410.

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

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

[11] J. Liu and N. Zheng, "Gait history image: A novel temporal template for gait recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Beijing, China, Jul. 2007, pp. 663–666.

[12] W. Kim, J. Lee, M. Kim, D. Oh, and C. Kim, "Human action recognition using ordinal measure of accumulated motion," *Eur. J. Adv. Signal Process.*, vol. 2010, pp. 1–12, Apr. 2010.

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

[14] T. H. W. Lam, K. H. Cheung, and J. N. K. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recognit.*, vol. 44, no. 4, pp. 973–987, Apr. 2011.

[15] W. K. Wong, H. L. Lim, C. K. Loo, and W. S. Lim, "Home alone faint detection surveillance system using thermal camera," in *Proc. Int. Conf. Comput. Res. Develop.*, Kuala Lumpur, Malaysia, May 2010, pp. 747–751.

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

[17] D. Zhang, Y. Wang, and B. Bhanu, "Ethnicity classification based on gait using multi-view fusion," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 108–115.

[18] R. B. Rusu, J. Bandouch, Z. C. Marton, N. Blodow, and M. Beetz, "Action recognition in intelligent environments using point cloud features extracted from silhouette sequences," in *Proc. 17th IEEE Int. Symp. Robot Hum. Interact. Commun.*, Munich, Germany, Aug. 2008, pp. 267–272.

[19] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Jun. 2014, *arXiv:1406.2199*. [Online]. Available: https://arxiv.org/abs/1406.2199

[20] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 4, pp. 498–509, Aug. 2016.

[21] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 416–425.

[22] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proc. ACM Int. Conf. Multimedia*, Amsterdam, The Netherlands, Oct. 2016, pp. 102–106.

[23] 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.

[24] 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.

[25] I. Misra, C. L. Zitnick, and M. Hebert, "Unsupervised learning using sequential verification for action recognition," Mar. 2016, *arXiv:1603.08561v1*. [Online]. Available: https://arxiv.org/abs/1603.08561v1
[26] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[27] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1110–1118.
[28] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016, pp. 3697–3703.
[29] 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.*, Santa Rosa, CA, USA, Mar. 2017, pp. 148–157.
[30] 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.
[31] 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: https://arxiv.org/abs/1707.02356
[32] B. Brattoli, U. Büchler, A.-S. Wahl, M. E. Schwab, and B. Ommer, "LSTM self-supervision for detailed behavior analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3747–3756.
[33] 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.
[34] M. Baccouche, F. Mamalet, C. Wolf, C. García, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. Int. Workshop Hum. Behav. Understand.*, Amsterdam, The Netherlands, Nov. 2011, pp. 29–39.
[35] 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: https://arxiv.org/abs/1611.06067
[36] A. Al-Naji, K. Gibson, S.-H. Lee, and J. Chahl, "Real time apnoea monitoring of children using the Microsoft Kinect sensor: A pilot study," *Sensors*, vol. 17, no. 2, p. 286, 2017.
[37] S. Liu, L. Kong, and H. Wang, "Human activities recognition based on skeleton information via sparse representation," *J. Comput. Sci. Eng.*, vol. 12, no. 1, pp. 1–11, 2018.
[38] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 1297–1304.
[39] C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Shanghai, China, Mar. 2016, pp. 2712–2716.
[40] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," Nov. 2017, *arXiv:1611.08050*. [Online]. Available: https://arxiv.org/abs/1611.08050
[41] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," Jan. 2016, *arXiv:1602.00134*. [Online]. Available: https://arxiv.org/abs/1602.00134
[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.*, Philadelphia, PA, USA, Jul. 2017, pp. 167–176.
[43] (2019). *Dongguk CNN Stacked LSTM and CycleGAN for Action Recognition, Generated Data, and Dongguk Activities & Actions Database (DA&A-DB2)*. [Online]. Available: http://dm.dgu.edu/link.html
[44] FLIR Systems. (2019). *FLIR Tau 2*. [Online]. Available: https://www.flir.com/products/tau-2/
[45] J. H. Lee, J.-S. Choi, E. S. Jeon, Y. G. Kim, T. T. Le, K. Y. Shin, H. C. Lee, and K. R. Park, "Robust pedestrian detection by combining visible and thermal infrared cameras," *Sensors*, vol. 15, pp. 10580–10615, May 2015.
[46] 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: https://arxiv.org/abs/1703.10593
[47] T. Abu-Ain, S. N. H. S. Abdullah, B. Bataineh, and K. Omar, "A fast and efficient thinning algorithm for binary images," *J. ICT Res. Appl.*, vol. 7, no. 3, pp. 205–216, 2013.
[48] M. H. Nguyen, S.-H. Kim, H. J. Yang, and G. S. Lee, "Stroke width based skeletonization for text images," *J. Comput. Sci. Eng.*, vol. 8, no. 3, pp. 149–156, 2014.
[49] T. Y. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," *Commun. ACM*, vol. 27, no. 3, pp. 236–239, 1984.
[50] Z. Guo and R. W. Hall, "Parallel thinning with two-subiteration algorithms," *Commun. ACM*, vol. 32, no. 3, pp. 359–373, 1989.
[51] 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: https://arxiv.org/abs/1603.08155
[52] A. Ray, S. Rajeswar, and S. Chaudhury, "Text recognition using deep BLSTM networks," in *Proc. 8th Int. Conf. Adv. Pattern Recognit.*, Kolkata, India, Jan. 2015, pp. 1–6.
[53] D. Liu, M. Ye, X. Li, F. Zhang, and L. Lin, "Memory-based gait recognition," in *Proc. Brit. Mach. Vis. Conf.*, York, U.K., Sep. 2016, pp. 82.1–82.12.
[54] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3156–3164.
[55] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition," Oct. 2016, *arXiv:1610.09975*. [Online]. Available: https://arxiv.org/abs/1610.09975
[56] L. Wu, C. Shen, and A. van den Hengel, "Convolutional LSTM networks for video-based person re-identification," Jun. 2016, *arXiv:1606.01609v1*. [Online]. Available: https://arxiv.org/abs/1606.01609v1
[57] A. Zhao, L. Qi, J. Dong, and H. Yu, "Dual channel LSTM based multi-feature extraction in gait for diagnosis of Neurodegenerative diseases," *Knowl. Based Syst.*, vol. 145, pp. 91–97, Apr. 2018.
[58] (2005). *Recognition of Human Actions*. [Online]. Available: http://www.nada.kth.se/cvap/actions/
[59] (2007). *Actions as Space-Time Shapes*. [Online]. Available: http://www.wisdom.weizmann.ac.il/ vision/SpaceTimeActions.html
[60] (2018). *IXMAS Actions—New Views and Occlusions*. [Online]. Available: http://cvlab.epfl.ch/data/ixmas10
[61] Nvidia. (2019). *NVIDIA Titan X*. [Online]. Available: https://www.nvidia.com/en-us/geforce/products/10series/titan-x-pascal/
[62] (2019). *Keras: The Python Deep Learning Library*. [Online]. Available: https://keras.io/
[63] OpenCV. (2019). *OpenCV: Open Source Computer Vision*. [Online]. Available: http://opencv.org/
[64] Wikipedia Foundation. (2019). *Mean Squared Error*. [Online]. Available: https://en.wikipedia.org/wiki/Mean_squared_error
[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[66] 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, p. 1010, 2016.
[67] 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.
[68] (2019). *FLIR Thermal Dataset*. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/
[69] (2019). *Bertin Tech. and ECA (Dataset #1 for Thermal Images)*. [Online]. Available: https://robin.inrialpes.fr/testsdef1.php
[70] (2019). *FLIR System AB Thermal Dataset*. [Online]. Available: http://www.csc.kth.se/~atsuto/IRimagesDataset/pages/downloads.html
[71] (2019). *Visible-Infrared Database*. [Online]. Available: http://www02.smt.ufrj.br/~fusion/
[72] (2019). *CVonline: Image Databases*. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm
[73] (2019). *OTCBVS Benchmark Dataset Collection*. [Online]. Available: http://vcipl-okstate.org/pbvs/bench/
[74] (2019). *AAU VAP Trimodal People Segmentation Dataset*. [Online]. Available: https://www.kaggle.com/aalborguniversity/trimodal-people-segmentation
[75] (2019). *KAIST Multispectral Pedestrian Detection Benchmark*. [Online]. Available: https://soonminhwang.github.io/rgbt-ped-detection/

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 Paichai University, Daejeon, South Korea, in 2014, and the Ph.D. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2019, where he has been a Professor with the Division of Electronics and Electrical Engineering, since March 2019. His research interests include biometrics and pattern recognition. He designed the entire system and wrote the original draft of the paper.

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

TUYEN DANH PHAM received the B.S. degree in electronics and telecommunications from the Hanoi University of Science and Technology (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, where he has been a Professor with the Division of Electronics and Electrical Engineering, since March 2017. His research interests include image processing, biometrics, and deep learning. He helped experiments and analysis.

CHANHUM PARK received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2018, where he is currently pursuing a combined course of his M.S. and Ph.D. degrees in electronics and electrical engineering. His research interests include image processing, biometrics, and deep learning. He helped with the experiments and training of GAN.

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