International Journal of Innovative
Computing, Information and Control
Volume **19**, Number **6**, December **2023**

ICIC International ©2023 ISSN 1349-4198
pp. 1861–1875

# A CNN-BASED METHOD FOR HUMAN ACTION ANALYSIS USING NIGHTTIME INFRARED IMAGES

Yaru Liu, Kai Matsui, Yoichi Kageyama*, Hikaru Shirai and Chikako Ishizawa

Graduate School of Engineering Science
Akita University
1-1 Tegata Gakuen-machi, Akita-shi, Akita 010-8502, Japan
d8522008@s.akita-u.ac.jp; kmatsui19006@gmail.com; {shirai; ishizawa}@ie.akita-u.ac.jp
*Corresponding author: kageyama@ie.akita-u.ac.jp

Received March 2023; revised July 2023

Abstract. *Compared with daytime, more pedestrian-vehicle traffic accidents occur late at night. To reduce the number of accidents, we studied the reasons behind this trend. We analyzed specific human actions using thermal infrared images because they capture object characteristics at night more effectively than visible images. We propose a novel convolutional neural network (CNN)-based model (i.e., VGG16-based) that effectively detects pedestrian actions using image enhancement for infrared image data. The images were acquired using a thermal infrared camera in an outdoor environment in Japan at night. Initially, our proposed method achieved a high detection precision of approximately 84% and 96% on average for infrared images captured at high and low temperatures, respectively. Using image standardization, we managed to improve the detection precision at high temperatures to be more than 90%. Furthermore, using 15187 processed images after standardization, our proposed method achieved an average precision of 0.909, outperforming the Faster R-CNN model, with an improvement of 0.107.*
**Keywords:** Human action analysis, Infrared image, Nighttime, CNN, Image enhancement, Outdoor environment

1. **Introduction.** Japan has the highest elderly adult population (65 years and above) to total population ratio worldwide [1-4]. By 2030, about one-third of the total population will be aged over 65 years [1,5,6]. Consequently, the number of people aged 65 years and older will grow the fastest in the driving population [7,8]. However, the rates of motor vehicle traffic accidents are higher for this segment and increase exponentially for those aged 75 or older [7,9,10]. In particular, under the complex conditions of the nighttime environment, the elevated risk of fatal crashes is two to four times greater than that during daytime [11-15]. According to analyses of accident databases, the reason for the high incidence of traffic accidents at night is poor visibility due to bright road lighting and glare conditions of the nighttime driving environment, rather than increased fatigue or alcohol consumption [11,12,16]. Therefore, to reduce the number of vehicle traffic accidents, there is an urgent need for developing and improving existing advanced driver assistance systems (ADASs) [17] and enhancing road visibility for the elderly.

Previous studies have investigated the ways to reduce vehicle traffic accidents and deaths due to speed-limit violations at night. For example, researchers introduced a novel method that extracts speed-limit signs using the Hough transform [18] and recognizes them through template matching. Experimental results have verified the effectiveness of this method in recognizing the speed limits on signs, with high sensitivity to changes

DOI: 10.24507/ijicic.19.06.1861

in the states of speed-limit signs at night [19-21]. However, to considerably reduce traffic accidents, identifying the speed limits on signs is insufficient because detecting pedestrians on the road is equally important. Therefore, other studies have developed methods for detecting pedestrians on the road using the histogram of oriented gradients (HOG) [22,23] and support vector machines (SVMs) [24-27]. Although results have demonstrated high accuracy for detecting pedestrians, owing to the large number of outdoor environments and human behavior patterns, detailed studies on human action analysis at night have not yet been completed.

In recent decades, deep learning models have achieved great success in the study of human behavior classification [28,29]. In particular, deep learning models using convolutional neural networks (CNNs) have been demonstrated to be powerful tools that eliminate the need for handcrafted feature extraction. For instance, Lv et al. proposed an efficient CNN-based approach for accurate crowd counting in complex crowd scenes [30]. However, most of these studies used visible images to train their models while ignoring illumination changes, poor imaging light, shadows, background clutter, and occlusion of objects. Tarmizi and Aziz developed a vehicle detection and computation model with modified CNN, which achieved an accuracy of 94.3% during the daytime but only 61.4% at night [31]. Compared with visible images, thermal infrared images can overcome the aforementioned challenges. For example, Dai et al. developed a self-learning softmax combined with a nine-layer CNN model to realize nighttime PR based on NIR images with testing accuracy as high as 94.49% [32]. Liu et al. formulated a novel IR pose dataset for self-regulated learning (IRPSRL) and proposed a simple CNN-based model to validate the developed dataset; the proposed EDE method obtained great accuracy in both robustness and prediction [33]. Nevertheless, although thermal infrared images significantly produce sufficient contrast between the object and the background images, the images are greatly influenced by outdoor temperature; particularly, when the temperature is high, such as in summer, human images are difficult to differentiate from the background. Furthermore, other problems such as vehicle driving and image noise occur in human action. Therefore, these problems were considered in our study because insignificant research has been conducted to improve action recognition at different temperatures in different seasons.

Human detection and action classification have the same principle as a general object. Dai et al. utilized the Faster R-CNN algorithm based on the ResNet-50 architecture for pedestrian detection, and they obtained an accuracy of 79.45%; the detection miss rate was 19.23% [34]. Dwi and Suryadiputra improved the performance of the original VGG16 model to detect and classify vehicles at nighttime, with the validation accuracy increasing from 86.40% to 98.30% [35]. Therefore, the present study used a CNN algorithm based on the VGG16 architectural approach to detect humans and classify actions at nighttime. In summary, the main contributions of this paper are as follows.

i) We acquired datasets for four actions (standing, squatting, bending, and walking) from four subjects in various environments in Japan at night: turned-on and -off streetlights, high temperature (24.6°C-26.8°C), and low temperature (3.8°C-5.5°C) to perform human action analysis at nighttime.

ii) We propose a model that improves the performance of the original VGG16 model. In the proposed model, the intermediate layers can effectively select human action features from infrared image data in different environments (varying illumination conditions and temperatures).

iii) We performed image standardization [36] on infrared images to improve robustness for a high-temperature environment; subsequently, we evaluated the accuracy for four types of actions from four subjects using cross-validation [37]. The validation accuracy increased from 80.20% to 90.90% using the proposed model.

Section 2 of this paper introduces the proposed method and VGG16-based model. Section 3 presents the infrared thermal camera, dataset, and experimental results. Finally, Section 4 presents the conclusions of this study.

## 2. Materials and Methods.

2.1. **Overview of the proposed method.** The flowchart of the proposed method is illustrated in Figure 1. First, the region of the person in the thermal infrared data was detected and extracted using the HOG-SVM method. Next, gamma correction [38,39] was used as a preprocessing step to clarify the outline of the person. Data augmentation [40,41] techniques were applied to expanding the limited datasets and utilizing the capabilities of big data, thus improving the model performance. Then we trained the VGG16 CNN model on our dataset to learn human action features [42,43]. The obtained recognition CNN model was used to determine human actions (standing, squatting, and walking). Furthermore, to improve the detection precision at high temperatures, we performed image standardization. Finally, we used cross-validation to evaluate the performance of the proposed model.

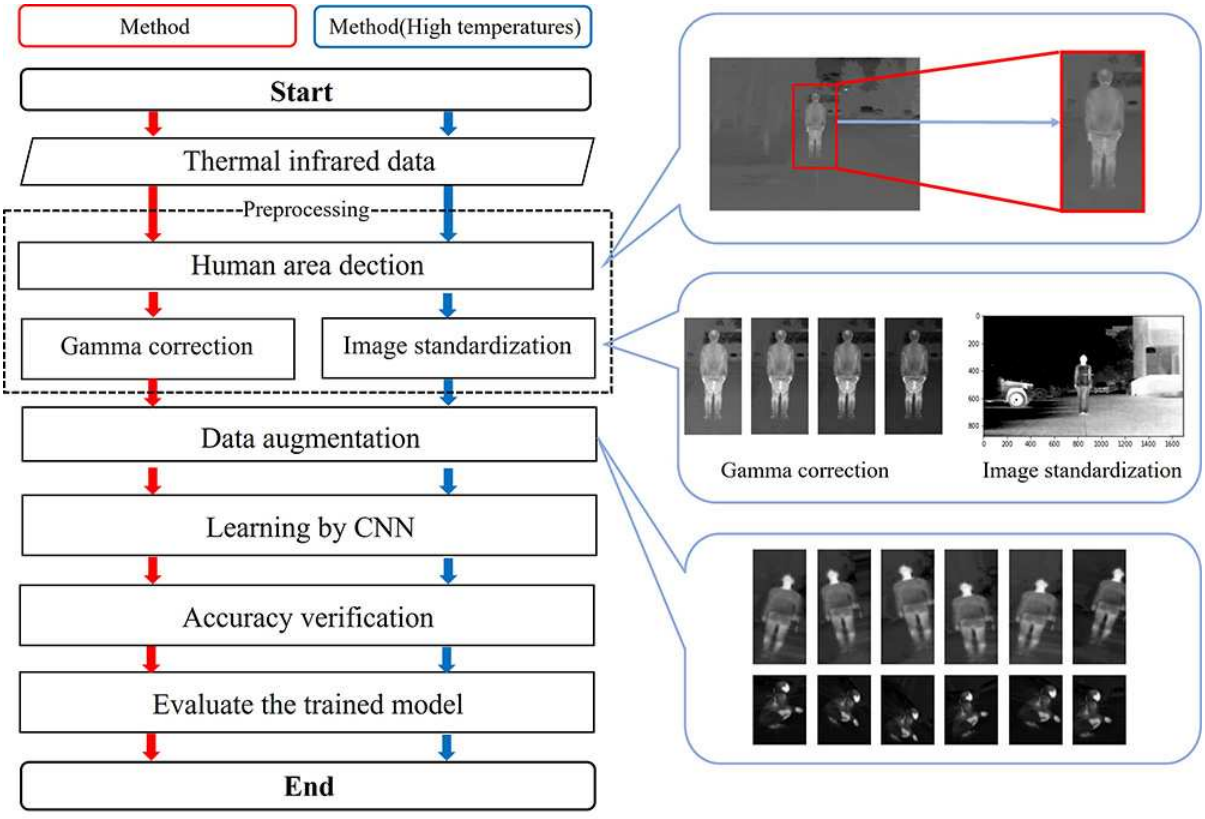


FIGURE 1. Flowchart of the proposed method

2.2. **Human area detection.** In this study, the HOG-SVM method was used to determine the person region in the thermal infrared image. HOG is a feature descriptor used in computer vision and image processing for object detection. We divided the target image into $8 \times 8$ pixel cells and then computed gradient histograms for each pixel within these cells (the gradient direction was divided into 9 directions). Subsequently, we considered groups of $2 \times 2$ cells as a block and normalized the image within each block. This process involves extracting features from all the blocks to generate the HOG feature vector [22,23].

SVM is a supervised machine learning method for pattern recognition. Once the image features were extracted into HOG feature values, we utilized these feature values to train a pre-packaged SVM classifier provided by OpenCV [24-27]. Figures 2(a)-2(f) show an example of the human area detection data.

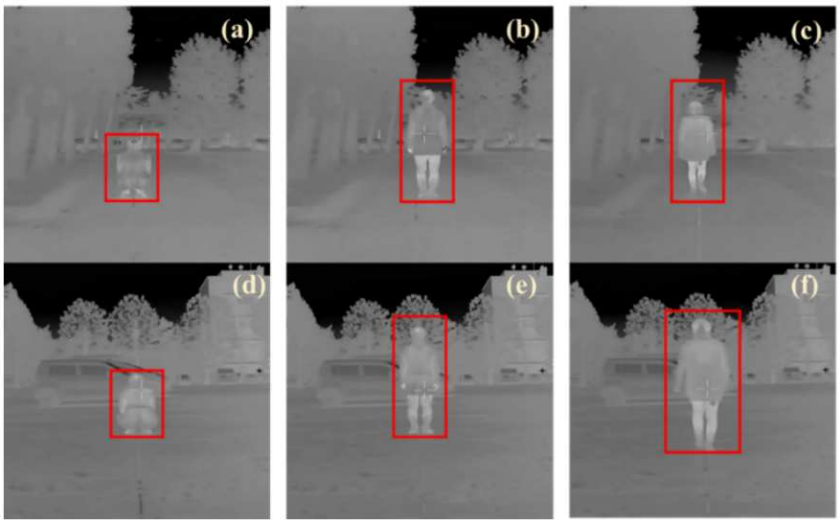


FIGURE 2. An example of human detection in the thermal infrared image when the temperatures are low. (a)-(c): with streetlights on; (d)-(f): with streetlights off.

2.3. **CNN-based model.** CNNs have been implemented in image classification, segmentation, and object detection. CNNs automatically detect important features without human supervision and are computationally efficient. Several CNN architectures have been proposed in [30,31,42,43]. In particular, in the ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC-2014), the VGG16 architecture achieved 92.7% top-5 test accuracy and secured the first and second places in the localization and classification tracks, respectively. Moreover, compared to other architectures, VGG16 shows better generalization capability, thus overcoming the problem of overfitting. Owing to its practical effectiveness, our method is based on the VGG16 architecture by adjusting its hidden layers for thermal infrared images. In our model, we changed the number of layers and the layer parameters. Additionally, the convolutional layers (filter size: $3 \times 3$) were utilized for the resized input image ($150 \times 150$ pixels) to improve the learning speed. Three types of output exist for the convolution layer, which were set to 8, 16, and 32. For the pooling layer, max pooling (filter size: $2 \times 2$) was employed. After the two fully connected layers, the state of human action (stand, squat, and walk) was discriminated in the output layer.

Figure 3 illustrates the VGG16-based model adopted in this study. The model structure is listed in detail in Table 1. The rectified linear unit (ReLU) was used as an activation function in each hidden layer [44], with the calculated results input to the next hidden layer. The activation function determines how the weighted sum of the input is transformed into an output from a neuron in a layer to the next. Nevertheless, deep learning models face the common problem of overfitting, in which the model learns the noise in the training data, and thus cannot generalize to unknown data [45,46]. To address this problem, we used the dropout regularization technique [47,48]. The dropout rate was set to 0.5 and was installed after each fully connected layer. Finally, in the output layer, the softmax function [49,50] was used to determine the human action.

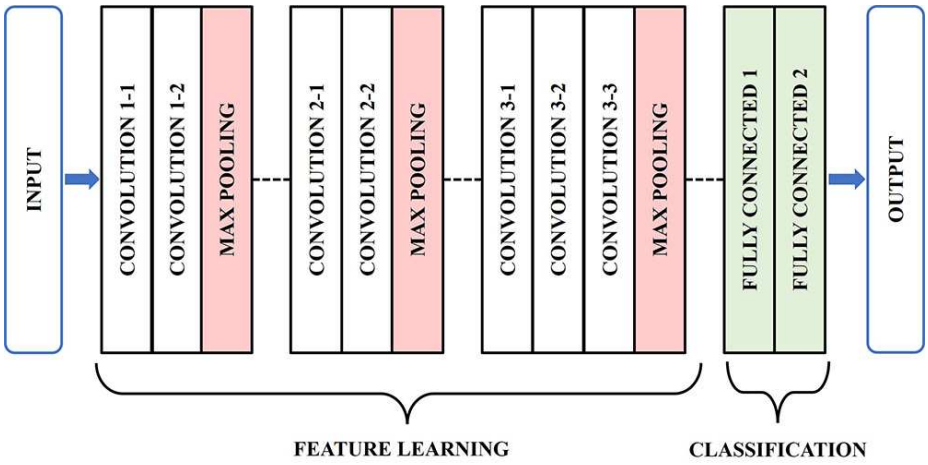


FIGURE 3. Our VGG16-based model for human action detection. It comprises seven convolutional, three pooling, and two fully connected layers.

TABLE 1. Structure of the VGG16-based model for human action analysis

| Layer | | Feature map | Output size | Filter size | Activation function |
|---|---|---|---|---|---|
| Input | Image | 1 | $150 \times 150 \times 1$ | – | – |
| 1 | $2 \times$ Con2 | 8 | $148 \times 148 \times 8$ | $3 \times 3$ | ReLU |
| | Max pooling | 8 | $73 \times 73 \times 8$ | $2 \times 2$ | ReLU |
| 2 | $2 \times$ Con2 | 16 | $71 \times 71 \times 16$ | $3 \times 3$ | ReLU |
| | Max pooling | 16 | $34 \times 34 \times 16$ | $2 \times 2$ | ReLU |
| 3 | $3 \times$ Con2 | 32 | $32 \times 32 \times 32$ | $3 \times 3$ | ReLU |
| | Max pooling | 32 | $14 \times 14 \times 32$ | $2 \times 2$ | ReLU |
| 4 | Fully-Connect | – | 128 | – | ReLU |
| | Dropout | – | 128 | – | – |
| 5 | Fully-Connect | – | 256 | – | ReLU |
| | Dropout | – | 256 | – | – |
| Output | Fully-Connect | – | 3 | – | Softmax |

In this study, we developed a new model tailored for training thermal infrared-image data based on the VGG16 model. In this VGG16-based model, we adjusted the number of layers in the intermediate layer to effectively capture features of the targets present in thermal infrared images. Additionally, we fine-tuned the training parameters of the VGG16-based model, significantly enhancing the efficiency of machine learning.

2.4. **Human motion analysis.** First, to refine the outline of the person in thermal infrared images, gamma correction (gamma value: 3.1-3.5, in 0.1 increments) was applied as a preprocessing step as per the following equation:

$$Y = 225 \times \left(\frac{x}{225}\right)^{\gamma}, \tag{1}$$

where $\gamma$ is the gamma value, and $x$ and $Y$ represent the image pixels before and after correction. Moreover, thermal infrared images captured at high temperatures were preprocessed by image standardization using the following equations:

$$Standard = \frac{(x - mean)}{adjusted\_stddev}, \tag{2}$$

$$adjusted\_stddev = \max\left(stddev, \frac{1}{sqrt(N)}\right), \tag{3}$$

where $x$ represents the pixels of a thermal infrared image, *mean* is the average pixel value of the thermal infrared image, *stddev* denotes the standard deviation of pixel values, $N$ is the image element count, and *sqrt* is the square root of the image element count.

After preprocessing, the obtained images were used as training data for the proposed VGG16-based model. To increase the diversity of data available for training and make the model invariant to geometric transformations of the input, data augmentation techniques (e.g., rotation, enlargement, and reduction) were applied. Thereby, our dataset expanded to include 51892 (with streetlights on) and 51676 (with streetlights off) images at high temperatures, and 49075 (with streetlights on) and 53940 (with streetlights off) images at low temperatures. Finally, the model was applied to the test dataset to verify the results.

## 3. Experiment and Results.

3.1. **Camera.** An infrared thermal imaging camera (manufactured by D-eyes Co., Ltd., "ultra-high sensitivity + far-infrared" 2-in-1 camera WCAM001-AU, 640 × 480 pixels) was used to acquire video data in an outdoor environment at night. The infrared thermal camera was equipped with a Microbolometer (uncooled) sensor and could acquire moving images up to 640 × 480 (VGA format). An external view of the equipment is shown in Figure 4. The main specifications of the equipment are listed in Table 2.


FIGURE 4. External view of the infrared thermal imaging camera and switcher (standard type)

TABLE 2. Main specifications of the equipment

| Sensor | | Camera | |
|---|---|---|---|
| Type | Microbolometer (uncooled) | Lens | f:19 mm |
| Pixel | 640 × 480 (VGA format) | Output | HDMI |
| Pixel size | 17 μm × 17 μm | Power | DC 12 V, 400 mA |
| Temperature resolution | < 50 mK°/−30°C-+150°C<br>< 30 mK°/normal temperature | Operating temperature | −20°C-+45°C |
| Sensitivity wave | 8-14 μm | Save temperature | −30°C-+70°C |

3.2. **Data acquisition.** Outdoor temperatures in September and November 2021 were between 24.6°C-27.8°C and 3.8°C-5.5°C, respectively. The camera shooting distance was 5.0-15.0 m, illumination intensity was 0.17-0.68 lux with streetlights off and 3.23-3.73 lux with streetlights on. We recruited two subjects of East Asian descent: a male and a female in their 20 s. The images were acquired at different temperatures with the subjects performing three actions: standing, squatting, and walking. During the data acquisition, the subjects were instructed to emulate the movement postures of elderly individuals. The standing and squatting actions were captured with the subjects 10.0 m from the camera in the front, back, right-side, and left-side directions. Walking was captured at 5.0-15.0 m from the camera in the front and back directions. Considering the complex situations in the streets, the data were captured with streetlights on and off. Figure 5 shows the data-acquisition environment of the experiment.

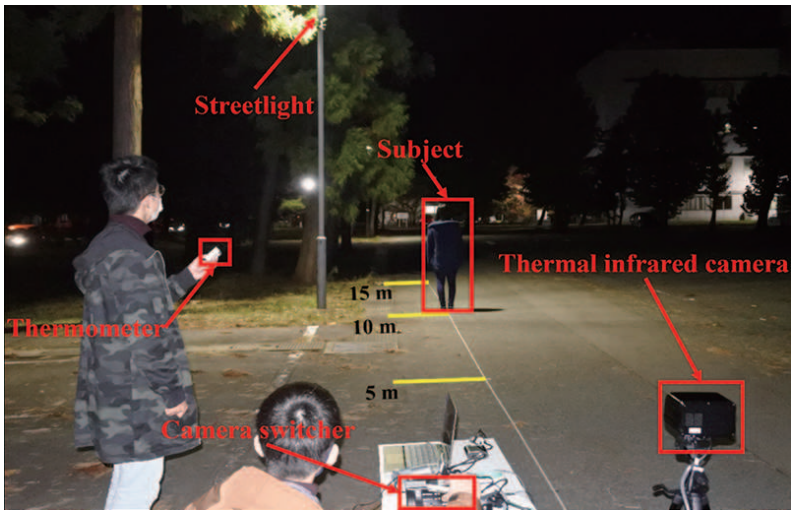


FIGURE 5. Data-acquisition environment (streetlight on; low temperature; subject standing)

The process of data collection was as follows.

1) The subject stood at 10.0 m from the camera (front, back, right-side, and left-side directions).

2) The subject squatted at 10.0 m from the camera (front, back, right-side, and left-side directions).

3) The subject walked at 5.0-15.0 m from the camera (front and back directions).

4) All aforementioned three steps were repeated under scenarios: streetlights next to the subject were on or off.

The data used in this investigation were acquired in accordance with ethical regulations concerning studies involving humans at Akita University, Japan.

3.3. **Examination.** The proposed model was trained for 100 epochs using an Intel Core i7-9700k 3.6 GHz processor with NVIDIA GeForce RTX 2080 Ti 27GB GPU, programming language used Python and learning frameworks used Keras. The features of acquired infrared images might provide different types of information in complicated outdoor environments. Additionally, input infrared images might be affected by the model, thus reducing the precision of human action analysis. Therefore, the discriminant precision of the proposed model and the analysis of human actions were examined with regard to differences in outdoor conditions. Specifically, as the test dataset, data collected at low temperatures were validated using 7200 images each for the standing and squatting

actions and 1000 images for the walking actions, whereas data collected at high temperatures were validated using 6662, 6254, and 1700 images for the standing, squatting and walking actions, respectively.

3.4. **Results and evaluation.** The proposed VGG-16-based model was evaluated on the test dataset. True positives (TP) represent outcomes where the model correctly identified the human action. False positives (FP) represent outcomes where the model incorrectly predicted the positive examples. Precision is calculated as

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

For data collected at high temperatures (24.6°C-27.8°C), Table 3 lists the precision results for the model on the test dataset with streetlights on and off. The model achieved precision of 0.825 or higher for every action. The average precision results for human action detection (standing, squatting, walking) were approximately 0.848 and 0.841 with and without streetlights, respectively, with a total average of 0.844.

Table 3. Precision results for the test dataset at high temperatures

| Action | Streetlights on | Streetlights off | Average |
|---|---|---|---|
| Standing | 0.861 | 0.854 | 0.857 |
| Squatting | 0.832 | 0.825 | 0.828 |
| Walking | 0.853 | 0.844 | 0.848 |
| Average | 0.848 | 0.841 | 0.844 |

For data collected at low temperatures (3.8°C-5.5°C), Table 4 lists the precision results for the model on the test dataset with streetlights on and off. The model achieved precision of 0.954 or higher for every action. The average precision results for human action detection (standing, squatting, walking) were approximately 0.962 and 0.961 with and without streetlights, respectively, with a total average of 0.961.

Table 4. Precision results for the test dataset at low temperatures

| Action | Streetlights on | Streetlights off | Average |
|---|---|---|---|
| Standing | 0.961 | 0.954 | 0.957 |
| Squatting | 0.956 | 0.968 | 0.962 |
| Walking | 0.971 | 0.962 | 0.966 |
| Average | 0.962 | 0.961 | 0.961 |

Thereby, the proposed method achieved high discrimination results (precision greater than 0.841) for each action. These results suggest that the proposed method can distinguish between the human actions of standing, squatting, and walking in an outdoor environment at different temperatures and under varying lighting conditions at night. However, focusing on the difference in precision based on temperature, the detection precision for data collected at high temperatures was 0.117 lower on average than that of data collected at low temperatures. For data collected at high temperatures, the discrimination precision was lower because of the small temperature difference between the human body and the external environment and the similar pixel value between the person and the infrared image background. Conversely, for data collected at low temperatures, the temperature difference was large; thus, the person's outline in the background of the image became clearer, making it easier to discriminate. There were also cases in which the standing state was erroneously identified as a walking state. This is because of the

similarity between the standing and walking states. Additionally, since the subjects were far from the camera, the model could not accurately capture the action features. Furthermore, since the features of the image background are similar when extracted from the hidden layers in CNN models, the squatting state was occasionally recognized as a standing state.

3.5. **Data standardization.** As the detection precision of data collected at high temperatures was lower than those collected at low temperature, the high-temperature data were optimized by image standardization to refine the person's outline area. Furthermore, to evaluate the usability of the performed image standardization, the high-temperature infrared image data were enlarged to include four types of actions (i.e., standing, squatting, bending, and walking) from four Subjects A-D. The model was then trained and evaluated on the standardized dataset using cross-validation, the precision results were compared to those of the unstandardized dataset.

3.5.1. *Data collection.* We expanded our database by adding this data-collection experiment. The data were collected in the same environment as presented in Subsection 3.2. In this data collection experiment, we added four additional subjects of East Asian descent: two men and two women in their 20 s. Figure 6 shows an example of the acquired data. To capture more features of pedestrians while walking, we made some adjustments. During the walking of the subjects, we added the scenario of walking sideways (Figure 6(d)). Therefore, to capture the complete motion state, we adjusted the distance to be between 10 m and 15 m. The details are as follows.

1) The subject stood at 10.0 m from the camera (front, back, right-side, and left-side directions).

2) The subject squatted at 10.0 m from the camera (front, back, right-side, and left-side directions).

3) The subject bended at 10.0 m from the camera (front, back, right-side, and left-side directions).

4) The subject walked in the front, back and side directions at 10.0-15.0 m from the camera.

5) All aforementioned four steps were repeated under two scenarios: streetlights next to the subject were on or off.

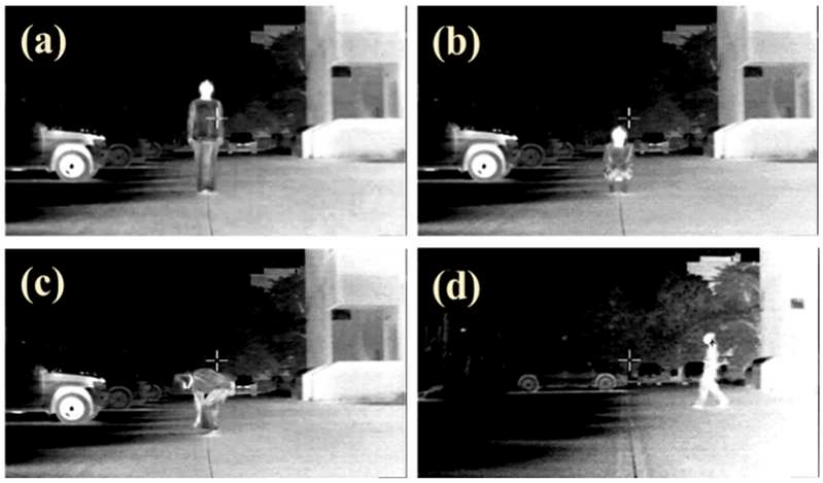


FIGURE 6. Data collection for four actions: (a) Standing, (b) squatting, (c) bending, and (d) walking

3.5.2. *Implementation.* To evaluate the usefulness of the proposed standardization process, an action detection experiment was conducted using the data obtained in Subsection 3.5.1 by cross-validation. Cross-validation was performed on eight different subsets by changing the combination of Subjects A-D in the training and test datasets. For implementation, three subjects were set as the training dataset, and the remaining subjects were set as the test dataset. The precision results for the original images (i.e., before standardization) and processed images (after standardization) are listed in Tables 5 and 6, respectively. The average precision for processed images captured at high temperatures (0.909) improved by 0.074 compared to that of original images (0.835). Table 7 compares the average detection precision for original and processed images by each action. It shows that for each action, the detection precision can be improved when the standardization process is applied under the two scenarios: streetlights are on or off.

These results suggest that applying the standardization process for data with low temperature differences is effective in improving detection precision.

Table 5. Precision results for original images

| Action | Streetlights on | | | | Streetlights off | | | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | A | B | C | D | precision |
| Standing | 0.852 | 0.834 | 0.825 | 0.861 | 0.856 | 0.867 | 0.849 | 0.893 | 0.854 |
| Squatting | 0.822 | 0.818 | 0.838 | 0.857 | 0.828 | 0.824 | 0.804 | 0.859 | 0.831 |
| Bending | 0.801 | 0.795 | 0.789 | 0.848 | 0.812 | 0.796 | 0.778 | 0.825 | 0.805 |
| Walking | 0.849 | 0.844 | 0.832 | 0.873 | 0.838 | 0.842 | 0.854 | 0.881 | 0.851 |
| Average | | | | | | | | | 0.835 |

Table 6. Precision results for processed images

| Action | Streetlights on | | | | Streetlights off | | | | Average |
|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | A | B | C | D | precision |
| Standing | 0.937 | 0.976 | 0.933 | 0.931 | 0.952 | 0.966 | 0.947 | 0.911 | 0.944 |
| Squatting | 0.917 | 0.888 | 0.866 | 0.853 | 0.892 | 0.874 | 0.862 | 0.856 | 0.876 |
| Bending | 0.893 | 0.889 | 0.851 | 0.932 | 0.870 | 0.891 | 0.858 | 0.903 | 0.885 |
| Walking | 0.951 | 0.953 | 0.895 | 0.922 | 0.917 | 0.956 | 0.921 | 0.936 | 0.931 |
| Average | | | | | | | | | 0.909 |

Table 7. Comparison between original and processed images by average precision

| Average | Streetlights on | | | | Streetlights off | | | |
|---|---|---|---|---|---|---|---|---|
| precision | Standing | Squatting | Bending | Walking | Standing | Squatting | Bending | Walking |
| Original images | 0.843 | 0.833 | 0.808 | 0.849 | 0.866 | 0.828 | 0.802 | 0.853 |
| Processed images | 0.944 | 0.881 | 0.890 | 0.930 | 0.944 | 0.871 | 0.880 | 0.931 |
| Difference | **+0.101** | **+0.048** | **+0.082** | **+0.081** | **+0.078** | **+0.043** | **+0.078** | **+0.078** |

3.6. **Comparison results with Faster R-CNN.** Using Faster R-CNN, which is an end-to-end object detection algorithm [51], region proposals are learned by the network. Faster R-CNN usually uses architectures such as ResNet [52] for feature extraction, which is highly generalizable in detection tasks. Therefore, the proposed method was compared with Faster R-CNN.

We set up the annotations and labeled the human actions in the infrared images: standing, squatting, bending, and walking. Furthermore, 15187 processed images after standardization were selected as a database from the total data collected at high temperatures. To evaluate the Faster R-CNN model, the data were separated according to the streetlight conditions: on or off. For data collected with streetlights on, 5439 images were set as the training dataset, and 2332 images were set as the test (validation) dataset. For data collected with streetlights off, 5191 images were set as the training dataset, and 2225 images were set as the test (validation) dataset. Tables 8 and 9 list the detection precision results for Faster R-CNN and the proposed model, respectively, for each action. Compared with Faster R-CNN, the average detection precision of the proposed model for standing, squatting, bending, and walking improved by 0.128, 0.074, 0.086 and 0.139, respectively. However, the proposed method achieved an average precision of 0.909, which is slightly higher than that of Faster R-CNN, with an improvement of 0.107. These results suggest that the proposed method outperforms Faster R-CNN in thermal infrared images and can be used to detect human actions at nighttime in diverse environments with high precision.

Table 8. Detection precision results using Faster R-CNN

| Action | Streetlights on | Streetlights off | Average |
|---|---|---|---|
| Standing | 0.801 | 0.831 | 0.816 |
| Squatting | 0.803 | 0.801 | 0.802 |
| Bending | 0.784 | 0.814 | 0.799 |
| Walking | 0.806 | 0.778 | 0.792 |
| Average | | | 0.802 |

Table 9. Detection precision results using the proposed model

| Action | Streetlights on | Streetlights off | Average |
|---|---|---|---|
| Standing | 0.944 | 0.944 | 0.944 |
| Squatting | 0.881 | 0.871 | 0.876 |
| Bending | 0.890 | 0.880 | 0.885 |
| Walking | 0.930 | 0.931 | 0.931 |
| Average | | | 0.909 |

4. **Conclusions.** In this study, we proposed a VGG16-based detection method for human action analysis in various environments and under varying conditions (illumination and temperature) at nighttime. Our results showed that the proposed method can identify human actions in an outdoor environment at night at different temperature and under varying illumination conditions. However, the detection precision for human actions in images captured at high temperatures was 0.117 lower on average than that at low temperatures. By applying image standardization to process infrared image data at high temperatures before training our model, we were able to clarify the contour of a person and improve the detection precision. Finally, our proposed method showed superior performance compared to other methods such as Faster R-CNN.

Our study has some limitations. Although we used infrared images taken at different temperatures in the current study, we wish to point out that the infrared image data could be affected by weather conditions other than temperature, which may reduce detection precision. Therefore, for practicality, it is necessary to use infrared image data acquired under different weather conditions, such as rain and snow. Furthermore, to improve the safety of road users and contribute to ADASs, it is important to detect persons and identify their actions in real time. In the future, we will use deep learning to train a real-time detection model and improve the computational speed.

**Acknowledgment.** This study was supported by a Grant-in-Aid for Scientific and Technological Research from the Suzuki Foundation.

## REFERENCES

[1] M. Naoko and H. Akiyama, Japan: Super-aging society preparing for the future, *The Gerontologist*, vol.51, no.4, pp.425-432, 2011.
[2] N. Kyoko and A. Koizumi, Strategy against aging society with declining birthrate in Japan, *Industrial Health*, vol.54, no.6, pp.477-479, 2016.
[3] N. Mayuko, S. McLean, D. Miyamori, Y. Kakiuchi and H. Ikegaya, Isolation and unnatural death of elderly people in the aging Japanese society, *Science & Justice*, vol.56, no.2, pp.80-83, 2016.
[4] Statistics Bureau of Japan, *Population Estimates: Result of the Population Estimates*, https://www.stat.go.jp/english/data/jinsui/2022np/index.html, Accessed on September 22, 2023.
[5] Y. Ouchi, H. Rakugi, H. Arai, M. Akishita, H. Ito and K. Toba, Redefining the elderly as aged 75 years and older: Proposal from the Joint Committee of Japan Gerontological Society and the Japan Geriatrics Society, *Geriatr. Gerontol. Int.*, vol.17, no.7, pp.1045-1047, 2017.
[6] World Health Organization, *World Health Statistics 2015*, 2015.
[7] K. J. Anstey, J. Wood, S. Lord and J. G. Walker, Cognitive, sensory and physical factors enabling driving safety in older adults, *Clinical Psychology Review*, vol.25, no.1, pp.45-65, 2005.
[8] S. Lyman, S. A. Ferguson, E. R. Braver and A. F. Williams, Older driver involvements in police reported crashes and fatal crashes: Trends and projections, *Injury Prevention*, vol.8, no.2, pp.116-120, 2002.
[9] J. H. Guerrier, P. Manivannan and S. N. Nair, The role of working memory, field dependence, visual search, and reaction time in the left turn performance of older female drivers, *Applied Ergonomics*, vol.30, no.2, pp.109-119, 1999.
[10] *National Institute of Population and Social Security Research: "Projection: Population & Household in Japan"*, http://www.ipss.go.jp/ppshicyoson/j/shicyoson18/1kouhyo/gaiyo_a.pdf, Accessed on March 7, 2023.
[11] J. A. Kimlin, A. A. Black and J. M. Wood, Nighttime driving in older adults: Effects of glare and association with mesopic visual function, *Investigative Ophthalmology & Visual Science*, vol.58, no.5, pp.2796-2803, 2017.
[12] M. C. Puell, A. R. Barrio, C. Palomo-Alvarez, F. J. Gomez-Sanz, A. Clement-Corral and M. J Pérez-Carrasco, Impaired mesopic visual acuity in eyes with early age-related macular degeneration, *Investigative Ophthalmology & Visual Science*, vol.53, no.11, pp.7310-7314, 2012.
[13] K. Lahav, H. Levkovitch-Verbin, M. Belkin, Y. Glovinsky and U. Polat, Reduced mesopic and photopic foveal contrast sensitivity in glaucoma, *Archives of Ophthalmology*, vol.129, no.1, pp.16-22, 2011.
[14] National Highway Traffic Safety Administration, *Motor Vehicle Traffic Crash Fatality Counts and Injury Estimates for 2004 (DOTHS 809923)*, Department of Transportation, Washington, D.C., US, 2005.
[15] D. A. Owens and M. Sivak, Differentiation of visibility and alcohol as contributors to twilight road fatalities, *Human Factors*, vol.38, no.4, pp.680-689, 1996.
[16] J. M. Sullivan and M. J. Flannagan, The role of ambient light level in fatal crashes: Inferences from daylight saving time transitions, *Accident Analysis & Prevention*, vol.34, no.4, pp.487-498, 2002.
[17] O. J. Gietelink, *Design and Validation of Advanced Driver Assistance Systems*, TRAIL Research School, 2007.
[18] M. Takagi and H. Shimoda, *Shinpen-Gazō-Kaiseki-Handobukku: Handbook of Image Analysis*, University of Tokyo Press, Tokyo, Japan, 2004.

[19] Y. Kageyama, H. Kameya, M. Nishida and C. Ishizawa, Recognition of speed limit signs, in night scene images in Japan, *IEEJ Transactions on Electrical and Electronic Engineering*, vol.8, no.S1, pp.88-94, 2013.
[20] Y. Kageyama, K. Suzuki, C. Ishizawa and T. Suzuki, Extraction and recognition of speed limit signs in night-scene videos, *Journal of the Institute of Industrial Applications Engineers*, vol.6, no.1, pp.29-33, 2018.
[21] T. Suzuki, Y. Kageyama and C. Ishizawa, Recognition method for speed limit signs and its applicability in recognition of vehicle entry prohibition signs at night, *IEEJ Transactions on Electrical and Electronic Engineering*, vol.15, no.10, pp.1448-1456, 2020.
[22] M. Bilal and M. S. Hanif, Benchmark revision for HOG-SVM pedestrian detector through reinvigorated training and evaluation methodologies, *IEEE Transactions on Intelligent Transportation Systems*, vol.21, no.3, pp.1277-1287, 2019.
[23] R. P. Yadav, V. Senthamilarasu, K. Kutty and S. P. Ugale, Implementation of robust HOG-SVM based pedestrian classification, *International Journal of Computer Applications*, vol.114, no.19, 2015.
[24] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol.1, pp.886-893, 2005.
[25] Y. Lin, Y. Chan, L. Chuang, L. Fu, S. Huang, P. Hsiao and M. Luo, Near-infrared based nighttime pedestrian detection by combining multiple features, *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp.1549-1554, 2011.
[26] P. Govardhan and U. C. Pati, NIR image based pedestrian detection in night vision with cascade classification and validation, *IEEE Int. Conf. on Advanced Communications*, Ramanathapuram, India, 2015.
[27] W. S. Noble, What is a support vector machine?, *Nature Biotechnology*, vol.12, no.12, pp.1565-1567, 2006.
[28] M. C. Kwon, M. Ju and S. Choi, Classification of various daily behaviors using deep learning and smart watch, *2017 9th International Conference on Ubiquitous and Future Networks (ICUFN)*, pp.735-740, 2017.
[29] S. S. Basha, V. Pulabaigari and S. Mukherjee, An information-rich sampling technique over spatio-temporal CNN for classification of human actions in videos, *Multimedia Tools and Applications*, vol.81, no.28, pp.40431-40449, 2022.
[30] M. Lv, K. Zhang, X. Zheng, W. Yang and M.-Z. Lu, Leverage multi-scale dilated convolutional neural network with global attention feature fusion for crowd counting, *International Journal of Innovative Computing, Information and Control*, vol.18, no.4, pp.1147-1162, 2022.
[31] I. A. Tarmizi and A. A. Aziz, Vehicle detection using convolutional neural network for autonomous vehicles, *International Conference on Intelligent and Advanced System (ICIAS)*, 2018.
[32] X. Dai, Y. Duan, J. Hu, S. Liu, C. Hu, Y. He, D. Chen, C. Luo and J. Meng, Near infrared nighttime road pedestrians recognition based on convolutional neural network, *Infrared Physics & Technology*, vol.97, pp.25-32, 2019.
[33] H. Liu, Y. Chen, W. Zhao, S. Zhang and Z. Zhang, Human pose recognition via adaptive distribution encoding for action perception in the self-regulated learning process, *Infrared Physics & Technology*, vol.114, 103660, 2021.
[34] X. Dai, Y. Duan, J. Hu, S. Liu, C. Hu, Y. He, D. Chen, C. Luo and J. Meng, Multi-task Faster R-CNN for nighttime pedestrian detection and distance estimation, *Infrared Physics & Technology*, vol.115, 103694, 2021.
[35] S. R. Dwi and L. Suryadiputra, Detection and classification of moving vehicle at night with visible camera using deep learning model, *ICIC Express Letters*, vol.17, no.3, pp.339-347, 2023.
[36] M. J. Weinberger, G. Seroussi and G. Sapiro, The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS, *IEEE Transactions on Image Processing*, vol.9, no.8, pp.1309-1324, 2000.
[37] P. Refaeilzadeh, L. Tang and H. Liu, Cross-validation, *Encyclopedia of Database Systems*, pp.532-538, 2009.
[38] H. Farid, Blind inverse gamma correction, *IEEE Transactions on Image Processing*, vol.10, no.10, pp.1428-1433, 2001.
[39] Y. S. Chiu, F. C. Cheng and S. C. Huang, Efficient contrast enhancement using adaptive gamma correction and cumulative intensity distribution, *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pp.2946-2950, 2011.

[40] J. Perez and J. Wang, The effectiveness of data augmentation in image classification using deep learning, *arXiv Preprint*, arXiv: 1712.04621, 2017.
[41] L. Taylor and G. Nitschke, Improving deep learning with generic data augmentation, *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp.1542-1547, 2018.
[42] H. Qassim, A. Verma and D. Feinzimer, Compressed residual-VGG16 CNN model for big data places image recognition, *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pp.169-175, 2018.
[43] A. Canziani, A. Paszke and E. Culurciello, An analysis of deep neural network models for practical applications, *arXiv Preprint*, arXiv: 1605.07678, 2016.
[44] K. Hara, D. Saito and H. Shouno, Analysis of function of rectified linear unit used in deep learning, *2015 International Joint Conference on Neural Networks (IJCNN)*, pp.1-8, 2015.
[45] S. Yeom, I. Giacomelli, M. Fredrikson and S. Jha, Privacy risk in machine learning: Analyzing the connection to overfitting, *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp.268-282, 2018.
[46] R. Roelofs, V. Shankar, B. Recht, S. Fridovich-Keil, M. Hardt, J. Miller and L. Schmidt, A meta-analysis of overfitting in machine learning, *Advances in Neural Information Processing Systems*, vol.32, 2019.
[47] P. Christoffersen and K. Jacobs, The importance of the loss function in option valuation, *Journal of Financial Economics*, vol.72, no.2, pp.291-318, 2004.
[48] S. Wang and C. Manning, Fast dropout training, *International Conference on Machine Learning PMLR*, pp.118-126, 2013.
[49] R. Memisevic, C. Zach, M. Pollefeys and G. E. Hinton, Gated softmax classification, *Advances in Neural Information Processing Systems*, vol.23, 2010.
[50] X. Liang, X. Wang, Z. Lei, S. Liao and S. Z. Li, Soft-margin softmax for deep classification, in *Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science*, D. Liu, S. Xie, Y. Li, D. Zhao and ES. El-Alfy (eds.), Cham, Springer, 2017.
[51] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in Neural Information Processing Systems*, vol.28, 2015.
[52] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

## Author Biography

**Yaru Liu** received the B.E. degree in Computer Science and Engineering from Qingdao Agricultural University, China, in 2019, and he received the M.E. degree in Computer Science and Engineering from Akita University, Japan, in 2022. He is now enrolled in a doctoral program with the Graduate School of Engineering Science in Akita University. His research interests include human sensing and image processing.

**Kai Matsui** received the B.E. and M.E. degrees in Computer Science and Engineering and the Dr. Eng. degree from Akita University, Japan, in 2017, 2019, and 2022, respectively. He is currently working as as an engineer at Suzuki Motor Corporation. His research interests include remote sensing and image processing.

**Yoichi Kageyama** received the B.E. and M.E. degrees in Computer Science and Engineering and the Dr. Eng. degree from Akita University, Japan, in 1995, 1997, and 2001, respectively. He joined Akita University as a Research Associate in 1997. He became an Assistant Professor in 2001 and an Associate Professor in 2004. He is now a Professor with the Department of Mathematical Science and Electrical Electronic Computer Engineering, Graduate School of Engineering Science. His research interests include human sensing, remote sensing, and image processing.

**Hikaru Shirai** received the B.E., M.E., and Dr. E. degrees in Computer Science and Engineering from Akita University, Japan, in 2011, 2013, and 2017, respectively. He joined Akita Electronics Systems Co., Ltd. in 2013. He joined Ricoh IT Solutions Co., Ltd. in 2017. He joined Akita University as a Technical Staff in 2019. He became an Assistant Professor in 2020. He is now a Lecturer. His research interests include remote sensing and image processing.

**Chikako Ishizawa** received the B.E. degree in Chemical Engineering for Resources from Akita University, Japan, in 1992, and joined FUJIFILM Software Co., Ltd. She joined Akita University in 1995. She received a Dr. Eng. degree from Akita University in 2012. She is now a Professor with the Department of Mathematical Science and Electrical Electronic Computer Engineering, Graduate School of Engineering Science. Her research interests include visual information processing and log analysis.