Master's Thesis

Title

Fusion of Thermal and Visible Videos for Heat Trace Detection on Touched Areas

Supervisor

Professor Masayuki Murata

Author

Kento Noguchi

February 2nd, 2023

Department of Information Networking

Graduate School of Information Science and Technology

Osaka University

Master's Thesis

Fusion of Thermal and Visible Videos for Heat Trace Detection on Touched Areas

Kento Noguchi

Abstract

Recently, image recognition AI has been introduced in various industries, such as manufacturing and logistics, for uncrewed and labor-saving operations. Sensors used to capture image data include not only RGB cameras but also thermal cameras and near-infrared cameras. Since these sensors have become relatively inexpensive, the appropriate sensor is selected according to the purpose. Furthermore, some image processing methods that fusion images obtained from sensors capturing different wavelength ranges have also been proposed.

The novel coronavirus (COVID-19) spread has raised considerable interest in countermeasures against infectious diseases. One of the routes of infection is contact transmission. It is known that the virus can remain on touched surfaces for several hours to several days, depending on the material touched, so it is important to avoid touching the areas where other people touched or to disinfect the area by detecting and warning of the touched points.

Several image processing methods using a thermal camera have been proposed for detecting touched points. It is known that thermal videos show a sign that the temperature of the touched area rises, which is called heat trace. Although humans can appear in thermal and RGB videos, heat traces appear only in thermal videos. Therefore, the detection accuracy of heat traces can be improved by combining thermal and visible light videos to distinguish between people and heat traces. In addition, since heat traces disappear and diffuse over time, temporal thermal change is considered effective for heat trace detection.

In this study, we propose the heat trace detection method by fusion of RGB and thermal videos with deep learning models and investigate the effectiveness of temporal and spatial features, contributing to the development of a heat trace detection system to

avoid contact transmission of viruses causing infectious diseases. We designed a simple dataset and 2DCNN and 3DCNN models to investigate the influence of the type of image bands, the time length, and the spatial size of input videos. The results show that when the spatial area is strongly restricted, the accuracy of thermal images is higher than that of the thermal and RGB images. On the other hand, when the spatial size was larger and more human and environmental information was included, the accuracy improved to 93.3 % when both were used, compared to 91.0 % when using only the thermal images. The accuracy of videos was higher than that of single images in any experiment. From these results, we confirmed that the fusion of RGB and thermal images improves the detection accuracy if a certain spatial area can be captured, and temporal information is an important feature for detecting heat traces. The obtained knowledge is helpful for designing a more optimal network to detect heat traces, thus we propose a new network architecture for future work.

Keywords

Image fusion
Multi-band image processing
Thermal imaging
3DCNN
Heat Trace
Coronavirus

Contents

List of Figures

List of Tables

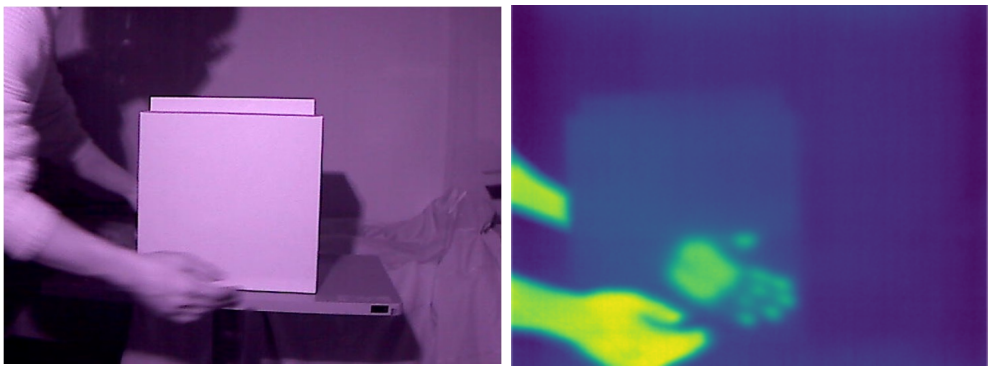
1 Introduction

Image recognition AI is becoming widely penetrating in various industries, including manufacturing and logistics, for automated and labor-saving operations. Various sensors can respectively capture particular wavelength ranges, for example, visible light (VIS) cameras, near-infrared (NIR) cameras, long-wave infrared (LWIR) cameras, multispectral cameras, and so on. Since these sensors are now relatively inexpensive, they can solve a wide range of problems, such as detecting abnormalities in machinery using a thermal camera or measuring the vitality of crops using a multispectral camera.

Additionally, image processing methods that combine sensors of different wavelength ranges to enhance the advantages and compensate for the disadvantages of each have been studied [1]. In this paper, a single wavelength range is called a single band, and the fusion of multiple bands is multi-band. This study proposes a multi-band image processing method using visible light and thermal images.

Meanwhile, the spread of novel coronavirus (COVID-19) infection from 2019 onward has led to a significant increase in interest in infectious disease countermeasures. Contact infection, one of the infections ways, is caused when a non-infected person touches a residual virus on the touched area by an infected person, and the virus enters the body. COVID-19 is reported to survive the longest on plastic object surfaces for about 72 hours, followed by stainless steel surfaces for about 48 hours [2]. In addition to COVID-19, Severe Acute Respiratory Syndrome (SARS) in 2002 and Middle East Respiratory Syndrome (MERS) in 2012 survived on plastic surfaces for up to 9 days and 2 days, respectively [3]. As seen from the outbreak years of SARS and MERS, many infectious diseases existed before the COVID-19 outbreak, and outbreaks are repeated every 5 to 10 years. To avoid contact infection, it is necessary to develop a system that can accurately recognize when and where a person has touched, warn by visualizing the touched areas, and promptly tell us to clean up the areas.

Some methods for touch detection using a thermal camera are proposed [4–6]. When a person touches the surface of an object, the heat signature transferred from the body temperature remains for a certain period, called a heat trace. Figure 1 shows the heat trace in the thermal image and the NIR image at the same time. In the NIR image shown



(a) NIR image (b) Thermal image and a heat trace

Figure 1: Image pairs after touching

in Figure 1a, you can see a human hand reaching a box, but in the thermal image shown in Figure 1b, not only the hand but also heat traces where the person touched the box remain. In this study, we are trying to detect touch with the surface of an object by detecting heat traces.

As shown in Figure 1, a person can appear in both thermal and RGB/NIR images, but heat traces appear only in thermal images, so to distinguish between a person and heat traces, image processing is performed in combination with RGB/NIR images rather than using thermal images alone. Therefore, to distinguish heat traces from persons, the recognition accuracy of heat traces can be improved by combining thermal images with RGB or NIR images rather than using only thermal images. In addition, since heat traces appear only after a person touches an object and shrink over time due to heat dissipation, temporal features are considered effective for heat trace recognition. Here, temporal features are obtained by extracting features in the time direction, such as the shrinking of heat traces, changes in background temperature, and the entry and exit of people into and out of the imaging area. Therefore, it is expected that the recognition accuracy of heat traces can be improved by simultaneously processing multiple images in a time-series sequence. Furthermore, spatial features such as the shape of heat traces and their positional relationship with people, which are obtained by feature extraction from a single image, can also be considered to have an impact on the identification of heat traces and are called spatial features.

For this purpose, to make the comparison as easy as possible, we design a dataset and 2DCNN and 3DCNN models that can be verified and investigate the influence of the combination of bands of images used as input, temporal features (frame size, frame rate, etc.), and spatial features (size of the spatial domain), and investigate the impact from each feature on the identification system of heat traces.

In Chapter 2, we introduce the application examples of multi-band image processing, contact detection, and related works on 3DCNN to understand the background of our proposed method; in Chapter 3, we describe the design of our proposed method and the deep learning model we use; and in Chapter 4, we give an overview of our experiments from dataset collection to learning and inference. The results obtained in Chapter 5 are evaluated in terms of band combinations, spatial features, and temporal features. Using the obtained results, we discuss the issue and future work in Chapter 6 and the conclusion in Chapter 7.

2 Related Work

In this section, we introduce related work to our study, including applications of multi-band image processing in various tasks, studies in the fields of human interface, and 3DCNN models using spatio-temporal information for action recognition.

2.1 Multi-Bandwidth Image Processing

Currently, image sensors that capture various wavelength ranges, such as RGB, near-infrared (NIR), and far-infrared (FIR, thermal) cameras, are widely used. Light refers to electromagnetic waves in a specific wavelength range, especially visible light in 380–780 nm, NIR light in 700–2,500 nm, and thermal light in 7.5–13.5 µm. Figure 2 shows the spectral transmittance in the atmosphere. In the atmosphere, light is diffused by collisions with water and carbon dioxide particles, which attenuates light in some wavelength regions. On the other hand, visible, NIR, and thermal wavelengths are relatively susceptible to attenuation, so they are employed as sensors for capturing images. In particular, infrared radiation is emitted from objects with an absolute temperature of more than 0 K, and its emission increases as the temperature rise, making it possible to calculate the temperature from the amount of infrared radiation. In this way, thermal cameras measure an object's temperature in a non-contact manner. Because certain information can be extracted from each of these different wavelength ranges, multi-band image processing combining various sensors is expected to improve the accuracy of image processing techniques that have been performed with a single band and to extract information not captured with a single band.

This section introduces several studies on the applications of multi-band image processing.

Speth et al. proposed the human detection method using aerial images for monitoring and patrolling by drones in disaster relief operations [7]. People captured by drones are so small that it is difficult to find all of them from RGB images. On the other hand, the person's edges are indistinct in thermal images, but its presence is easy to detect because of the high-temperature enhancement. A combination of RGB and thermal images was used to take advantage of these features to improve the human detection accuracy of people with a temperature above a certain level. Compared to the input of single band images

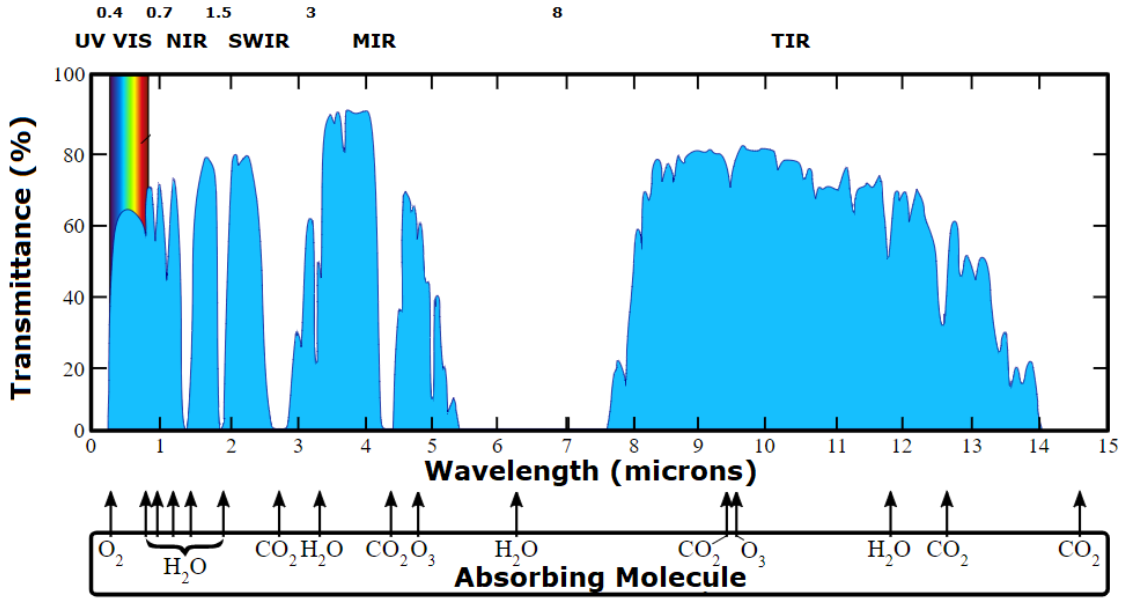


Figure 2: Spectral atmospheric transmittance[1]

to the object detection model YOLOv3 [8] by rectangles or combining their outputs, the highest score was reported when using fused images in both bands before input to the network.

Huo et al. proposed a glass segmentation method by combining multiple band images [9]. The glass transmits visible light, making it difficult to detect with conventional visible light image-based detection methods. In contrast, the glass reflects most of the FIR rays. Therefore, by simultaneously convolving RGB and thermal images and extracting these features, they have achieved robust detection of transparent glass without affecting the background.

Furthermore, a crack detection method in outdoor infrastructure, such as the ground, has been proposed [10]. RGB cameras can depict detailed spatial information under ideal lighting conditions, but their performance is known to deteriorate in low-light environments. Thermal cameras, on the other hand, have relatively low resolution but are robust to changes in lighting conditions. Since it is difficult to have ideal lighting conditions outdoors, the authors report that a combination of RGB and thermal images can detect

[1]Cited from "Atmospheric Transmittance," `https://www.usna.edu/Users/oceano/pguth/md_help/remote_sensing_course/atmos_transmit.htm`, accessed on January 26th, 2023.

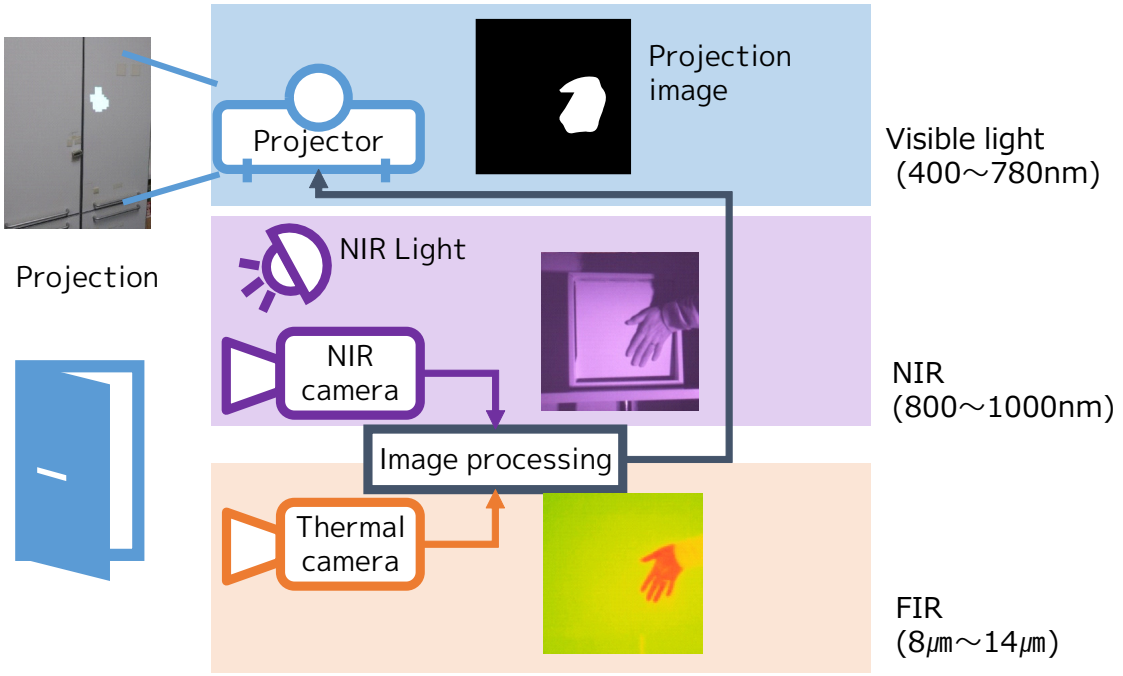


Figure 3: Real-time heat trace visualization system

cracks and distinguish damage types with high accuracy, even under unstable outdoor lighting conditions.

An image fusion framework based on multi-band image processing has also been proposed [11]. Importance maps are computed from images in each band based on a learning-based approach. By reconstructing the image using the created importance maps, the features of each band image are fused into a single image. Experimental results show that this method is effective for various applications such as depth enhancement and haze removal.

This research also aims to combine features of different bands to achieve highly accurate image processing and detect thermal traces that appear in thermal images but not in RGB images.

2.2 Touch Detection and related Studies with Heat Traces in the Field of Human Interface

Our research group has developed a system for real-time detection and visualization of heat traces using NIR and thermal images [4]. Figure 3 shows the system overview. When

a person touches an object, a person and heat traces appear in the thermal image, but only a person appears in the NIR images. Therefore, when the previously acquired background image is subtracted from the images, the thermal image emphasizes the area of the person and heat traces. In contrast, the NIR image emphasizes the area of the person. In this way, among the areas highlighted in the thermal image, areas that overlap to a certain extent with those highlighted in the NIR image can be considered human, and others can be heat traces. The thermal traces detected in this way are projected onto the target object using a projector to visualize and warn of contact in real-time, thereby avoiding contact infection and helping to prevent the spread of COVID-19 disease. The system has been extended to detect disinfected areas as well [5]. When disinfection is performed, the disinfectant gradually evaporates on the surface of the disinfected object. The object's temperature decreases over ten to several tens of seconds due to the heat of vaporization. This change in the thermal image during disinfection is called a cold trace. In the same way as with thermal traces, the cold trace area can be identified by performing the same image processing on the temperature decrease. This can facilitate more efficient cleaning activities by recording the areas that have been disinfected.

As another example of research in touch detection using heat traces, a method for detecting heat traces appearing in thermal video using U-Net [12] has been proposed [6]. U-Net is an Encoder-Decoder model that identifies object types and backgrounds by labeling each pixel in the input image. Each frame of the thermal video is input to U-Net and segmented into heat traces, people, and background. Only the regions labeled as heat trace are summed in the temporal direction to detect regions in the thermal video where touch has occurred. However, since the model uses frame images as input and assumes that a thermal trace is successfully detected if it is detected at least once during the video, it does not take into account the disappearance or temporal features of heat traces.

In the field of human interface, it has been known that heat traces are generated by touching objects, and many studies utilize this phenomenon. For example, by analyzing the characteristics of heat traces, the authors are trying to develop a wall display that can be operated by touching and projecting the control screen directly on the wall [13, 14]. From the security point of view, Abdelrahman et al. have suggested that Personal Identification Number (PIN) codes and pattern locks, which are widely used as PINs for

smartphones, can be identified from thermal traces by using a thermal camera to monitor a smartphone after the PIN has been entered. They present an example of designing more secure passwords that is difficult to identify even if the phone is monitored with a thermal camera [15].

In this research, we target the identification of heat traces from the appearance to the disappearance and confirm whether CNNs can identify heat traces and what features are effective in identifying heat traces with high accuracy.

2.3 3DCNN

Recently, 3Dimentional CNN (3DCNN), a temporal extension of CNN for image recognition, has been attracting attention in action recognition. Conventional CNNs used for image recognition cannot learn temporal features because they learn images one by one as if they were completely different images, even if they are consecutive frames. Convolution in the temporal dimension effectively improves the accuracy of tasks like action recognition using videos. There are roughly two types of 3DCNN models: those that use a two-dimensional velocity vector field (optical flow) generated by calculating pixel-by-pixel movement vectors from neighboring video frames and those that convolve video inputs in the spatio-temporal direction simultaneously. In this section, we introduce several models that have contributed to the development of 3DCNNs. However, in this study, CNNs that use single frames as input will be referred to as 2DCNNs.

In 2012, Ji et al. are the first to propose a 3DCNN model with multiple consecutive images in the temporal direction as input [16]. The proposed model treats each frame as a channel, like the RGB channel of a visible light image. The input data to the model is generated from a grayscale image and 2-dimensional optical flows. The authors attempt to extract temporal features for each generated data by convolving consecutive 3 scenes with the same kernel.

As a 3DCNN using optical flow, Simonyan et al. proposed in 2014 a Two-Stream ConvNet that exploits spatio-temporal information by training different CNNs with the spatial and temporal information of input videos and integrating their output [17]. First, one frame is selected, and horizontal and vertical optical flows between consecutive frames are generated from the input video. By integrating these outputs, action recognition is

performed by understanding the position and motion of objects in the video. However, since spatial stream ConvNet takes each frame of the optical flow as a channel, the networks that make up Two-Stream ConvNet are all 2DCNNs.

In 2015, Tran et al. proposed C3D, a 3DCNN model that simultaneously convolves in spatio-temporal directions by extending all the layers of 2DCNN to three dimensions [18]. Like a simple 2DCNN, C3D performs feature extraction using eight 3D convolutional layers and 3D pooling layers. However, since the number of frames in the temporal direction is generally smaller than the spatial size of the input video, pooling in the temporal direction is often not applicable. Because of the large number of parameters required for 3D convolution, the authors concluded that the network could not be very deep, and a larger video dataset is necessary for network training.

To solve the problem of C3D, Carreira et al. proposed I3D [19], a 3DCNN model with a much-reduced number of parameters compared to C3D by embedding the Inception module used in GoogLeNet [20]. In addition, they have published a large dataset Kinetics for action recognition tasks. They reported that Two-Stream I3D, which trains videos and their optical flows in parallel to different I3Ds and integrates their outputs, is the most accurate compared to conventional 3DCNN methods. Furthermore, Kinetics enabled more complex training of 3DCNN models, which is a significant contribution to the field of action recognition.

Feichtenhofer et al. proposed SlowFast Networks, a 3DCNN model focusing on the mechanism of the human eye [21]. Two neural pathways transmit signals from the human retina to the brain: the parvocellular pathway (parvo) and the magnocellular pathway (magno) [22]. The former has high spatial resolution but low temporal resolution, while the latter has low spatial resolution but high temporal resolution. Parvo is generally involved in morphological processing, while magno is involved in motion processing.

The network structure of SlowFast Networks is based on these neural systems and consists of two networks: The slow pathway that captures semantic information and the Fast pathway that captures motion. In the slow pathway, the input video's frames are reduced, and the channel size is increased to extract semantic features. On the other hand, the fast pathway uses more frames as input than the slow pathway to capture more motion instead of reducing the channel size. The model achieved State-of-the-Art (SOTA)

on an action classification task using the large video dataset Kinetics-400 in January 2019, making it the current de facto standard for 3DCNN models.

Heat trace, the target of this study, appears through a series of actions: a person appearing in the angle of view, touching an object, and then moving out of the angle of the thermal camera. In addition, since heat traces diffuse or shrink after emerging until disappearance, temporal information may be effective for identification of them. In this study, we design a 3DCNN model based on C3D that simultaneously convolves the temporal and spatial directions and investigate how the temporal features affect heat trace identification.

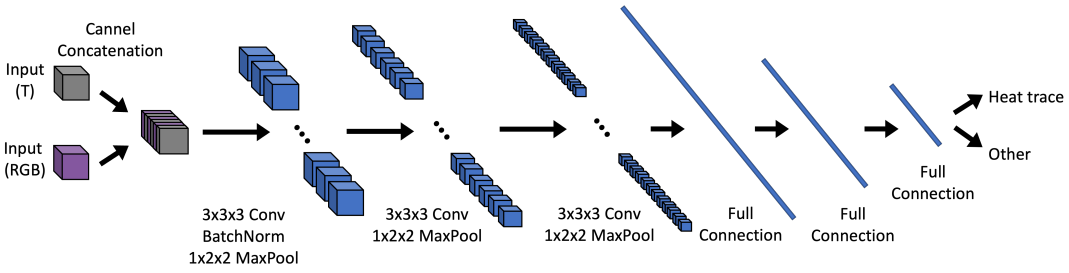


Figure 4: 3DCNN architecture using RGB and thermal videos

3 Proposed Method

This section describes our proposed 3DCNN model for heat trace detection using two band images, thermal and RGB images. Figure 4 shows the network structure when RGB video and thermal video are input simultaneously. Although several methods exist to combine thermal and visible light images as input, we first chose to combine them by channel concatenation. When RGB images are used as input to the CNN, each color R, G, and B are considered 3 channels, each color having a different feature. Considering thermal images as T, in the same way, the T channel can be used as the 4th channel for the input to the network. On the network side, data with any channel size can be input by adjusting the channel size of the input layer.

Since videos are input to each CNN model, designing a model that performs feature extraction in the spatial and temporal directions is necessary. To compare which spatial or temporal features are more effective in detecting heat traces, this study builds a model based on C3D, which does not distinguish between the temporal and spatial dimensions and performs convolution processing simultaneously. Taking a series of f frames as input, convolution by $3\times3\times3$ kernels in the spatio-temporal directions and the $1\times2\times2$ maximum pooling are repeated 3 times. The features obtained in this way are processed in all the 3 full connection layers, and finally, the output layer estimates whether the video has heat traces. The model is trained using batches of batch size bs.

We constructed a 2DCNN with the same structure as this model to confirm effectiveness using the temporal features. Since the 2DCNN does not require convolution and pooling in the temporal dimension, we replaced the convolution and pooling layers for 2D networks. We set the frame size to $f = 1$ so images can be input instead of videos.

The goal of my research group is to develop a contact detection system using heat traces. Therefore, we first identify issues in heat trace identification by conducting experiments under simple conditions, then explore optimal network structures and data input methods. We then apply this knowledge to data collected against various backgrounds to extend the system to real-world sensing environments. To accomplish this, we divided the experiment into the following phases.

Step 1: Investigate whether heat traces can be detected under the simplest conditions and what features are used in the detection

Step 2: Confirm whether the findings obtained in Step 1 are effective for sensing in real environments

Step 3: Extend from heat trace identification to detection and develop a real-time contact detection system

In this study, we focus on Step 1. First, we took videos under the simplest conditions, and training and test datasets were created through several preprocessing steps. We evaluate heat trace identification with some scores by training and inference with each model and discuss the results regarding bands used, temporal features, and spatial features. Furthermore, we also experimented with some channel size patterns to optimize the networks.

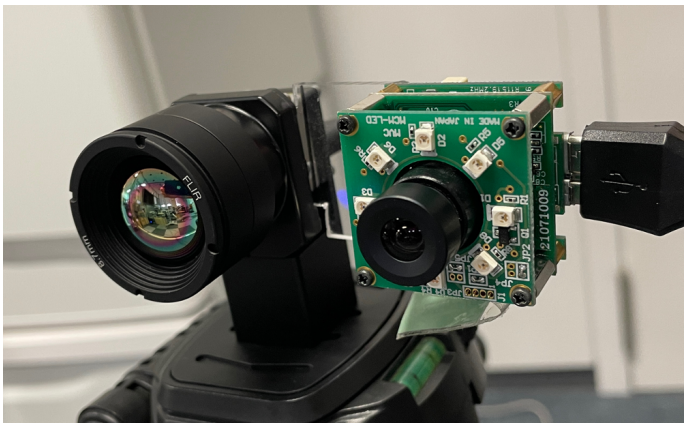


Figure 5: The thermal (left) and RGB (right) cameras

4 Experimental Outlines

This section describes the experimental steps, from capturing the video to detecting heat traces. This study requires RGB and thermal videos captured from the same angle, but now there are no open datasets supporting multi-band image processing. In this study, We generated training and test datasets by capturing videos of people touching a target and heat traces appearing at the touched surface, and we annotated and preprocessed the videos.

4.1 Data Collection

As shown in Figure 5, using the thermal and RGB cameras on the heat trace visualization system [4], we have taken videos to create datasets [4]. As a first step in the experiment, the ability of the CNN to discriminate heat traces was confirmed by comparing only the bands and spatial and temporal features. However, it is difficult to make rigorous comparisons because of noises from backgrounds if videos are shot against various backgrounds. Therefore, we chose the background of a white wall for the touch target.

The specifications of the cameras are shown in Figure 1. Since these cameras have different frame rates, the frame rate of the thermal camera was standardized by decimating some frames of thermal videos to match that of the visible light camera, which has a lower frame rate. Figure 6 shows examples of RGB and thermal images in our dataset.

We have captured 1 and 3 contact in a 1 movie. A series of actions generate heat

Camera	Band	Image size [pixels]	Frame rate [fps]
FLIR Boson 640	NWIR	512×640	60
Gazo MCM-320	Visible	480×640	30

Table 1: RGB and thermal cameras' specifications

(a) RGB image (b) Thermal image

Figure 6: Example images in our dataset

traces: a person entering the camera's angle of view, touching the target, and moving out of the imaging area. On the other hand, a person may behave in a non-touching manner, such as crossing between the cameras and the target or stopping in front of the target for several seconds. Therefore, assuming that people behave in various ways in a real environment, we also captured videos with a mixture of touching and fake information, such as a person passing in front of the camera or staying in front of the camera without touching it. Finally, we took 14 videos of about 10 minutes each, one of which was used for the inference dataset and the others for the training dataset.

4.2 Annotation

We manually annotated heat traces by enclosing each of them in a bounding box, as shown in Figure 7. It is difficult to define disappearance visually, so we defined disappearance as the time when the temperature difference between the maximum and the minimum values

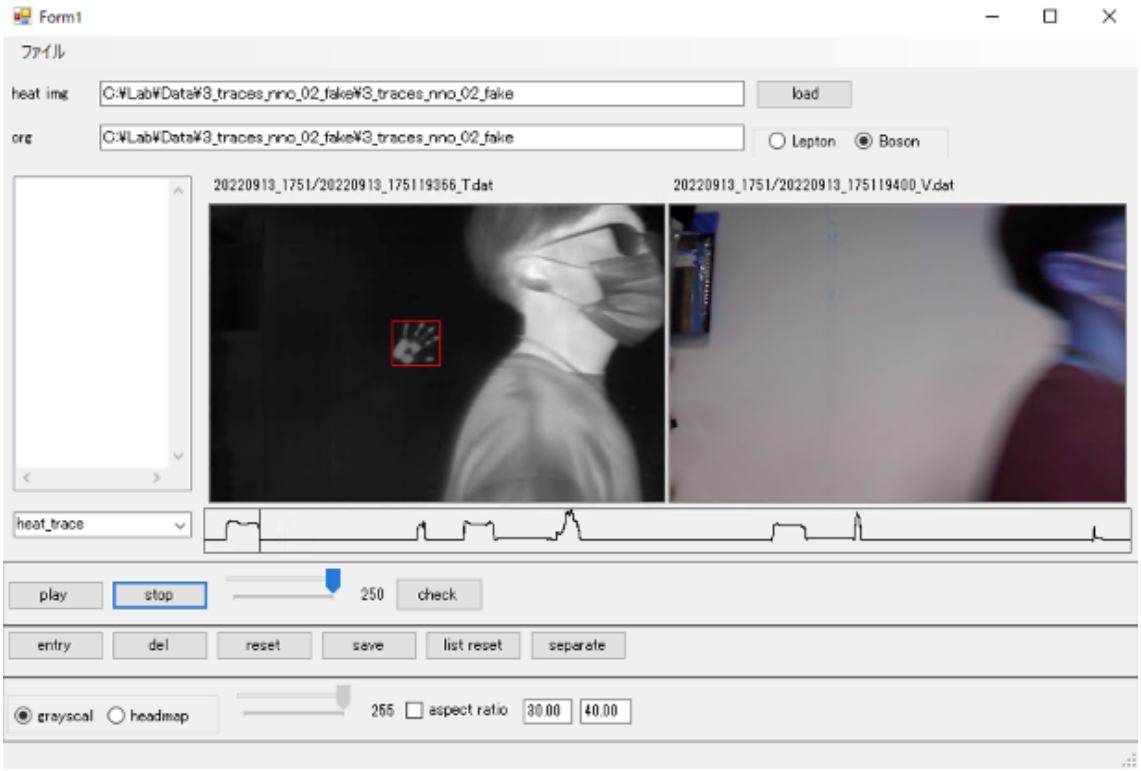


Figure 7: A screenshot of the annotation tool

in the bounding box is less than 0.1 degree Celsius. Because the bounding box contains both heat traces and background, the maximum value can be approximated as the temperature of the heat traces and the minimum value as the temperature of the background. The temperature change of heat traces over time and the time of disappearance are shown in Figure 8. Each maximum pixel intensity (maxpi) and minimum pixel intensity (minpi) refers to the maximum and minimum temperature, respectively, and mean pixel intensity (meanpi) is the average pixel intensity of the image.

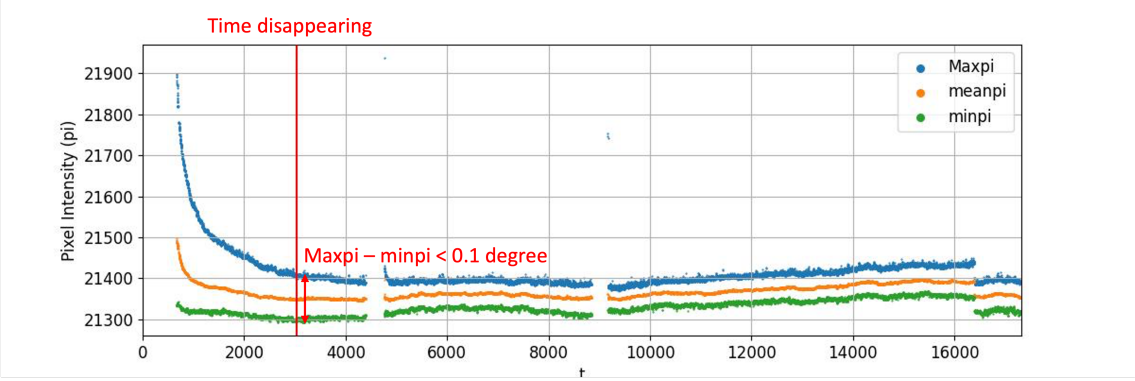


Figure 8: Attenuation performance and disappearance threshold of heat traces

4.3 Preprocessing

This section describes the preprocessing steps involved in creating a dataset.

4.3.1 Calibration

Two types of cameras are used in this study, and the difference in the angle of view and the installation position of the cameras causes the misalignment. In Figure 6, 6a has a narrower angle of view and a smaller person area than 6b. Therefore, before creating the dataset, we used perspective transformation to align the images as a preprocessing step. Affine transformation is a method of aligning image A with image B by setting three or more corresponding points on A and B, calculating a perspective matrix representing linear transformations (scaling, shearing, rotation) and translations, and multiplying it by A. The image obtained by perspective transformation is shown in Figure 9. In this study, the RGB images were transformed to match the thermal images. The regions of the person are well synchronized, which reduces noise due to misalignment that may occur when training the CNN.

4.3.2 Trimming and Resizing in Spatial Dimension

To confirm the influence of spatial features on heat trace identification, in this study, we cropped images with the 60×60 (narrow) and the 200×200 (wide) squares to include the heat traces within the region. The image cropping can exclude regions with no values generated by the affine transformation. The images obtained by cropping with the narrow

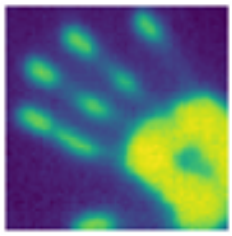
Figure 9: RGB and thermal images aligned by affine transformation

and wide box are shown in Figure 10.

4.3.3 Data Division in Temporal Dimension

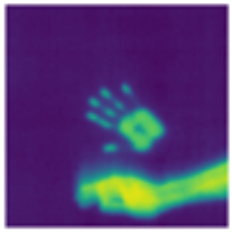
Several time-related parameters must be specified to create a video to be input to 3DCNN. In this study, captured videos are divided using the frame size f, the slide size s, and the frame interval i as parameters. For example, consider the case where $f = 3$, $s = 5$, and $i = 2$. Figure 11 shows how the video is divided according to these parameters. The frame rate of collected videos is $30fps$ because the frame rate of the thermal camera is aligned according to the RGB camera, but since the image is resampled by $i = 2$, that is, every 3 frames, the real frame rate of 1 video generated by the division is $15fps$. Finally, the video dataset is created with frame size $f = 3$, 0.2 seconds in duration, and slide size $s = 5$, 0.33 seconds. Finally, a video dataset is generated by splitting the image with frame size $f = 3$ and slide size $s = 5$, i.e., with 0.2 seconds width being shifted every 0.33 seconds.

(a) RGB image (narrow)



(b) Thermal image (narrow)

(c) RGB image (wide)



(d) Thermal image (wide)

Figure 10: Examples of RGB and thermal images in dataset

4.3.4 Training and Inference

The datasets used have extremely different numbers of data per label. Therefore, when training the network, we implemented a batch sampler that randomly selects the same number of videos of heat trace labels as those of other labels so that the training is less affected by the difference in the number of data per label. An example of the contents of the image and video datasets used for training and inference is shown in Table 2. In this study, we use NVIDIA RTX 3090 as the GPU for training and inference with the datasets.

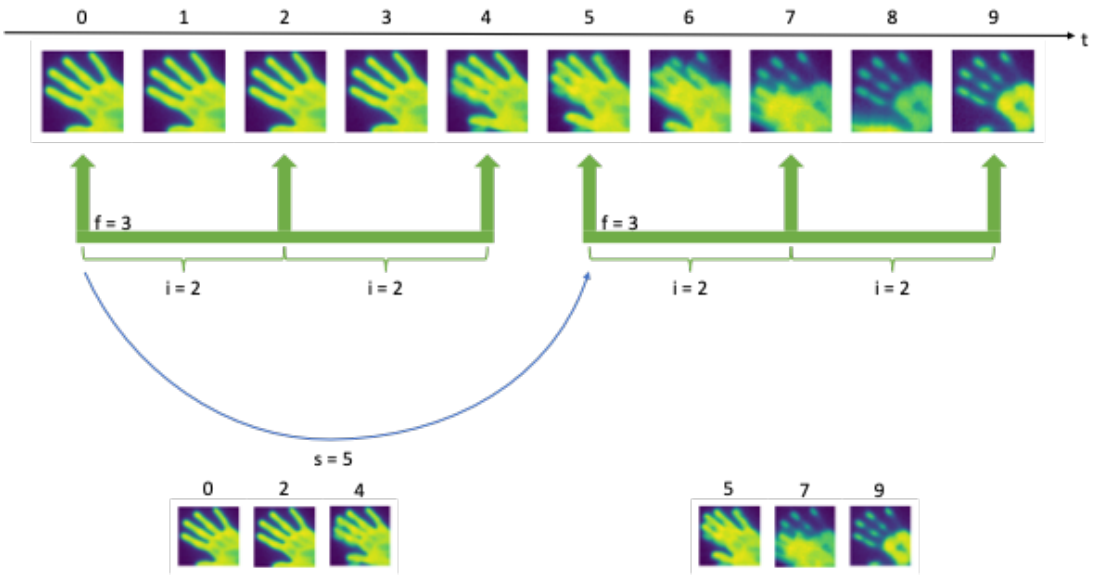


Figure 11: Example of video division

(in case of $f = 3, s = 5, i = 2$)

				Train		Test	
CNN	f	bs	$size$	T	F	T	F
2D	1	256	narrow	22733	10371	1542	1438
			wide	19331	13173	1542	1438
3D	10	256	narrow	22630	10424	1529	1448
			wide	19822	13232	1529	1448

T: heat trace, F: others

Table 2: Dataset overview

5 Evaluation

In this section, we evaluate the ability to identify heat traces using four scores: accuracy, precision, recall rate, and f1 score, calculated from the confusion matrix obtained from the experiments in section 4, to compare three perspectives: the band types, time length, and space size. In this study, we will mainly discuss the accuracy and f1 score to see if we can detect the emergence and disappearance of heat traces. Here, we call image names as below:

- T: using only thermal images
- V: using only RGB images
- TV: using fused images with thermal and RGB images

In addition, we call frame size $f = 1$ as "2D", $f = 10$ as "3D", and image size 60×60 as "narrow," 200×200 as "wide." We set the channel sizes 32, 64, and 128 for 2DCNN and 64, 128, and 256 for 3DCNN in order from the lower convolutional layer.

5.1 Difference in Band types

Figure 12 shows the scores obtained for fixed spatial and frame sizes, allowing comparison by bands of input data with the model in section 3 and collected data in section 4. First, accuracy was as low as 50 % in all experiments using V. The recall rate was very low for narrow images because most of the inputs were predicted not to be heat traces, while for wide images, the recall rate was very high because heat traces were predicted to exist. These show that the RGB images do not identify heat traces well because they do not contain any information about the heat traces.

Second, comparing T and TV, T is much more accurate than TV when using narrow images, whereas TV is more accurate than T when using wide images. In other words, V was ineffective in identifying heat traces when the location of their appearance was known in advance. In contrast, if the location of the appearance was not known in advance, V could be used to recognize the imaging environment and help detect heat traces.

Figure 13 and Figure 14 show the inference results using T and TV for the narrow-3D dataset for validation. For the vertical axis, 1 means that the correct label has heat traces,

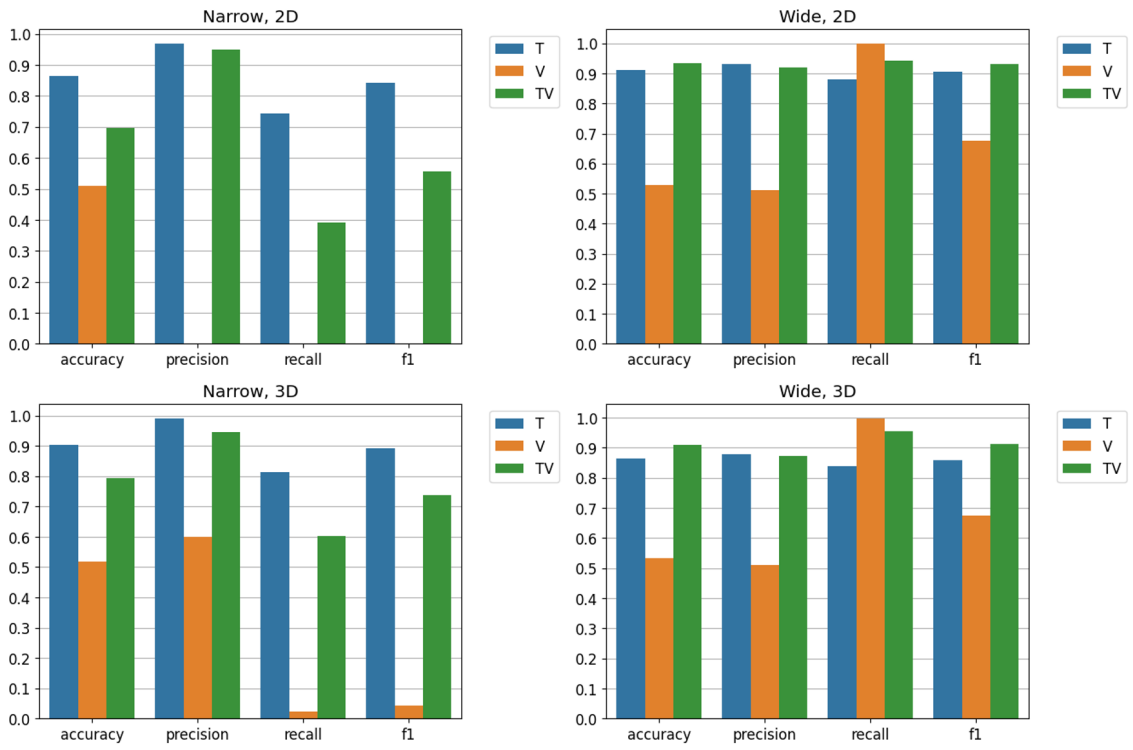


Figure 12: Scores in difference from used band

0 means others, and the horizontal axis indicates the index in chronological order. For each input, a blue dot indicates a correct prediction and a red one a failure. In both cases, the inferences for the data before and after the appearance of heat traces are generally correct, and this trend is also true for the results using other inputs. On the other hand, most errors occur before and after the disappearance of the heat traces. Because heat traces diffuse or shrink continuously over time, there is no precise timing of disappearance, making it difficult to recognize the time when heat traces disappear according to the label attached.

5.2 Difference in Temporal Features

Figure 15, recapitulating Figure 12 with frame sizes, shows the scores obtained by fixing the spatial size and band types (only T and TV) so that they can be compared concerning the frame size. From now on, we will compare the scores for T and TV, excluding V. The score for 3D was higher than that for 2D when narrow images were used, while it was lower than the score for 2D when wide images were used. The heat traces appeared very

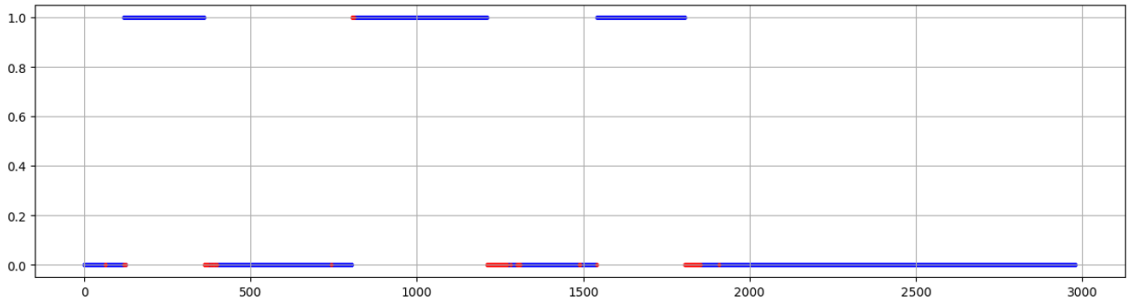


Figure 13: Inference result using T

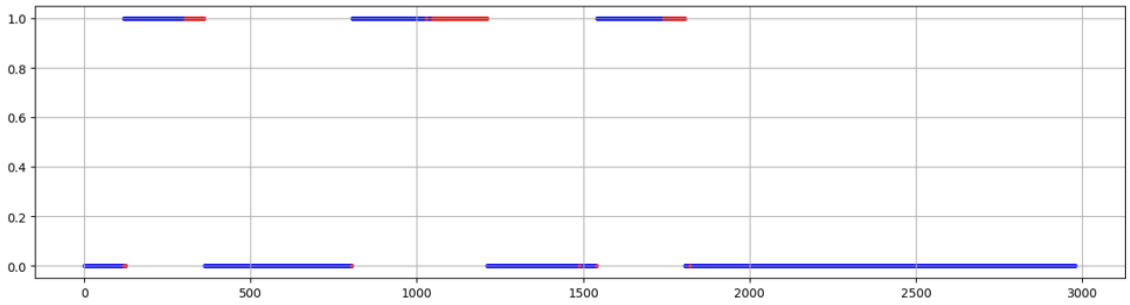


Figure 14: Inference result using TV

small in the wide images, which may have made it difficult to detect heat traces based on time information. On the other hand, if the locations where heat traces appeared were known in advance, the time information would be effective in identifying heat traces.

5.3 Difference in Spatial Features

In Figure 16, recapitulating Figure 12 with frame sizes, each score obtained by fixing the number of frames and the band types (only T and TV) is displayed so that it can be compared according to the spatial sizes. The wide images' scores were higher than those of narrow images although those of 3D-T are almost equal. In particular, the difference in scores between using narrow and wide images was large for TV, indicating that the value of spatial information became higher by combining V. Although the low resolution of thermal images makes it difficult to recognize the background, the supplementary use of high-resolution RGB images helped CNN understand the background.

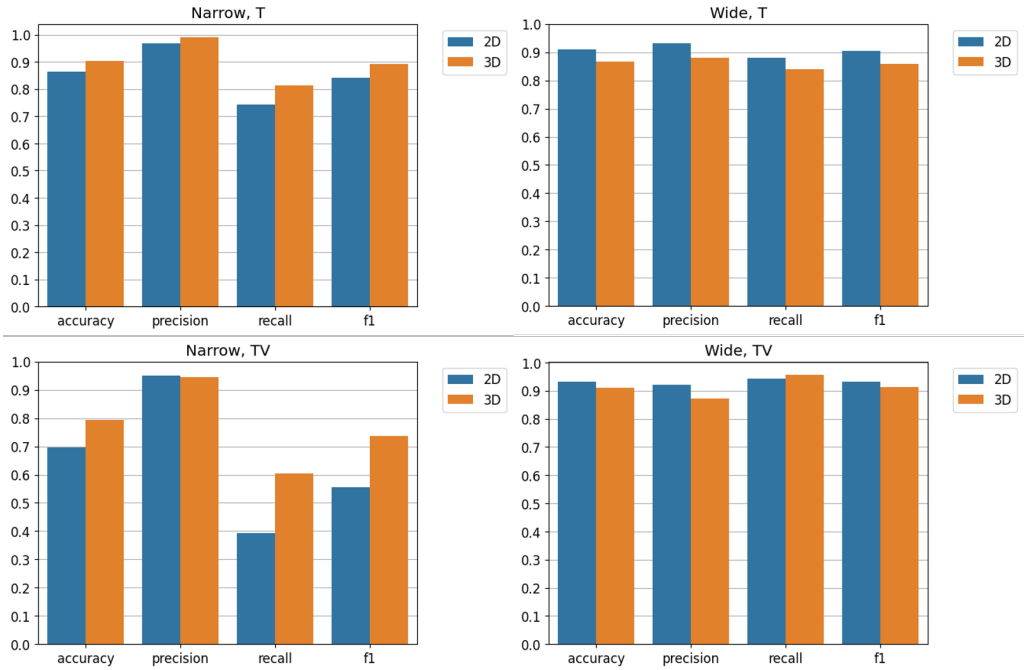


Figure 15: Scores in difference from frame sizes

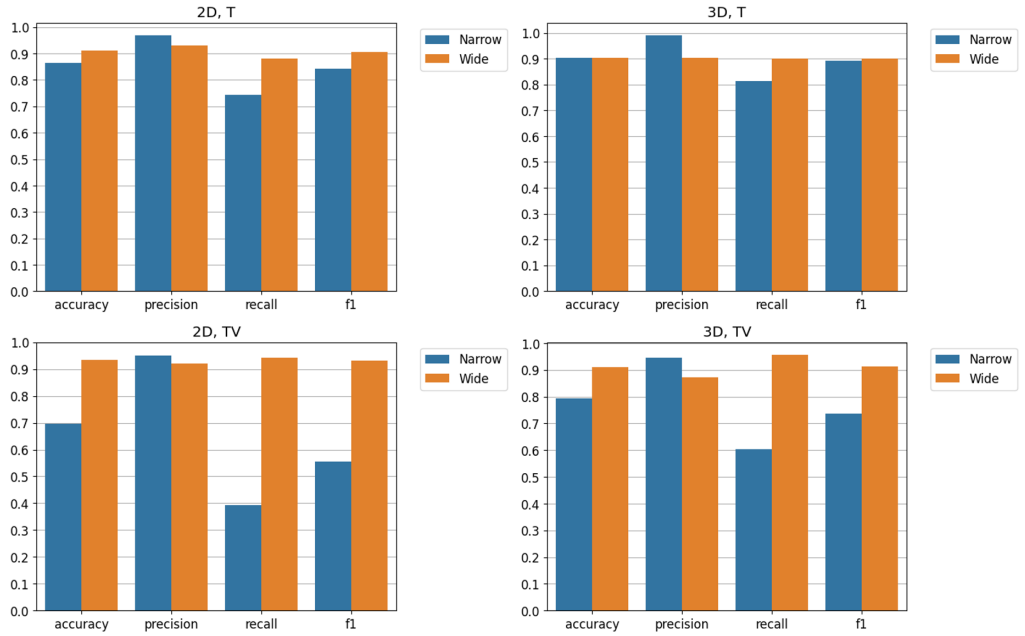


Figure 16: Scores in difference from spatial sizes

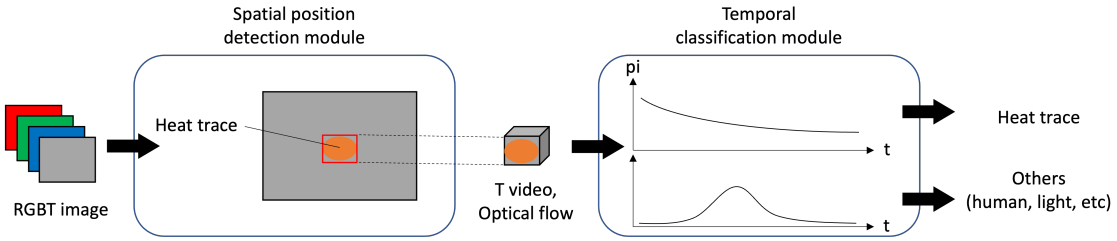


Figure 17: New network processing temporal and spatial information separately

6 Discussion

The results in section 5 indicate that information from RGB images is effective when a large area is captured because it is not known where heat traces will appear in advance. In contrast, using temporal information improves detection accuracy when the locations where heat traces appear are restricted.

Our study uses a model that simultaneously processes all channel, spatial, and temporal information. However, we expect that a model that makes more efficient use of each type of information can be constructed by dividing the processing into two stages as described below:

Stage 1: Identify the appearance area of heat traces by 2DCNN using the thermal and RGB images capturing a large space, including the object or the background to be touched as input (detection task).

Stage 2: Classify whether it is a heat trace or not by 3DCNN using the thermal video generated by trimming output images in the predicted area and combining them in fixed frames from Stage 1 (identification task).

Figure 17 shows the newly proposed network architecture constructed by these two stages. We would like to confirm whether it is possible to simultaneously detect the location and time of appearance and disappearance of heat traces with high accuracy by using a network that processes temporal and spatial information separately. In the second stage, the optical flow of the input T videos can be also effective for temporal feature extraction because it means attenuation characteristics of the heat trace, human motion, and so on. In addition, because we used a very designed dataset with a white wall as the

touch target, the identification accuracy was higher even when only T was used. If the spatial information becomes more complex by setting various objects as touched targets, the effect of combining V is likely to be more significant. Therefore, we are planning to investigate whether this study's findings effectively detect heat traces by using a more realistic dataset with various backgrounds and objects.

7 Conclusion

This paper proposed a method for detecting heat traces appearing on touched points by fusing thermal and RGB videos. We compared the results using several evaluation metrics calculated from the confusion matrix of the test dataset regarding used band types, temporal information, and spatial size. The results show that when the spatial size is strongly limited, accuracy is improved by using temporal information. On the other hand, when input images can be imaged in a wide area, the combination of both types of images results in higher detection accuracy. As a result example, the accuracy was improved from 91.0 % when using only the thermal images to 93.3 % when both were used. This suggests that heat traces may be detected more accurately by processing the band combination, temporal information, and spatial information separately rather than simultaneously.

Considering this, we discussed a new model composed of two modules. The first module detects the locations where heat traces appear using spatial information with thermal and RGB images that capture a large space. The second module uses temporal information to classify whether the input is a heat trace with thermal videos generated by collecting a fixed number of thermal images trimmed in the specific area in the first module.

The proposed method can be helpful for the heat trace detection system to avoid contact infection from COVID-19 or other viruses. Furthermore, obtaining information about image processing by the fusion of images captured in different wavelength areas also be effective for the development of image recognition by multi-band image processing.

In our future work, we would like to investigate the effectiveness of the proposed method and the newly proposed model for datasets with various backgrounds and objects since the dataset used in this study was designed under simple conditions. In addition, the number of parameters and training time for 3DCNN are very large due to hardware resource limitations. We would like to construct a network that enables faster learning and inference by adopting effective resource-saving techniques proposed so far.

Acknowledgments

First, I would like to express my deepest gratitude to my master's thesis supervisor, Professor Masayuki Murata of the Graduate School of Information Science and Technology, Osaka University, for his valuable comments, insights and continuous encouragement.

I would show my greatest appreciation to Specially Appointed Associate Professor Yasue Kishino of the Department of Information Networking in the Graduate School of Information Science and Technology, Osaka University, also Senior Research Scientist of NTT Communication Science Laboratory, NTT Corporation, for her continuing encouragement, valuable discussions, academic advising, providing many opportunities for growth as a researcher and various supports throughout my studies and the preparation of this manuscript as a chief investigator.

I also would like to express my gratitude to Associate Professor Shin'ichi Arakawa, Associate Professor Yuichi Ohsita, Assistant Professor Daichi Kominami, Assistant Professor Tatsuya Otoshi and Specially Appointed Assistant Professor Masaaki Yamauchi of the Graduate School of Information Science and Technology, Osaka University for giving me helpful comments.

I would also like to show great appreciation to Dr. Shin Mizutani, Research Scientist of NTT for Communication Science Laboratory, NTT Corporation, for constructive discussions, especially when we participated in the workshop in Hakodate, Hokkaido.

I would like to thank all Advanced Network Architecture Research Laboratory members. At the end of my acknowledgment, I thank my parents for supporting my life as a master's student.

References

[1] Y. Yu, "RGB-thermal based denoising methods: a review of deep learning based image denoising algorithm and application," *IEEE Transactions on Multimedia*, vol. 14, no. 8, 2022.

[2] N. van Doremalen, T. Bushmaker, D. H. Morris, M. G. Holbrook, A. Gamble, B. N. Williamson, A. Tamin, J. L. Harcourt, N. J. Thornburg, S. I. Gerber, J. O. Lloyd-Smith, E. de Wit, and V. J. Munster, "Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1," *New England Journal of Medicine*, vol. 382, no. 16, pp. 1564–1567, 2020.

[3] G. Kampf, D. Todt, S. Pfaender, and E. Steinmann, "Persistence of coronaviruses on inanimate surfaces and their inactivation with biocidal agents," *Journal of Hospital Infection*, vol. 104, no. 3, pp. 246–251, 2020.

[4] Y. Shirai, Y. Kishino, Y. Yanagisawa, K. Ohara, S. Mizutani, and T. Suyama, "Alertable Surfaces: an actual environment that alert virus attachment (in Japanese)," in *Proceedings of Workshop on Interactive System and Software (WISS 2020)*, 2020.

[5] Y. Kishino, Y. Shirai, Y. Yanagisawa, H. Sugawara, K. Ohara and S. Mizutani, "Alertable Surfaces: an actual environment that alert virus attachment by recognizing human touch and sanitization (in Japanese)," in *Proceedings of Workshop on Interactive System and Software (WISS 2022)*, 2022.

[6] G. Ma, W. Ross, M. Tucker, P. Hsu, D. M. Buckland, and P. J. Codd, "Touch-point detection using thermal video with applications to prevent indirect virus spread," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1–11, 2021.

[7] S. Speth, A. Gonçalves, B. Rigault, S. Suzuki, M. Bouazizi, Y. Matsuo, and H. Prendinger, "Deep learning with RGB and thermal images onboard a drone for monitoring operations," *Journal of Field Robotics*, vol. 39, pp. 840–868, 2022.

[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv, 1804.02767, 2018.

[9] D. Huo, J. Wang, Y. Qian, and Y. Yang, "Glass segmentation with RGB-thermal image pairs," arXiv, 2204.05453, 2022.

[10] Q. G. Alexander, V. Hoskere, Y. Narazaki, A. Maxwell, and B. F. Spencer, "Fusion of thermal and rgb images for automated deep learning based crack detection in civil infrastructure," *AI in Civil Engineering*, vol. 1, no. 1, p. 3, 2022.

[11] T. Shibata, M. Tanaka, and M. Okutomi, "Unified image fusion framework with learning-based application-adaptive importance measure," *IEEE Transactions on Computational Imaging*, vol. 5, no. 1, pp. 82–96, 2019.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[13] E. Larson, G. Cohn, S. Gupta, X. Ren, B. Harrison, D. Fox, and S. Patel, "Heatwave: Thermal imaging for surface user interaction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, 2011, pp. 2565–2574.

[14] A. Sahami Shirazi, Y. Abdelrahman, N. Henze, S. Schneegass, M. Khalilbeigi, and A. Schmidt, "Exploiting thermal reflection for interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, 2014, p. 3483–3492.

[15] Y. Abdelrahman, M. Khamis, S. Schneegass, and F. Alt, "Stay cool! understanding thermal attacks on mobile-based user authentication," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, 2017, p. 3751–3763.

[16] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[17] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems (NIPS)*, vol. 27, pp. 568–576, 2014.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015, pp. 4489–4497.

[19] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the Kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4724–4773.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[21] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019, pp. 6202–6211.

[22] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth: Anatomy, physiology, and perception," *Science*, vol. 240, no. 4853, pp. 740–749, 1988.