IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON HUMAN-CENTERED SMART SYSTEMS AND TECHNOLOGIES

Received December 22, 2017, accepted February 18, 2018, date of publication March 2, 2018, date of current version March 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2810951

AHD: Thermal Image-Based Adaptive Hand Detection for Enhanced Tracking System

EUNGYEOL SONG, HYEONGMIN LEE, JAESUNG CHOI, AND SANGYOUN LEE, (Member, IEEE)
Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea
Corresponding author: Sangyoun Lee (syleee@yonsei.ac.kr)

ABSTRACT Thermal sensors are robust to conditions that are constraints for visual sensors, such as illumination changes. Based on this effectiveness, studies on hand tracking have enabled the demonstration of higher performance with low computation times. In this paper, we propose a novel hand detection method and tracking framework based on information extracted from thermal images. An adaptive hand detection (AHD) method was designed with five models that use temperature analysis to obtain the region of interest of the hand. To improve performance, we introduced an AHD-based automatic tracking-by-detection algorithm using the Kernelized correlation filters tracker. Finally, we combined the hand detection and tracking algorithms into a single framework, a guidance framework for tracking by detection (GFTD). To verify the performance, we evaluated the accuracy of the hand detection using success rate (Intersection over Union) and the trajectory using object tracking error (OTE). The proposed GFTD improves the performance in terms of success rate and OTE by 15% and 16.3%, respectively, compared with the conventional methods.

INDEX TERMS Human computer interaction, detection algorithms, hand detection, tracking loops.

I. INTRODUCTION

A Thermal sensors in the infrared subdivision are classified as long-wavelength infrared (LWIR), which has a wavelength range of 8-12μm. Radiant heat from the human body is sensed through LWIR sensors and visualized using infrared radiation detection video equipment. This method is employed in studies on hand tracking and hand gesture interfaces [1], [17]. A hand gesture interface enables interactions between a user and a computer in places where smart environments are needed, using simple and intuitive gestures [2], [3]. A hand gesture interface is more natural and intuitive than a traditional keyboard or mouse. Thus, studies have been conducted on the interface [4], [5].

Existing hand gesture studies have concentrated on improvements in the gesture recognition rate through algorithm development and technical implementation using improved recognition. Recently, studies on more intuitive and natural human-oriented hand gesture elicitation have been conducted [6]. To improve the gesture recognition rate, studies have been conducted on gesture interfaces based on sensors or images [7]. These sensors identify gestures by recognizing hand movements, as they are attached to the fingers or back of the hand. Studies have been conducted by attaching sensors to each joint of the hand, thereby accurately identifying not only hand movements but also the shape of the hand to recognize gestures [8], [9].

Studies have been conducted based on images (i.e., two-dimensional (2D) shapes) and the three-dimensional (3D) shape of the hand, including the concept of depth [2], [10]–[12]. More recently, studies on thermal sensors were conducted in addition to studies on visible light and depth, in which the hand's temperature was visualized to detect only high-temperature objects [13]. Most vision-based methods assume that the light axis of the camera is perpendicular to the palm to minimize hidden phenomena. This assumption also reduces the degrees of freedom (DoF) of the hand by fixing the palm's pose, thereby making hand tracking faster and more robust [11] From a technical point of view, there are two types of hand tracking approaches: model-based and appearance-based. Since the model-based method performs hand tracking by changing the hand model variably and comparing it with features acquired through images, it can perform accurate tracking without limitations related to the palm. However, numerous computations are required to transform the hand model and create hand candidates for hand tracking. The appearance-based method, on the other hand, requires less computations than the model-based method, since it selects an image whose probability is the largest by using

hand candidate images and feature matching in the image sequence [14], [15].

We propose a hand tracking method that uses information detected in thermal and visible light images. Compared to previous studies, our research is novel in the following three ways. First, we propose an adaptive hand detection (AHD) algorithm. AHD generates five models of temperature by analyzing different regions of interest. This model is used for accurate hand detection. Second, whereas existing trackers are semi-automatic, we have linked AHD with the tracker to implement automatic tracking by detection. Furthermore, we have implemented every single object to be specified in the thermal camera environment. Lastly, we have improved the performance of automatic tracking by detection. We designed the proposed AHD algorithm to be referenced as the guide in the middle of tracking. We call this the guidance framework for tracking by detection (GFTD), and its performance was improved by correcting the hand coordinates from AHD through the GFTD. The remainder of this paper is organized as follows. Section 2 introduces thermal cameras and their features and reviews studies related to hand tracking and studies using thermal sensors. Section 3 describes in detail the proposed detection method, AHD, and the principle of the GFTD, which integrates with AHD and the tracker. Section 4 describes the experimental environments, datasets, evaluation indices, and evaluation results. Finally, Section 5 presents the conclusions and suggestions for future work.

II. RELATED WORK

A. THERMAL CAMERA APPLICATIONS

A thermal camera is a camera that can visualize objects via thermal images obtained by sensing radiation emitted from the objects' surfaces.

For example, it can be used in health-care, surveillance, monitoring, and safety by visualizing special information that cannot be seen by human eyes, or by detecting specific objects [1], [16], [18]. There are a number of visualization methods, but the color map matching method is the most widely used. A color map is color information that is mapped and represented with a color space to visualize thermal data. There are five types of color maps that can be used to depict thermal information: gray, iron, rainbow, saturation, and blue-red blended. In the experimental environment, data were acquired by setting the temperature between 23°C and 37°C, and a color distribution map was represented using a normalization technique employing the maximum and minimum values. The color space used in the experiment was a gray space. Figure 1 shows color images acquired using a FLIR Systems ThermalCAM S65. As shown in Figure 1, color images can have various appearances.

B. HAND TRACKING USING THE THERMAL CAMERA

Hand-gesture-based user interfaces can be divided into two modes: touch and touchless. To acquire a user's motion data, touch and touchless modes are implemented by using motion recognition devices and cameras, respectively. Although a number of methods have been studied using color or depth information [12], [19], they suffer from inherent problems, such as illumination changes that are difficult to overcome with the sensors used.

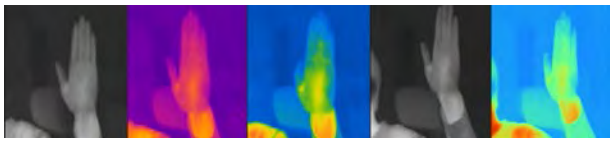


FIGURE 1. Five color images acquired with the FLIR S65 camera: (a) gray, (b) iron, (c) rainbow, (d) saturation, (e) blue-red blended.

TABLE 1. Representative works in vision-based human detection.

Research Study	Sensor	Core Algorithm	Target Region	Note
Zeng et al. [20]	Visible, Thermal	head template tracking method	Bounding box	Hand Tracking, Recognition (HMM, FSM)
Hanif et al. [21]	Visible, Thermal	DWT, Gabor Filter	Bounding box	Face Recognition
Leykin et al. [22]	Visible, Thermal	Particle Filter	Bounding box	Pedestrian Tracking, Visible (Grayscale)
Li et al. [23]	Visible, Thermal	Laplacian Sparse representation	Bounding box	Object Tracking, Pedestrian Estimation
Kim et al. [13]	Visible, Thermal	TLD, Thermal Guided Joint Filter	Bounding box	Hand Tracking, Gesture Recognition
Dias et al. [25]	Thermal	Geometrical Histogram	Contour	Hand Pose Recognition
Ours	Thermal	TLD, KCF, MIL, BOOSTING, MEDIANFLOW	Bounding box, Contour	Hand Tracking

Recently, thermal-based methods were employed to solve problems related to illumination changes using temperature information. Zeng et al. [20] proposed a hand pose detection algorithm and demonstrated substantial results even when using a beam projector as the light source. They fused thermal and visual cameras together; therefore, it was robust against illumination changes. Since the hand tracking algorithm requires visual sensor information, Zeng et al. [20] limited the tracking range of the beam projector to remain within the rectangular area. However, the range of tracking is limited to the region of the beam projector, and the hand tracking algorithm still requires visual sensor information.

Hanif et al. [21] fused visible light and thermal images together to implement a new face recognition system. Gabor filtering technology was introduced in this paper; it can extract facial features and improve performance in face recognition. Leykin et al. [22] proposed a pedestrian tracker using a particle filter. With this method, they generated a background

model through a filter, where each pixel is represented as a multimodal distribution with a different number of modalities for both grayscale and thermal input. Li et al. [23] suggested a Laplacian sparse representation method in object tracking that can fuse different sources of data with the aim of generating a multimodal feature model. The multimodal feature model was developed to improve the occlusion and the local spatial information and to detect pedestrians by fusing thermal and visible light images.

More interestingly, Kim et al. [13] proposed an algorithm that efficiently matches two images, to overcome the drawbacks of existing methods. In their method, the visual and thermal sensors are calibrated. In addition, a thermally guided joint filter capable of representing the contours of images was introduced to solve the problem of identifying hand movements in dark locations and over long distances. On the other hand, Dias et al. [25] proposed the open gestures recognition engine (OGRE) system for hand movement detection at a short distance, which can convert hand poses into Portuguese sign language. In this system, the histogram of the distances and the angle between the contour edges are used to derive a contour signature, called the pairwise geometrical histogram. Additionally, classification is performed by comparing the pairwise geometrical histograms representing the gestures.

As described above (and summarized in Table 2), most studies have attempted to gain advantages through the fusion of visual and thermal sensors. Thus, the existing methods employ visual images to solve the problems by relearning feature maps by fusing sensors. However, these methods increase the amount of computation as they require image fusion, including calibration and registration procedures.

III. PROPOSED METHOD

We propose AHD, an algorithm that is robust to illumination and human skin color in a thermal sensor environment. An experiment was conducted involving the fusion of five trackers that are currently widely used. As Figure 2 shows, in AHD hand detection coordinates are used as the initial input to the tracker. The present paper also proposes GFTD. With existing methods, trackers accumulate error in the detected hand region over time. To solve this problem, GFTD tracks hand movements and detects a hand coordinate region additionally detected by the AHD algorithm. Since detection error accumulates through comparisons with the tracker coordinates, the hand location can be recognized anew for each frame cycle, and accurate coordinate correction is thereby accomplished.

A. ADAPTIVE HAND DETECTION (AHD)

In this section, the AHD algorithm to detect a hand efficiently from thermal images is briefly explained using pseudocode as well as through the block diagram shown in Figure 2.

The following three constraints should be satisfied to ensure robust hand detection in a thermal sensor environment. First, the hand temperature and the temperature near the hand should be clearly distinguished. Second, computation cost should be low to detect the hand in real time. Finally, since the temperature changes with changes in hand conditions, it should be adaptive to temperature environments.

The proposed algorithm employs a method of estimation that learns the hand temperature from the initial thermal image and infers a temperature distribution from the learning model. Thus, it is a nonparametric method, and a Gaussian model was adopted. The proposed algorithm is presented with pseudocode in Algorithm 1.

Algorithm 1 Adaptive Hand Detection(AHD)

Input: Thermal image T
Generate Model: $M = m_1, m_2, m_3, m_4, m_5$
▷ $m_i : 7 \times 7$ *Matrix*
Output : Hand binary image
procedure AHD (T, M)
    Grayscale norm: y = ((x-min)*255/(max-min))
    Get the element(m_i) at the position from the specified array
    for each 5 elements i from 0 to 4 **do**
        (G_i, Range(Threshold)) = Gaussian Model(m_i)
    end for
    for each 5 elements i from 0 to 4 **do**
        B_i = Binary image(T, $G_i Threshold$)
    end for
    MImage = Merge(B_i) ▷ Combine five images into one
    MNImage = NoiseReduction(*MImage*)
end procedure

In the pseudocode, after the function receives the input image (T), it returns the hand's binary images (MNImage). A number of problems can be easily resolved by using thermal images of a relatively uncomplicated heat map, since Gaussian model measures hand temperature to separate the hand from the background. However, it cannot define a specific temperature because of the unevenness in the hand temperatures. To overcome this shortcoming, a Gaussian model is employed, as presented in Equation (1).

Under controlled temperature conditions, surface temperatures of different individuals cluster in a small region in the color space. Hence, under certain temperature conditions, the surface temperature distribution of different individuals can be modeled by a multivariate normal distribution in normalized color space [26]–[28]. The temperature distribution of the surface is modeled by a Gaussian probability density function (pdf), defined as

$$F(T) = \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(T-\mu_m)^T \sum^{-1}(T-\mu_m)} \quad (1)$$

where T is a thermal vector and μm and $\sum$ are the mean vector and the diagonal covariance matrix, respectively. The model parameters are estimated from the training data by

$$\mu_m = \frac{1}{n}\sum_{j=1}^{n} T_j; \quad \Sigma = \frac{1}{n-1}\sum_{j=1}^{n}(T_j - \mu_m)(T_j - \mu_m)^T \quad (2)$$

where n is the total number of surface temperature pixels T_j. The parameters, μ_m and $\sum$, are estimated over all the temperature samples T_j from the training data using the maximum likelihood estimation approach. The probability $F(T)$ can be used directly as a measure of hand surface temperature likelihood, and the classification is normally obtained by comparing it to a certain threshold value estimated empirically from the training data.

The Gaussian model is made, and a reference point is determined. Temperatures for five regions of the palm are acquired, and each of the regions of interest (ROI) is calculated to obtain $F(T)$, as presented in Equation (1). Finally, a threshold value that can classify hand and background regions from the thermal image can be identified through $F(T)$. A binary image is created by using each of the five thresholds, and the five binary images are merged. The merged single binary image is used to detect the hand position through noise reduction and analysis of the convex points in the contours and the convexity defects [29].

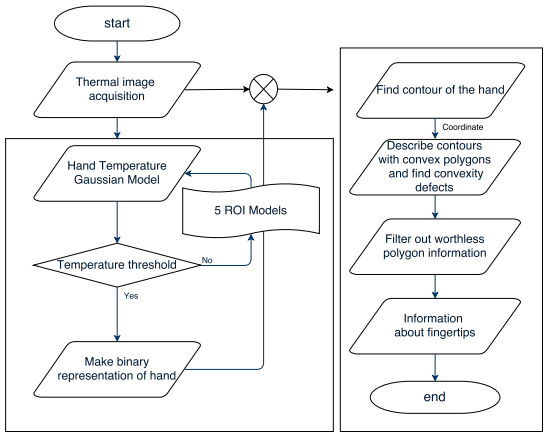


FIGURE 2. Block diagram of the proposed AHD method.

The input images of the thermal camera have a resolution of 640 × 480. The temperature of the hand should be measured to track the hand adaptively. The measurement method is as follows: five windows are created, as shown in Figure 3, and temperatures are acquired at places around the five locations on the palm.

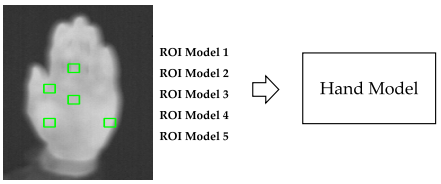


FIGURE 3. Locations indicated for creating a model from hand temperature images at each of five designated blocks.

A thermal image is characterized by changes in intensity, as it is normalized according to acquired temperature information, unlike visible light or depth images. Thus, since each binary image created using the five models has a different

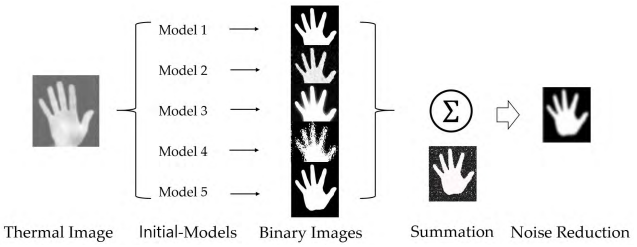


FIGURE 4. Four phases in creating a representative hand model.

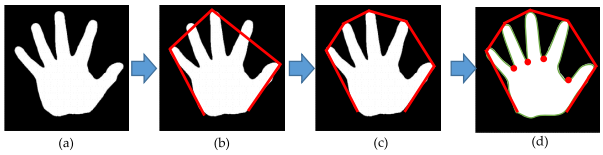


FIGURE 5. Four phases in determining a hand contour line: (a) binary image of hand, (b) result of hand convex points using the Ramer-Douglas-Peucker algorithm, (c) result of processing the convex hull, (d) result with convexity defects.

threshold, five different results of various shapes are revealed, as shown in the binary images in Figure 4. After this, a representative binary image is created by producing a merged image by summing all images to create a single binary image. Noise is removed from the representative binary image using a median filter, which is a noise reduction algorithm.

After noise removal, a hand is detected, as shown in Figure 5. The contours of the hand should be detected first to find the hand shape. Thus, a hand contour can be found via analysis using the convex hull and the convexity defects. The convex hull analysis searches for vertices that connect the contour line of points in 2D space, and a polygon of minimum size is calculated, thereby obtaining a result as shown in Figure 5(b). This method sets a threshold to find a hand contour. Since the algorithm proposed in Graham's scan can extract a hand contour only when the length of the contour line exceeds a threshold of n pixels, a threshold is set using the data statistics of the pixels and based on a hand shape. When the fingers are spread out, the hand is the largest, and when the hand is clenched into a fist, the hand is the smallest [30].

The convexity defects are obtained using the convex hull and the contour line. The convex hull is a polygon that connects the outermost contour lines, and the convexity defects are the areas between the fingers within the convex hull of the hand, as shown in Figure 5(c), In Figure 5(d) the red line indicates the convex hull, the green curve indicates the hand contour, and the red dots indicate the convexity defects of the hand. Fingers are easily detected by using the coordinates of the convexity defects.

B. GUIDANCE FRAMEWORK FOR TRACKING-BY-DETECTION

This section describes the increase in the tracking speed obtained by handing over the bounding box coordinates

detected by the AHD algorithm to the tracker, which is connected to the AHD. The structure of the GFTD is also explained(Figure 6).

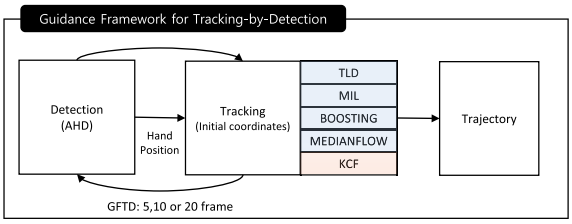


FIGURE 6. Structure of the system that tracks a hand and records a path by combining the proposed AHD method and five trackers.

There are five different trackers built into the GFTD, each of which uses one of the following different tracking algorithms: Tracking Learning Detection (TLD) [33], Multiple Instance Learning (MIL) [36], BOOSTING [39], MEDIANFLOW [40], and Kernelized Correlation Filters (KCF) [42].

For example, we introduced KCF into the GFTD. Henriques and Joao et al. [42] proposed a circulant structure with kernel (CSK) by using a circulant matrix and then a KCF tracker, which represents an extended version of CSK.

The KCF tracker models object appearance by using a correlation filter w trained on an image patch x of $M \times N$ pixels, where all the circular shifts of $x_{i,j}(i,j) \in \{0, 1, \ldots, M-1\} \times \{0, 1, \ldots, N-1\}$, are generated as training samples with the Gaussian function label $y_{i,j}$. The goal is to find the optimal weights w in Equation (3).

$$w = \min_{w} \sum_{i}^{n} (f(x_{i,j}) - y_{i,j})^2 + \lambda ||w||^2 \tag{3}$$

where w is the model parameter matrix, and $x_{m,n}$ denotes each image patch used to train the model. The λ is a regularization parameter that controls over-fitting. Regression target $y_{i,j}$ is the desired output of $f(x_{i,j})$.

To utilize the kernel trick and reformulate the Ridge Regression problem in dual space, the model can be written as $f(z) = w^T z = \Sigma_{i=1}^{m,n} \alpha_{i,j} \kappa(z, x_{i,j})$ and applied to allow for a more powerful classifier. For the most commonly used kernel functions, the circulant matrix trick can also be used. The dual space coefficients α can be learnt as below:

$$\hat{\alpha} = \frac{\hat{y}}{\hat{k}^{\dot{x}\ddot{x}} + \lambda} \tag{4}$$

where $\hat{}$ denotes the discrete Fourier transform and α is a matrix consisting of coefficients $\alpha_{i,j}$. $\hat{k}^{\dot{k}\ddot{k}}$ represents the kernel correlation operation and is defined as follows:

$$\hat{k}^{\dot{x}\ddot{x}} = exp(-\frac{1}{\sigma^2}(\|\dot{x}\|^2 + \|\ddot{x}\|^2) - 2F^{-1}(\hat{\dot{x}}_d^* \cdot \hat{\ddot{x}}_d)) \tag{5}$$

where σ is the size of the Gaussian kernel, F^{-1} denotes the inverse discrete Fourier transform, and $*$ refers to complex conjugation. Here d th denotes the dimension of the image feature patch. Equation (4) can be used to initialize the coefficient matrix in dual space.

$$\hat{f}(z) = \hat{k}^{\mathring{x}p} \cdot \hat{\alpha} \tag{6}$$

where $\mathring{x}$ denotes the data to be learnt in the model. The location of the maximum element in f corresponds to the cyclic shift of z that is most similar to the current target appearance $\mathring{x}$

We looked at the main idea of KCF. Therefore, five hyper-parameters were used in the experiment. The KCF has a number of hyper-parameters including the padding size, desired Gaussian response width, learning rate, regularization and Gaussian kernel bandwidth. Table 1 summarizes the hyper-parameters of KCF.

TABLE 2. KCF hyper parameters.

Parameter	Value	note
Padding	1.0	extra area surrounding the target
sigma factor	1/16	spatial bandwidth (proportional to target)
Sigma	0.2	Gaussian kernel bandwidth
Lambda	1e-2	Regularization
Learning Rate	0.075	learning rate for appearance model update scheme

Our two contributions are as follows: First, the existing semi-automated tracking-by-detection algorithm was redesigned to be automatic by the implementation of a hand tracking algorithm that is specialized for thermal environments. Second, the GFTD system measures the effect of the intervention of AHD on the performance of each cycle and contributes to performance improvements through verification by means of the optimal parameter values. In addition, the GFTD parameter in AHD aims to create dense trajectory data by optimizing the relationship between the detection algorithm and the tracker. That is, dense trajectory data result in improved tracking performance.

IV. EXPERIMENTS AND DISCUSSION

Studies on hand gestures are generally conducted in indoor environments, and the detection performance varies significantly with the lighting. Thus, we prepared two experiments. First, we tested AHD and the five tracker algorithms with visible light images under different lighting conditions. Second, we tested AHD and the tracker algorithms using a thermal camera and applied GFTD for performance improvements.

A. HARDWARE AND SOFTWARE

We captured thermal and color images through VGA at 30 fps using the FLIR S65 (Figure 7(a)) and Microsoft LifeCam (Figure 7(b)), respectively. Data acquisition was implemented in OpenCV, and other modules, including the tracker, were implemented using C/C++ on a machine with an Intel i7 3.4 GHz processor and 20 G of physical memory (RAM). It also supports bindings to Python/MATLAB.

(a) (b)

FIGURE 7. (a) FLIR S65 thermal camera, (b) visual camera.

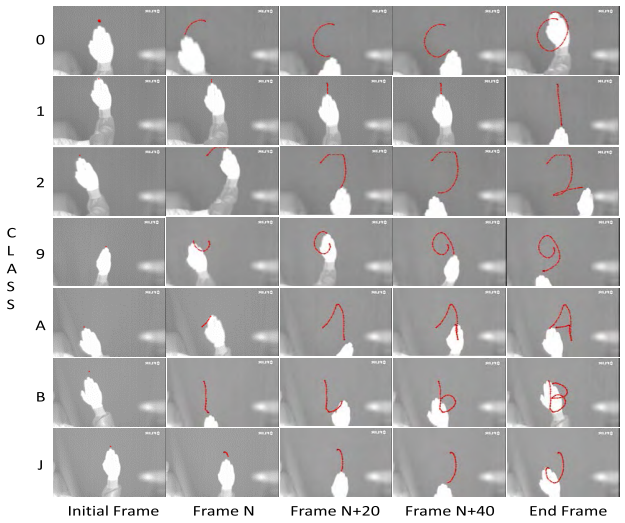


FIGURE 8. Sample of hand database consisting of data with numbers and characters, acquired using thermal camera.

B. DATABASE

In this study, we constructed our own hand tracking database using thermal cameras (Figure 8). In the experimental environment, the temperature range for camera acquisition was 23–37°C. with a resolution of 640 × 480. The dataset included data from a total of 20 female and male persons. The data consisted of a number dataset (0-9) and a character dataset (A-J), and numbers (0-9) and characters (A-J) were combined to create a total of 20 classes. Each class contained three sequences, and each sequence consisted of 100 frames (34 s) on average. The ground truth and trajectory were analyzed completely using a total of 1200 sequences, and detection and tracking were configured to be evaluated objectively using the dataset.

Additionally, we created a visual dataset (Figure 9) to determine whether the same result shown in the thermal sensor environment could be produced in the visual environment. To do this, we configured three environments (lighting ON, lighting OFF, and backlit) and acquired data with a resolution of 640 × 480 at the same location as that used for the thermal database (DB). The three environments were composed of ten classes from 0 to 9. The lighting-OFF subset, the lighting-ON subset, and the backlighting subset consisted of 425 frames, 445 frames, and 476 frames, respectively.

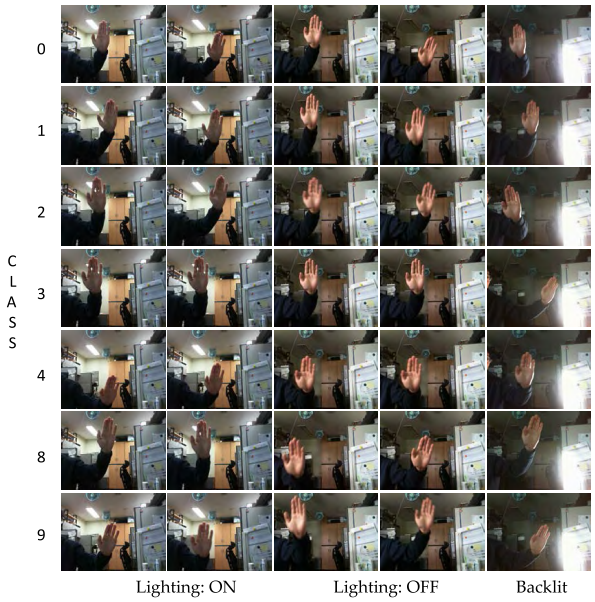


FIGURE 9. Sample of hand dataset consisting of data with numbers, acquired using visual camera.

C. EVALUATION METRICS

There are two main types of evaluation metrics: intersection over union (IOU) and object tracking error (OTE). In general, the success rate (IOU) method is employed to evaluate whether object detection has been done correctly. For this evaluation metric, the result is defined as good if the IOU of the bounding box where the confidence score is predicted and the IOU of the bounding box of the ground truth are the same. The success rate is between 0 (no overlap) and 1 (perfect overlap). The IOU index has five elements: x, y, width, height, and confidence score. (x, y) are the starting coordinates of the bounding box, and width and height refer to the predicted size of the hand with respect to the entire image. The confidence score is the IOU index between the prediction bounding box and the ground truth bounding box. In this study, the success rate(IOU) was calculated as presented in Equation (7).

$$IOU(p, g) = \frac{area(p \cap g)}{area(p \cup g)} \tag{7}$$

In Equation (7), p represents the prediction bounding box, which is a result of the tracking-by-detection algorithm, and g represents the ground truth. An overlap ratio is calculated and evaluated using the two coordinate values for the bounding boxes p and g.

The object tracking error (OTE) is an index that evaluates a hand's direction of motion over time. This OTE method can evaluate a path of motion by comparing it with the ground truth after detecting a fingertip.

$$OTE = \frac{1}{N_{tg}} \sum \sqrt{(x_i^g - x_i^t)^2 + (y_i^g - y_i^t)^2} \tag{8}$$

where N_{tg} represents the total number of overlapping fingertips between the ground truth and tracking system results;

x_i^g, y_i^g represent the coordinates of the centroid of the fingertip in the i point of the ground truth; and $x_i^t y_i^t$ represent the coordinates of the centroid of the fingertip in the i point of the GFTD result [45].

V. EXPERIMENTAL RESULTS

A. RUNNING TIME

Table 3 illustrates the running times of the detection and trackers, which were measured using the C++ language. With the GFTD(5) algorithm, MEDIANFLOW showed the best tracker performance, followed by KCF, BOOSTING, MIL, and TLD. Although there were differences among the algorithms, the algorithm without GFTD had the shortest running times on average. Processing time differences between GFTD(5) and GFTD(20) were very small. However, the difference was more pronounced with KCF. In this case, the tracker performed faster with GFTD than without since KCF iterates hand tracking when it fails at the detection.

MEDIANFLOW showed considerably better performance in terms of execution time than the other trackers. MEDIANFLOW is a median feature vector extraction algorithm that uses the Forward-Backward error to calculate the error of the feature vectors. The Forward-Backward error correctly eliminates 50% of the key points from the median flow tracker as odd points. Then, it finds the median value of the rest of the key points to determine the flow of the object. These factors allowed MEDIANFLOW to be faster than the other trackers.

TABLE 3. Processing times *(ms)* of trackers with guidance framework for tracking-by-detection (GFTD) methods.

Algorithm	BOOSTING	KCF	MEDIANFLOW	MIL	TLD	Average
Without GFTD	32.93	72.90	3.66	54.96	169.16	66.72
GFTD (20)	44.93(+12)	37.83(-35.07)	3.46(-0.2)	64.48(+9.52)	554.90(+385.74)	141.12(+74.4)
GFTD (10)	48.80(+15.87)	34.91(-37.99)	3.39(-0.27)	64.65(+9.69)	580.99(+411.83)	146.54(+79.82)
GFTD (5)	48.62(+15.69)	33.33(-39.57)	3.46(-0.2)	65.59(+10.63)	411.40(+242.24)	112.48(+45.76)
Average(GFTD (n))	47.45(+14.52)	35.36(-37.54)	3.44(-0.22)	64.90(+9.94)	515.76(+346.60)	133.38(+66.66)

B. INTERSECTION OVER UNION (VISUAL DATASET)

The visual dataset was acquired under three conditions: lighting ON, lighting OFF, and backlit; the success rate evaluation was conducted with all 1346 frames. According to the results (Table 4), all five trackers had degraded performance. The reason is that those were implemented in the lighting OFF and backlit environments. The lighting environment factor affects the performance of all trackers. As they extract feature vectors to track objects based on image intensity, noisy image intensity degrades all trackers' performance. KCF uses a cyclic method with a good detection rate, whereas TLD uses optical flow [33] with a low detection rate, and the experiment found that the cyclic method is more robust against noise and brightness than the optical flow algorithm.

Figure 10 and Table 4 shows the success rate(IOU) evaluation results for the visual dataset. The analysis results show that Color-AHD is characterized by a sudden decrease in the detection rate in the region of a 60%–80% overlap threshold. Furthermore, the best performing tracker was

TABLE 4. Success rate(IOU) results for visual dataset with five tracker.

Overlap Threshold	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Color-AHD	100.0	90.3	87.0	82.2	69.8	56.1	48.0	38.3	25.9	12.6	6.8
Color-BOOSTING	100.0	87.7	77.6	67.3	55.3	44.5	36.6	23.8	11.0	3.7	2.2
Color-KCF	100.0	88.8	80.2	70.6	60.9	53.4	45.9	37.5	22.3	7.1	2.8
Color-MEDIANFLOW	100.0	82.1	69.8	52.8	43.4	37.1	31.5	21.5	10.1	4.4	3.0
Color-MIL	100.0	85.7	73.5	60.4	44.4	34.4	25.6	17.4	9.2	3.0	2.4
Color-TLD	100.0	77.5	69.5	57.7	46.2	35.9	23.8	12.4	5.6	4.4	4.1

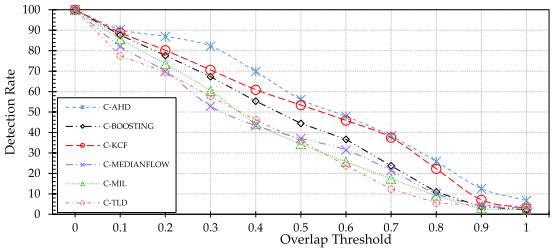


FIGURE 10. Success rate(IOU) results for visual dataset with five tracker.

C-KCF, followed by C-BOOSTING, C-MIL, C-TLD, and C-MEDIANFLOW. As verified through the experiment, as the overlap threshold became larger, the detection rate decreased linearly, and no significant difference in performance was found between trackers for the visual dataset. Thus, GFTD is expected to have an insignificant effect on visual images. Furthermore, hand gestures using colors in a drastically changing indoor environment are susceptible to illumination. An objective assessment can be made by comparing the visual and thermal datasets using the same evaluation metric.

C. INTERSECTION OVER UNION (THERMAL DATASET)

The success rate(IOU) evaluation results for the thermal dataset are shown in Table 5 and Figure 11. The proposed hand detection algorithm (AHD) shows a detection success rate of 86.8%. In addition, when GFTD was not run, KCF had the best performance, whereas TLD had relatively severely degraded performance. All five trackers improved their performance by running GFTD. Here, the TLD algorithm had the best improvement in performance: from 31.7% to 65.9% (GFTD(5)), which is an improvement of 34.2 percentage points. The KCF algorithm, which had the best performance, also improved from 73.2% to 82.4% (GFTD(5)), an improvement of 9.2 percentage points. The results in Figure 11 verify that GFTD helps to improve the performance of the tracker.

In particular, The performance of GFTD became more accurate by adopting the number of guidance trials 5 or 10 than 20 times.

This GFTD parameter exhibited a cycle that could repeatedly update the hand detection coordinates. Finally, there are two reasons for the high performance of KCF. Henriques et al. [42] proposed a method to resolve the ridge regression problem in the space domain using the optimal filter, and robust tracking was implemented using the HOG feature descriptor. Thus, high performance can be maintained

TABLE 5. Average success rate (IOU) (%) (thermal dataset).

GFTD: (n)	Without GFTD	GFTD (20)	GFTD (10)	GFTD (5)	Average (GFTD(n))
AHD	86.8	86.8	86.8	86.8	86.8
BOOSTING	63.8	74.1(+10.3)	75.7(+11.9)	76.9(+13.1)	75.6(+11.8)
KCF	73.2	79.6(+6.4)	81.3(+8.1)	82.4(+9.2)	81.1(+7.9)
MEDIANFLOW	62.1	75.7(+13.6)	76.5(+14.4)	77.1(+15)	76.4(+14.3)
MIL	65.7	75.1(+9.4)	76.6(+10.9)	77.7(+12)	76.5(+10.8)
TLD	31.7	56.9(+25.2)	62.5(+30.8)	65.9(+34.2)	61.8(+30.1)
Five-Tracker Average	59.3	72.3(+13)	74.5(+15.2)	76.0(+16.7)	74.3(+15)

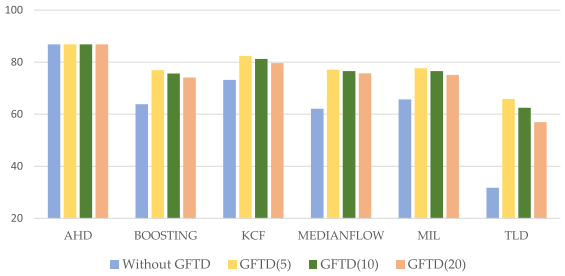


FIGURE 11. Average success rate (IOU) (%) (thermal dataset).

even with thermal images.

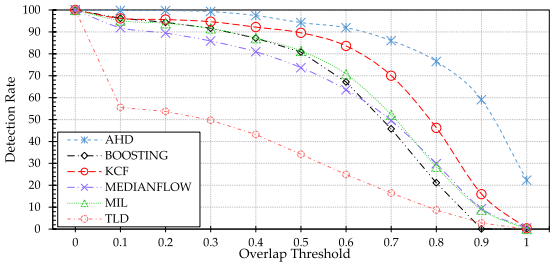


FIGURE 12. Success rate (IOU) results with five tracker.

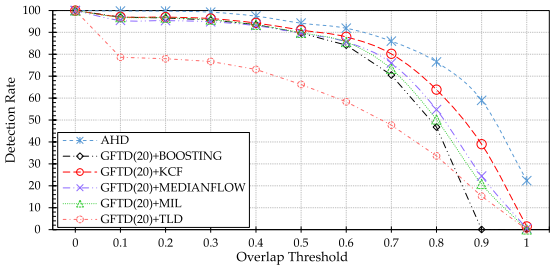


FIGURE 13. Success rate (IOU) results with GFTD(20).

Figure 12 shows the success rate(IOU) evaluation results when GFTD was not running, where the overlap threshold value was increased from 0 in steps of 0.1 to measure an evaluation value. As the overlap threshold approached 1, the hands were more precisely detected. Figure 13 shows the results when the GFTD parameter was 20; Figure 14 and Figure 15 show the evaluation results when the GFTD parameter was 10 and 5, respectively.

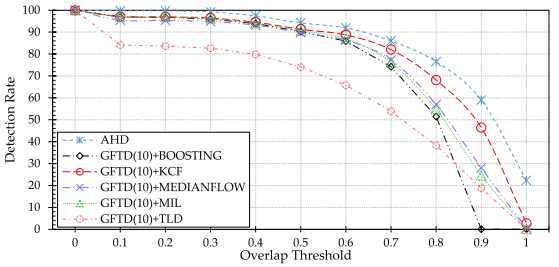


FIGURE 14. Success rate (IOU) results with GFTD(10).

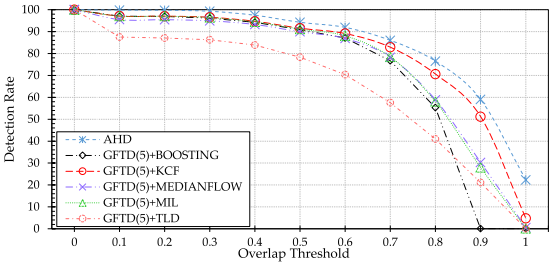


FIGURE 15. Success rate (IOU) results with GFTD(5).

D. OBJECT TRACKING ERROR (THERMAL DATASET)

The object tracking error (OTE) indicates the distance between the ground truth and the trajectory of the tracked object. As shown in Table 6 and Figure 16, the proposed AHD algorithm had the best performance when the location error threshold was cumulative average; the precision rate was 88.4%. In contrast, KCF (with GFTD (5)) had the best tracking performance in the OTE evaluation, 83%. The TLD algorithm, which performed worse than the other tracker algorithms, showed a performance of 68% with GFTD(5), a pattern similar result to that in the success rate (IOU) evaluation analysis. Figure 16 shows the OTE evaluation results when the GFTD was not running, where the distance threshold value was increased from 0 in steps of 10 to measure an evaluation value. As the distance threshold approached 1, the hands were more precisely detected. Figure 17 shows the results when the GFTD parameter was (20); Figure 18 and Figure 19 show the evaluation results when the GFTD parameter was 10 and 5, respectively. The performance of all trackers tended to increase when the GFTD iteratively updated the hand detection coordinates.

E. DISCUSSION

A visual sensor is a sensor that visualizes a visible wavelength. This sensor represents information in a manner similar to that of the human eye. Thus, a large number of researchers have concentrated on hand tracking using visual cameras. Moreover, the proposed method in the present study employs LWIR to track hands. The LWIR has a wavelength of 8–15μm, and a thermal camera was developed using sensors and a lens that can detect light within the range of the LWIR.

In this study, the FLIR Systems S65 model, which can sense radiant energy emitted from the human body and

TABLE 6. Average object tracking error (OTE)% (thermal dataset).

GFTD: (n)	Without GFTD	GFTD (20)	GFTD (10)	GFTD (5)	Average (GFTD (n))
AHD	88.4	88.4	88.4	88.4	88.4
BOOSTING	65.5	76.1(+10.6)	77.5(+12)	78.6(+13.1)	77.4(+11.9)
KCF	75.9	80.7(+4.8)	81.9(+6)	83.0(+7.1)	81.9(+6)
MEDIANFLOW	60.6	78.1(+17.5)	78.4(+17.8)	78.7(+18.1)	78.4(+17.8)
MIL	65.0	76.9(+11.9)	78.3(+13.3)	79.2(14.2)	78.1(+13.1)
TLD	31.3	60.0(28.7)	65.1(+33.8)	68.0(+36.7)	64.4(+33.1)
Five Tracker Average	59.7	74.4(+14.7)	76.2(+16.5)	77.5(+17.8)	76(+16.3)

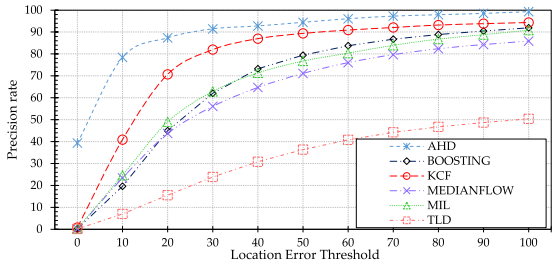


FIGURE 16. Object tracking error (OTE) results with without GFTD.

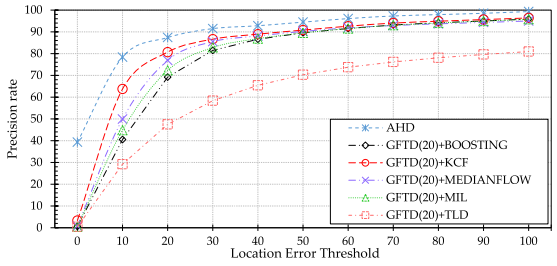


FIGURE 17. Object tracking error (OTE) results with GFTD (20).

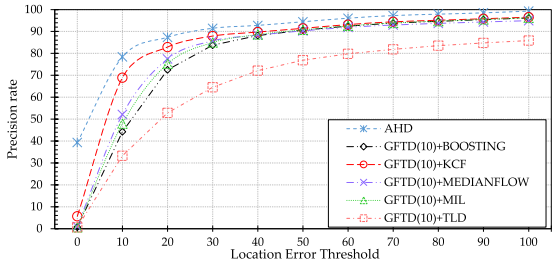


FIGURE 18. Object tracking error (OTE) results with GFTD (10).

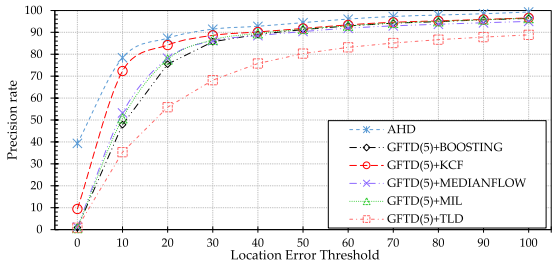


FIGURE 19. Object tracking error (OTE) results with GFTD (5).

visualize it into thermal images, was used. It provided a resolution of 640 × 480 with 30 frames per second. A user can set the temperature range and acquire images. In the present

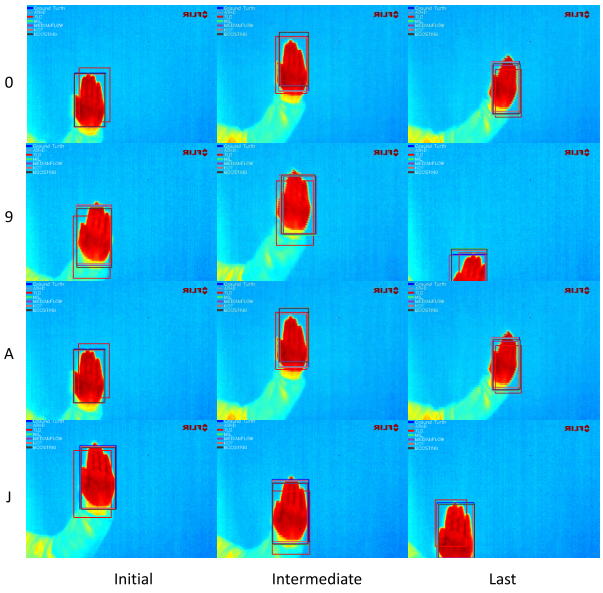


FIGURE 20. Hand tracking results.

study, we set the temperature range to 23–37°C to conduct the experiment. By using thermal images, the advantages of robustness to illumination changes as well as processing speed can be obtained. However, the thermal images showed differences in the hand temperatures relative to the surrounding temperature, thereby creating difficulty in hand detection.

This problem was a new one that had not occurred in existing studies with visible light images. To solve this problem, we proposed the AHD algorithm. This method involves creating five windows around the palm to create the initial model and using the model to generate five binary images. Then, the five images are merged into a single image to create a representative binary image, and hand positions are detected using the proposed function in AHD. Once the hand position is detected, hand tracking is performed to detect the hand's direction of motion. The hand tracking algorithm tracks hand movements. Although the algorithm was proposed in the present study to reduce the calculation time and ensure accurate tracking, the performance was not satisfactory. Thus, we selected five different tracking algorithms and proposed and evaluated GFTD.

The evaluation indices showed that the KCF algorithm exhibited the best performance. Note that the above-mentioned tracking algorithms were vulnerable to the error correction function. Accordingly, we redesigned the tracking algorithms to enable the AHD results to meddle with the tracker to correct the errors and connected AHD to the tracker. We then evaluated the effect of the changes on the tracking algorithm's performance.

The results of a comparison between the five tracker (BOOSTING, KCF, MEDIANFLOW, MIL, TLD) and GFTD showed that the performance with GFTD was better than that of the five tracker, as expected. The KCF algorithm with GFTD (5) showed the best performance with the

thermal cameras, as the thermal images were temperature-based images, which have relatively simple features. In other words, the more the algorithm depends on the presence of strong features, the greater the degradation in performance.

With thermal images, the performance of KCF and TLD differ significantly because KCF is a correlation-filter-based tracker. Thus, a correlation filter was designed to obtain a maximum value of correlation for hands. In brief, a response map can be created using the Zero Mean Normalized Cross Correlation method [42]. in the spatial domain, and a response map is calculated based on the mean in the search area and input image. The mean in the search area can have good performance if the temperature sensitivity is reduced by removing the mean temperature from the thermal image using a zero-mean method. In the TLD algorithm, the tracking and detection processes coexist, and the algorithm showed degraded performance. The analyses show that performance degradation occurs when the patches that are trained in the detection process become the target candidates. Here, performance degradation occurred when a variance filter was used to reduce the classification time by rapidly removing the patches whose variance was less than 50% of the target variance.

VI. CONCLUSIONS

In this paper, a single-sensor-based real-time hand tracking system was proposed (Figure 20).

The proposed AHD algorithm had the best performance when the success rate (IOU) was average GFTD (n); the precision rate was 86.8%. In contrast, KCF had the best performance, 82.4%, in the success rate (IOU) evaluation. The TLD algorithm, which performed worse than the other tracker algorithms, showed an success rate (IOU) performance of 65.9% in the GFTD (5) and a similar result pattern in the OTE evaluation. Finally, the performance of all trackers tended to increase when the GFTD iteratively updated the hand detection coordinates. In this system, AHD is used to detect hands in real time by reducing computation using a single sensor rather than fusing multiple sensors, and we proposed GFTD to enable accurate tracking and thereby improve hand tracking accuracy. We evaluated the accuracy of the hand detection using intersection over union (IOU) and the trajectory using object tracking error (OTE). The proposed GFTD improves the performance in terms of success rate (IOU) and OTE by 15% and 16.3%, respectively, compared to the conventional methods.

In future work, the advantages of the combination AHD and tracker proposed in this study will be investigated further by changing thermal-image-based hand shapes. Moreover, the detection of the skeleton of the hand using deep learning technology will be investigated by extending the proposed system to predict accurate hand positions and shapes, thereby increasing the utility of the system. Another research direction is to apply the proposed system to multimodal hand pose estimation systems by fusing LWIR and NIR (Near-infrared) sensors.

ACKNOWLEDGMENT

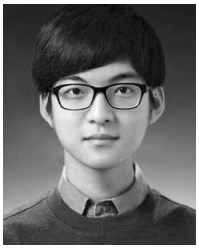
This work was supported by Institute for Information and communications Technology Promotion(IITP) grant funded by the Korea government (MSIP)(No.2016-0-00197, Development of the high-precision natural 3D view generation technology using smart-car multi sensors and deep learning)

REFERENCES

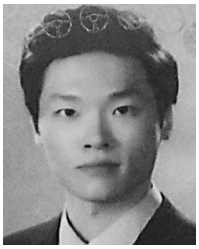
[1] R. Gade and T. B. Moeslund, "Thermal cameras and applications: A survey," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 245–262, 2014.
[2] J. P. Wachs *et al.*, "A gesture-based tool for sterile browsing of radiology images," *J. Amer. Med. Informat. Assoc.*, vol. 15, no. 3, pp. 321–323, 2008.
[3] H. I. Stern, J. P. Wachs, and Y. Edan, "Optimal consensus intuitive hand gesture vocabulary design," in *Proc. IEEE Int. Conf. Semantic Comput.*, Aug. 2008, pp. 96–103.
[4] A. S. M. M. Rahman, M. A. Hossain, J. Parra, and A. El Saddik, "Motion-path based gesture interaction with smart home services," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 761–764.
[5] A. R. Varkonyi-Koczy and B. Tusor, "Human–computer interaction for smart environment applications using fuzzy hand posture and gesture models," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 5, pp. 1505–1514, May 2011.
[6] E.-J. Choi, S.-H. Kwon, D.-H. Lee, H.-J. Lee, and M.-K. Chung, "Design of hand gestures for smart home appliances based on a user centered approach," *J. Korean Inst. Ind. Eng.*, vol. 38, no. 3, pp. 182–190, 2012.
[7] S. Chantasuban and S. Thiemjarus, "UbiBand: A framework for music composition with BSNs," in *Proc. 6th Int. Workshop IEEE Wearable Implant. Body Sensor Netw. (BSN)*, Jun. 2009, pp. 267–272.
[8] C. Zhu and W. Sheng, "Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 3, pp. 569–573, May 2011.
[9] V.-R. Jaijongrak, S. Chantasuban, and S. Thiemjarus, "Towards a BSN-based gesture interface for intelligent home applications," in *Proc. ICCAS-SICE*, Aug. 2009, pp. 5613–5617.
[10] C. Manresa, J. Varona, R. Mas, and F. J. Perales, "Hand tracking and gesture recognition for human-computer interaction," *Electron. Lett. Comput. Vis. Image Anal.*, vol. 5, no. 3, pp. 96–104, Jul. 2005.
[11] Y. Fang, K. Wang, J. Cheng, and H. Lu, "A real-time hand gesture recognition method," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2007, pp. 995–998.
[12] H. Guan, R. S. Feris, and M. Turk, "The isometric self-organizing map for 3D hand pose estimation," in *Proc. 7th Int. Conf. IEEE Autom. Face Gesture Recognit. (FGR)*, Apr. 2006, pp. 263–268.
[13] S. Kim, Y. Ban, and S. Lee, "Tracking and classification of in-air hand gesture based on thermal guided joint filter," *Sensors*, vol. 17, no. 1, p. 166, 2017.
[14] M. Donoser and H. Bischof, "Real time appearance based hand tracking," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2008, pp. 1–4.
[15] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision based hand pose estimation: A review," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 52–73, 2007.
[16] J. H. Lee *et al.*, "Robust pedestrian detection by combining visible and thermal infrared cameras," *Sensors*, vol. 15, no. 5, pp. 10580–10615, 2015.
[17] B. Chakraborty and B. K. Sinha, "Estimation of emissivity with the help of an infrared camera," *Sensors Transducers*, vol. 148, no. 1, p. 66, 2013.
[18] Y. Ma, X. Wu, G. Yu, Y. Xu, and Y. Wang, "Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery," *Sensors*, vol. 16, no. 4, p. 446, 2016.
[19] C.-S. Chua, H. Guan, and Y.-K. Ho, "Model-based 3D hand posture estimation from a single 2D image," *Image Vis. Comput.*, vol. 20, no. 3, pp. 191–202, Mar. 2002.
[20] B. Zeng, G. Wang, and X. Lin, "A hand gesture based interactive presentation system utilizing heterogeneous cameras," *Tsinghua Sci. Technol.*, vol. 17, no. 3, pp. 329–336, Jun. 2012.
[21] M. Hanif and U. Ali, "Optimized visual and thermal image fusion for efficient face recognition," in *Proc. 9th Int. Conf. IEEE Inf. Fusion*, Jul. 2006, pp. 1–6.

[22] A. Leykin and R. Hammoud, "Pedestrian tracking by fusion of thermal-visible surveillance videos," *Mach. Vis. Appl.*, vol. 21, no. 4, pp. 587–595, 2010.
[23] C. Li, S. Hu, S. Gao, and J. Tang, "Real-time grayscale-thermal tracking via laplacian sparse representation," in *Proc. Int. Conf. Multimedia Modeling*, 2016, pp. 54–65.
[24] M. Talha and R. Stolkin, "Particle filter tracking of camouflaged targets by adaptive fusion of thermal and visible spectra camera data," *IEEE Sensors J.*, vol. 14, no. 1, pp. 159–166, Jan. 2014.
[25] J. J. Dias, B. Bhowal, C. J. Wildin, and J. R. Thompson, "Assessing the outcome of disorders of the hand," *Bone Joint J.*, vol. 83, no. 2, pp. 235–240, 2001.
[26] T. S. Caetano, S. D. Olabarriaga, and D. A. C. Barone, "Performance evaluation of single and multiple-Gaussian models for skin color modeling," in *Proc. 15th Brazilian Symp. IEEE Comput. Graph. Image Process.*, Oct. 2002, pp. 275–282.
[27] M.-G. Kim and K. Suzuki, "A card playing humanoid for understanding socio-emotional interaction," in *Proc. Int. Conf. Entertainment Comput.*, 2010, pp. 9–19.
[28] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognit.*, vol. 40, no. 3, pp. 1106–1122, Mar. 2007.
[29] C. Davatzikos and J. L. Prince, "Convexity analysis of active contour problems," *Image Vis. Comput.*, vol. 17, no. 1, pp. 27–36, 1999.
[30] R. L. Graham, "An efficient algorithm for determining the convex hull of a planar set," *Inf. Process. Lett.*, vol. 1, pp. 132–133, 1972.
[31] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2756–2759.
[32] J. Bouguet, "Pyramidal implementation of the affine lucas Kanade feature tracker description of the algorithm," *Intel Corp.*, vol. 1, no. 2, pp. 1–9, 2001.
[33] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2011.
[34] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, 1997.
[35] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. NIPS*, 2003, pp. 577–584.
[36] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Proc. CVPR*, Jun. 2009, pp. 983–990.
[37] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proc. Brit. Mach. Vis. Conf*, vol. 1. 2006, pp. 47–56.
[38] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1417–1424.
[39] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 234–247.
[40] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 23–26.
[41] A. Varfolomieiev and O. Lysenko, "An improved algorithm of median flow for visual object tracking and its implementation on arm platform," *J. Real-Time Image Process.*, vol. 11, no. 3, pp. 527–534, 2016.
[42] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 702–715.
[43] L. Zhang and P. N. Suganthan, "Robust visual tracking via co-trained kernelized correlation filters," *Pattern Recognit.*, vol. 69, pp. 82–93, Sep. 2017.
[44] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. ECCV Workshops*, 2014, pp. 254–265.
[45] V. Manohar, P. Soundararajan, H. Raju, D. Goldgof, R. Kasturi, and J. Garofolo, "Performance evaluation of object detection and tracking in video," in *Proc. 9th IEEE Int. Workshop PETS*, 2006, pp. 151–161.

EUNGYEOL SONG received the B.S. degree from the Department of Electronic and Electrical Engineering, Dankook University, in 2012. He is currently pursuing the Ph.D. degree in electrical and electronic engineering with Yonsei University, Seoul, South Korea. His current research interests include hand tracking, 3-D reconstruction, and machine learning.



HYEONGMIN LEE is currently pursuing the bachelor's degree in electrical and electronic engineering with Yonsei University, Seoul, South Korea. His current research interest includes video understanding, hand gesture recognition, and machine learning.



JAESUNG CHOI received bachelor's degree in electrical and electronic engineering with Yonsei University, Seoul, South Korea, in 2013, where he is currently pursuing the Ph.D. degree in electrical and electronic engineering. His current research interests include signal processing algorithm and machine learning based on the data from any sensors for human recognition.

SANGYOUN LEE (M'04) received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 1987 and 1989, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1999. He is currently a Professor and the Head of electrical and electronic engineering with the Graduate School and the Head of the Image and Video Pattern Recognition Laboratory, Yonsei University. His research interests include computer vision, with a special focus on pattern recognition for face detection and recognition, advanced driverassistance systems, and video codecs.