IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 May 2022, accepted 12 June 2022, date of publication 22 June 2022, date of current version 27 June 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3185402

A Novel Fusion Method With Thermal and RGB-D Sensor Data for Human Detection

AHMET OZCAN AND OMER CETIN
Hezârfen Aeronautics and Space Technologies Institute, National Defence University, 34149 Istanbul, Turkey
Corresponding author: Ahmet Ozcan (iletisim@ahmetozcan.com)

ABSTRACT Human detection methods are widely used in various fields such as autonomous vehicles, video surveillance, and rescue systems. To provide a more effective detection system, different types of sensor data (i.e. optics, thermal, and depth data) may be used together as hybrid information. Fortifying object detection, based on optical data and additional sensor data, such as depth and thermal data, also represents information regarding the distance and temperature of classified objects that can be used for video surveillance, rescue systems, and various applications. In this study, a simple and effective method is introduced to fuse RGB-D and thermal sensor data to achieve a more accurate form of human detection. To accurately combine the sensors, they are physically fixed to each other, and the relationship between them is determined using a novel method. The feature points on the optical and thermal images are extracted and matched successfully using computer vision. The proposed method is completely brand-free, easy to implement, and can be used in real-time applications. Using both thermal and optical data, humans are classified as benefiting from a widely used object detection method. The performance of the presented method is tested with a newly generated dataset. The proposed method boosts human detection accuracy by 5% when compared to the use of only optical data and by 37% when compared to the use of thermal data with COCO Dataset upon YOLOv4 neural network weights. After training with the newly generated dataset, the detection accuracy increases by 18% compared with the best results of single sensor usage.

INDEX TERMS Data fusion, human detection, image processing.

I. INTRODUCTION

Human detection is widely used in surveillance, rescue, and security applications. Nevertheless, some conditions, such as nighttime, bad weather, and the presence of physical obstacles may cause difficulties in practice. Varying light and other disruptive conditions may reduce image quality, which results in diminished detection accuracy, especially in outdoor applications. However, these variable conditions are deemed natural and frequently observed. In addition, disaster zones affected by earthquakes, floods, hurricanes, and fires cause rapid changes in conditions. Whereas security and rescue duties are crucial, being dependent on the environmental conditions can prevent systems from using detection information in autonomous applications.

Reinforcing methods can be leveraged to improve accuracy of human detection algorithms. For example, optical cameras may provide high accuracy under good lighting conditions; however, their accuracy is poor under low lighting conditions. Similarly, thermal cameras are quite successful when there is no obstacle emitting heat but work with limited accuracy in the presence of obstacles such as glass and foil. Therefore, the fusion of different sensors improves the results into a better detection accuracy in places observed during day and night, in varying lighting conditions, or when there are too many obstacles in the field. Taking advantage of multiple data sources containing different features of the environment, in lieu of solving a specific problem by employing data from a single source, is generally more effective. Detailed information such as dimensions, color, distance, and temperature of the surrounding objects may be obtained by various types of sensors in real-time. In numerous human rescue studies, different types of unmanned vehicles perform the pre-rescue human detection tasks using a single sensor [1], [2]. However, the accuracy of the classification algorithm can be boosted using different sensors simultaneously, as in [3]. Moreover, the use of extra sensors can provide various benefits, such as the detection of a person and determining his/her body temperature simultaneously. The distance information can localize the observer or object, and the color data can reveal the

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

distinguishing features. Having all of this data simultaneously can provide detailed information about the environment and can carry out a much more precise analysis of the mission. For example, the thermal and depth information of any desired object or region in the environment can be determined.

The first condition of an effective detection system is to combine multiple sensor data obtained from different sensors in real-time. To provide a robust method and offer a rich variety of features from various sensors, a symbiotic relationship needs to be defined among the sensors. Registered data acquired from this relationship should provide actual information to improve the decision-making process of an autonomous robotic system. The relationships among the sensors can be predefined in some cases, or they can be composed instantly in accordance with the problem. In both cases, combining data costs computation time, making it difficult to implement in real-time. Additionally, the fusion method should work without delay and be applicable without knowing the physical properties of the sensors such as focal point and angle of view. In other words, the method should be independent of sensor architectures, manufacturers, and models. Thus, it may be possible to use the method resiliently in different duties and areas such as security or rescue.

Using sensor fusion is not solely enough for efficient human detection in places where varying conditions are present. In addition to having detailed information about the environment, to be able to detect humans in real-time, one of the most crucial issues for detection systems is having an effective and fast detection algorithm. With the rapid developments in deep learning methods and the use of graphic processors on small-sized boards, it is now possible to classify objects even with limited computing capabilities. Some classification algorithms such as those in the study of Alparslan and Cetin [4] use a single neural network not to exhaust computing power, though they produce quite satisfying results. Furthermore, it is necessary to generate datasets to improve the performance of artificial neural networks in accordance with fused data. There are several human datasets such as [5]–[8] available. However, utilizing a dataset with challenging conditions and in conjunction with sensor fusion may boost detection accuracy.

In this research, to have a more effective human detection system using a real-time fusion, a new method establishing relationships between RGB-D and thermal sensors is proposed and the performance of the method is compared with that of using only a single sensor. The contribution of the study is a real-time fusion method, which is independent of sensor features. Furthermore, a generated dataset obtained by this method is presented.

In the second part of the study, the existing literature is reviewed, and challenges are explained. In the third part, the data fusion method is presented without using the features of the sensors such as focal point and angle of view. The performance of the proposed method is tested with an experimental study, including a newly generated dataset in the fourth section. In the last section of work, the results obtained in the study are discussed and the possible future research directions and applications are presented.

II. RELATED WORK

Using proper sensor fusion methods is one of the key issues in robotics studies [8]–[10] and unmanned vehicles [11]–[13] to carry out effective and robust missions. By using sensor fusion, an autonomous vehicle can perceive its surrounding environment better, by which it can localize itself in an unknown environment in more detail [10], [14], [15] and detect environmental changes faster [16], [9]. Moreover, it can produce more detailed data for targets like humans, animals, etc. [17]–[20]. The most recent studies to fuse sensor data are far from a dynamic fusion approach and are designed for specific sensors [21], [22]. In addition, graphic processors may be needed for fusion studies to be compatible with deep learning methods [23]. The fast and massive data collection capabilities of the sensors and representation of the obtained large data in the memory with different data types are the main problems of the real-time fusion process, especially in mobile robotic systems that are configured with limited computing power processors. To that end the relationship between the sensors should include dynamic features and provide an adaptive approach by being free of sensor brands, types, and producers.

Considering the above expectations, the robust and fastest approach to data fusion may be achieved by using image processing techniques [24]. Although optical, thermal, and depth sensors' data are in different forms, they can be modeled and processed as images. The process of combining optical images obtained from two identical sensors in the same environment is defined as the stereo vision in the literature [25], which is also a form of data fusion. In recent years, large-scale research has been carried out related to stereo vision. For example, Tippetts and colleagues have demonstrated a comprehensive review of stereo vision algorithms for systems with limited resources. They collected and presented accuracy and runtime performance data for all stereo vision algorithms developed over the past decade [26]. Although stereo vision is a popular and successfully implemented sensor fusion technique, it is limited to the usage of identical sensors together [27]–[30].

In order to satisfy changing needs and provide better decision-making systems, different features are registered by using different sensors together. Nevertheless, a small number of researchers have studied heterogeneous sensor fusion from the perspective of stereo vision. For example, Yang *et al.* have successfully provided a structure that can present a fusion of depth, thermal, and optical data, even though it cannot produce real-time results [31]. In another study by Chen *et al.*, data gathered from various sensors looking at a specific region for a specific purpose were combined [32]. Yet, their method is based on feature extraction, and it is not applicable to different sensors, environments, and purposes. Vidas *et al.* proposed another method for producing real-time sensor fusion results. However, resource consumption increases with

TABLE 1. Comparison of the fusion method with existing methods.

Reference	Used Sensors	Environment	The need for feature extraction points in runtime	Dataset Generation	Real-Time
Yang et al. [31]	Thermal, RGB	Outdoor, Indoor	Yes	No	No
Chen et al. [32]	Thermal, RGB-D	Indoor	Yes	No	N/A
Vidas et al. [33]	Thermal, RGB-D	Indoor	Yes	No	Yes
Cao et al. [34]	Thermal, RGB-D	Indoor	Yes	No	Yes
John et al. [39]	Thermal, RGB	Outdoor, Indoor	Yes	Yes	Yes
Ours	Thermal, RGB-D	Outdoor, Indoor	No	Yes	Yes

recurring operations, and the number of feature points that will be determined in the environment needs to be predefined [33]. Cao *et al.* fixed two different sensors to each other and combined them, yet the ICP method used in the study needs several feature points and includes complex calculations [34].

There are many robotic research studies where different types of sensors are used together. In the study by Correa *et al.*, people and their faces were detected by robots using thermal and optical sensors together in domestic environments [35]. In that study, the sensors had a similar field of view and depth of field. Thus, it was ensured that the fields of view of the thermal and optical images were close to each other. The main drawback of this study is the use of sensors without any data fusion process. In another study where optical and thermal sensors were used together, Carrio *et al.* developed an obstacle detection system for small UAVs [36]. The synchronized version of the data from the sensors was used in the study and it allows to work under extreme illumination conditions such as direct sun exposure and during nighttime. However, the fusion method was not explained in the study.

Combining different sensor data with image processing methods is one of the main problems in the literature and numerous studies have been conducted in this subject. In most of these studies, the combined images were taken from different angles of optical cameras with the same features [27]–[30], whereas some studies have also used cameras with different features [37]. It is observed that for the combining of thermal images and color images, methods involving complex calculations were generally used. In the work of Ben-Artzi *et al.*, a 2-points approach was proposed and the method outperformed 8-points and 7-points approaches, which are commonly used to equalize the epipolar geometries of different images [38]. This study shows that images can be combined by determining 2-points on the epipolar plane. Yet, the method is carried out with the same sensor at different angles. In the study by John *et al.*, thermal, and optic sensors were fused. Feature points were created using a source disseminating heat and light, and these points were overlapped to obtain registered sensor data. Moreover, a comparison of person detection accuracy was made [39].

Table 1 shows the differences and similarities between the proposed method and the fusion studies where RGB-D and thermal sensors are used. In the table, preferred sensors in the studies, work environment, real-time availability, and the technique for the presented fusion methods are given. Those methods benefit from feature extraction techniques to fuse the sensors. This process is required to gather feature points for every frame. It may cause big challenges in working in real-world environments since it is not possible to detect feature points in every situation. On the other hand, our study provides calibration with pre-calculated values for sensor fusion and does not need feature extraction points for every frame.

Detection of humans in images has been repeatedly carried out since computer vision studies were introduced. The latest instances that target better accuracy in real-time generally rely on neural networks and deep learning [40]. The authors in [41] studied human detection for small-sized UAVs with limited computer power benefiting from a YOLO detection algorithm. In [42], the authors proposed real-time detection on embedded platforms by using deep learning.

Numerous studies have been conducted to improve human detection accuracy. Xue *et al.* proposed a variation of YOLO called MAF-YOLO to improve the pedestrian detection performance at nighttime [43]. They used a multi-modal feature extraction module and modal weighted fusion instead of simply direct concatenation. In [44], the authors used an adaptive Region-of-Interest (ROI) for image scaling, which improves the detection accuracy and reduces the detection time. In [45], the authors proposed a regression-based approach for human detection using thermal images. To detect low-power human thermal signatures from a distant viewpoint, a fully connected network model was used and a much lower computational cost than other similar architectures was obtained. While these studies were performed considering different aspects, unexpected varying conditions harmed the detection capability. The simultaneous use of different sensors at the same time provided a significant advantage to mitigate the

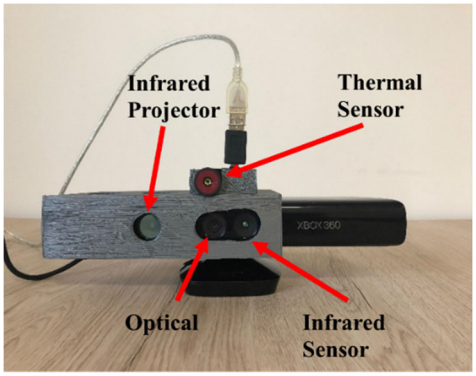

FIGURE 1. The appearance of sensors fixed to each other.

detection faults, because it is more resilient to changing factors.

III. SENSOR FUSION METHOD

It is valuable to use optical, thermal, and depth sensors together to extract many features of the environment and use them for real-time problems in computer vision. This study focuses on a novel method that provides a resilient combining approach that gathers data from various types of sensors in real-time. The process steps of the developed method within the scope of this study are explained step by step in this section.

It is possible to have many different brands and models of sensors for obtaining optical, thermal, and depth data and these sensors may have different features, such as frame rate, range, and image resolution value. However, the method in this study can be performed with any brand and model of the sensors.

In this study, Seek Compact Pro is used as a thermal sensor to detect thermal data in the environment and a Microsoft Kinect camera is used as an RGB-D sensor to obtain optical and depth data. These sensors are physically fixed to each other using an original component produced by a 3D printer and are shown in Fig. 1With a triple sensor platform, these parts can be moved together and work simultaneously. Using the Kinect sensor, with an image with a resolution of 640 x 480 pixels, including the distance data for each pixel, and a rate of 30 frames per second can be produced [46]. The Seek Compact Pro sensor can generate thermal data at a rate of 8 Hz [47] which can be represented as a visual snapshot of 320×240 pixels.

For real-time sensor fusion and to achieve rich detailed environmental data, it is necessary to establish relationships without the requirement of any special marks. In previous works, the design of the tools used for calibration was generally based on the selection of at least two special points that can be distinguished in thermal and optical data. Based on the positions of these points on the same plane and their relative positions, we aimed to determine the constant values to combine the images. Based on this idea, the calibration mechanism design created in this study consists of two black

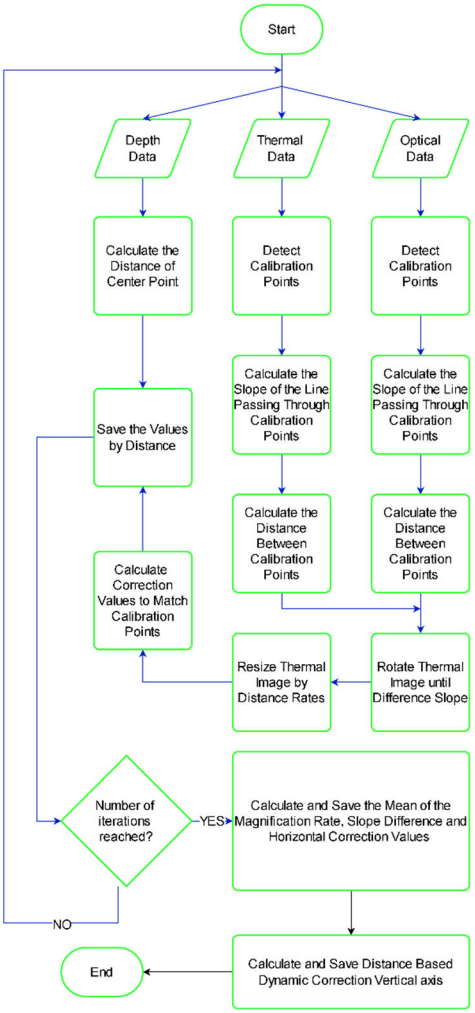

FIGURE 2. The determination of calibration values for the fusion of thermal and RGB-D sensor data.

circle drawings on a white background and incandescent bulbs that emit heat at the center of these circles. Thus, distinguishable common points in both the thermal and optical data are obtained. The flowchart in Fig. 2 shows the determination of calibration values. In order to determine the calibration values with the help of the calibration mechanism, the operations of detecting the circle center in the optical images and the center point of the heat sources in the thermal images are performed. These detected points are overlapped on the same plane to combine the data. During the calibration process, attention is given to whether there are heat sources of a similar density or similarly sized circular shapes in the field of view of the sensors.

The differences in distance and slope between the detected calibration points are calculated and compared to each other. For comparison, the magnification rate and slope difference values are calculated and recorded. After the thermal image is rotated to the point of the slope difference, magnification is applied as the distance ratio. Thus, both images have equal slope degrees and distances between the points. The horizontal and vertical correction values of the thermal image are

determined by finding the center points of the line segments passing through the points in images in order to exactly overlap the planes. In addition to the operations on the images, the distance from the calibration device is obtained from the depth data depending on the sensor features. All the computed calibration values are recorded based on the measured distances.

All of the steps need to be repeated several times for each distance value. All obtained calibration results are recorded with the distance values. The recorded results will be used for data fusion. When the results are examined, it is observed that the magnification ratio, slope difference, and horizontal correction values do not show significant changes for the different distance cases. Therefore, the mean of the relevant values is recorded as a single result for different distances. In contrast to these results, the vertical correction value varies significantly depending upon the distance. Distance values are determined in the optical images depending on the depth data. Appropriate calibration values are associated with this distance value. Thus, the thermal image is bound exactly to the optical image. The proposed method is described in detail in the following sections.

Calibration is performed based on the distance value of each pixel. The distance data corresponding to each pixel value acquired by the optical sensor on the Microsoft Kinect system is calculated using the Freenect library. The distance information is represented in memory as raw data with an 11-bits length value. The distance d from any desired pixel to the sensor in meters can be calculated with the help of the raw data r value received from the sensor and the k_1, k_2, $k_3$constants [43], as provided by (1).

$$\begin{aligned} d &= k_3 \times \tan\left(r \div k_2 + k_1\right) \\ k_1 &= 0.1236 k_2 = 2842.5 k_3 = 1.1863 \end{aligned} \tag{1}$$

After creating digital images with sensor data, the following operations are performed to find calibration points in the thermal and optical data. The Canny Edge detection method is used to determine the calibration points in the thermal image shown in Fig. 3(a). The points obtained from the thermal image are shown in Fig. 3(b). In the optical image shown in Fig. 3(d), circles containing the calibration points in the image are found using the Hough circle finding method and the centers of these circles are determined as calibration points, as shown in Fig. 3(e). Line segments are obtained and plotted on the thermal data, as illustrated in Fig. 3(c) and on the optical data as shown in Fig. 3(e). While the lengths of these two-line segments are determined by the Euclidean distance, the slopes of the lines between the points are calculated by the slope formula and stored to use in the combining process, as shown in Fig. 3(f).

The thermal image is rotated by as much as the difference in the computed slopes of the lines. Subsequently, the thermal image is resized by the ratio of the length of the lines. The midpoints of the lines and the distances between midpoints are determined in optical and thermal images to exactly overlap both images in the same plane. Thus, the position of the thermal image according to the optical image can be determined.

For the different scenarios and different distance values, the same steps are repeated for the desired number of times, and all the computed values are recorded as calibration data with the distance value. Magnification ratio, slope difference, and horizontal correction values are averaged from the obtained values. Vertical correction values vary significantly as the calibration distances change. The effect of the distance value on the vertical correction is modeled mathematically by examining the data from various distances and is represented by (2). By using this equation, the vertical correction value is prevented from undergoing a large change according to the distance. In (2), d_{max} represents the farthest calibration distance, and v_{max} shows the biggest value of vertical corrections.

Calibration processes are repeated five times for each 50 cm distance in the range of 50-250 cm, depending on the working range of the sensors. The obtained magnification and averages, the slope difference in radians and their averages, the horizontal corrections and averages in pixels, the vertical correction values in pixels, and the calculation time of the values in seconds are provided in Table 2 The computed values for the calibration processes performed at different distances are also shown.

$$f(x) = \begin{cases} v_{max} - \frac{d_{max}-d}{15}, & 100 < d < 250 \\ v_{max} - \frac{d_{max}-d}{15-(100-d)*0.4}, & 50 < d < 100 \end{cases} \tag{2}$$

Sensors' yaw and pitch differences on the plane cause the horizontal and vertical differences, whereas camera features constitute the magnification and slope differences. Looking at the values in Table 2, since the variation of values is small except for the vertical correction, the average values can be used in the method. The calculation times for the calibration process are also presented in Table 2. A calibration process consisting of five steps is approximately 1 s. Once the calibration process is completed, the mean values in Table 2 and (2) are used for the fusion of data from the sensors. This fusion process is executed in approximately 0.02 s. Thus, the acquired data from the sensors are combined with this process, causing a latency of 20 milliseconds, which allows the method to operate in real time.

IV. DATASET GENERATION

Since our method performs the fusion process while producing data, existing datasets containing optical, depth, and thermal images cannot be used to test our method. To verify and test the performance of the method and implement a realistic test scenario, an original dataset was prepared using RGB-D and thermal cameras. The sensors located on a fixed tripod are fused as mentioned in Section III. The view of the sensors is changed in horizontal and vertical axis to acquire data from different angles.

The dataset was generated in closed areas and outdoors, under different lighting conditions with 10 volunteers.

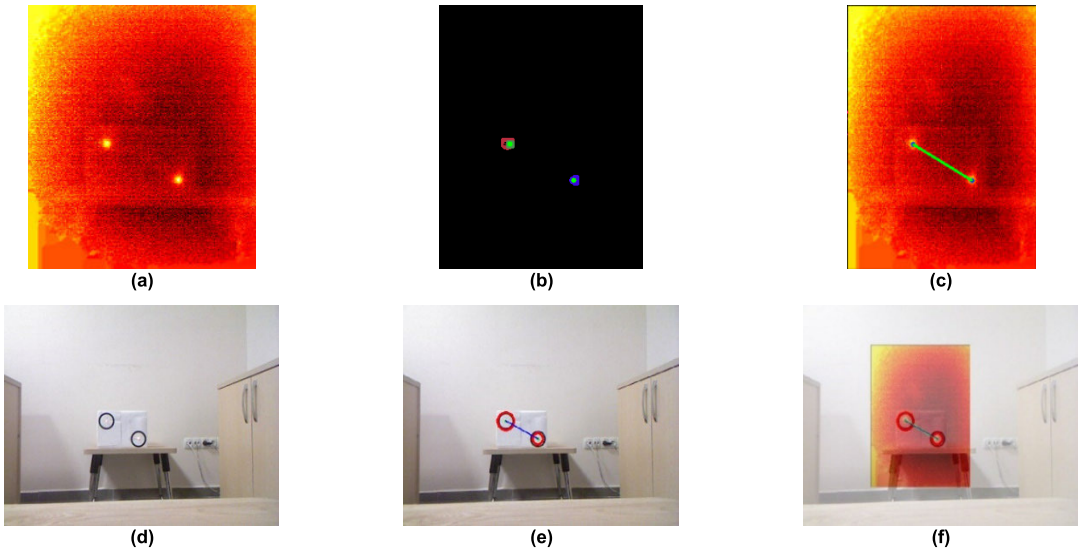

FIGURE 3. Calibration process for 150cm (a) thermal image (b) thermal image calibration points (c) thermal image calibration line (d) optical image (e) optical image calibration points and line (f) combined image.

TABLE 2. Calibration values by distances.

Description	Values [a]					
Calibration Distance	250	200	150	100	50	Mean
Magnification Rate	0.9564	0.9863	0.9721	0.9688	0.97680	0.9721132
Slope difference	-0.032	-0.059	-0.046	-0.074	-0.041	-0.050893
Horizontal Correction	158.91	154.38	156.38	156.42	158.95	157.0132566
Vertical Correction	95.591	84.809	85.058	83.326	62.719	
Calculation Time	0.1910	0.1959	0.1891	0.1850	0.1892	0.1900656

[a] Distance metric is in centimeters, slope difference metric is radian, point metrics are pixel and time metric are seconds.

TABLE 3. Gathered data by location.

Environment	Number of Frames
Remote Sensing Laboratory	134
Cafeteria	273
Camellia	519
Hallway	494
UAV Laboratory	709
Study room	250
Classroom	408
Distance Education Class	523
Total:	3310

3310 frames from each sensor were taken in eight different locations at 11 different times and the number of frames for each location is provided in Table 3. Instead of images, raw data acquired from the sensors was stored to avoid data loss and test the different combining conditions.

As shown in Fig. 4, under different lighting and physical conditions, different sensor data can be useful for detecting objects. In Fig. 4(a), the optical sensor cannot detect the human because of inadequate light, whereas the thermal sensor can easily classify it as human as in Fig. 4(b). On the other hand, in Fig. 4(g), the thermal sensor cannot detect the human because of a transparent physical obstacle, whereas this obstacle does not prevent the optical sensor from detecting the object. Depth data can be seen in Fig. 4(c) and 4(i) which show a display similar to the one in the optical image. The ROI of the images is shown in Fig. 4(d), Fig. 4(e), Fig. 4(f), Fig. 4(j), Fig.4(k), and Fig. 4(l).

A. PRODUCING OF REGISTERED IMAGES

After gathering data from the sensors, interest regions are assigned to overlap data from different sensors using the method described in Section III. Examples of ROI images produced from the raw data are shown in Fig. 5. To use in the training and testing processes, objects on the images are

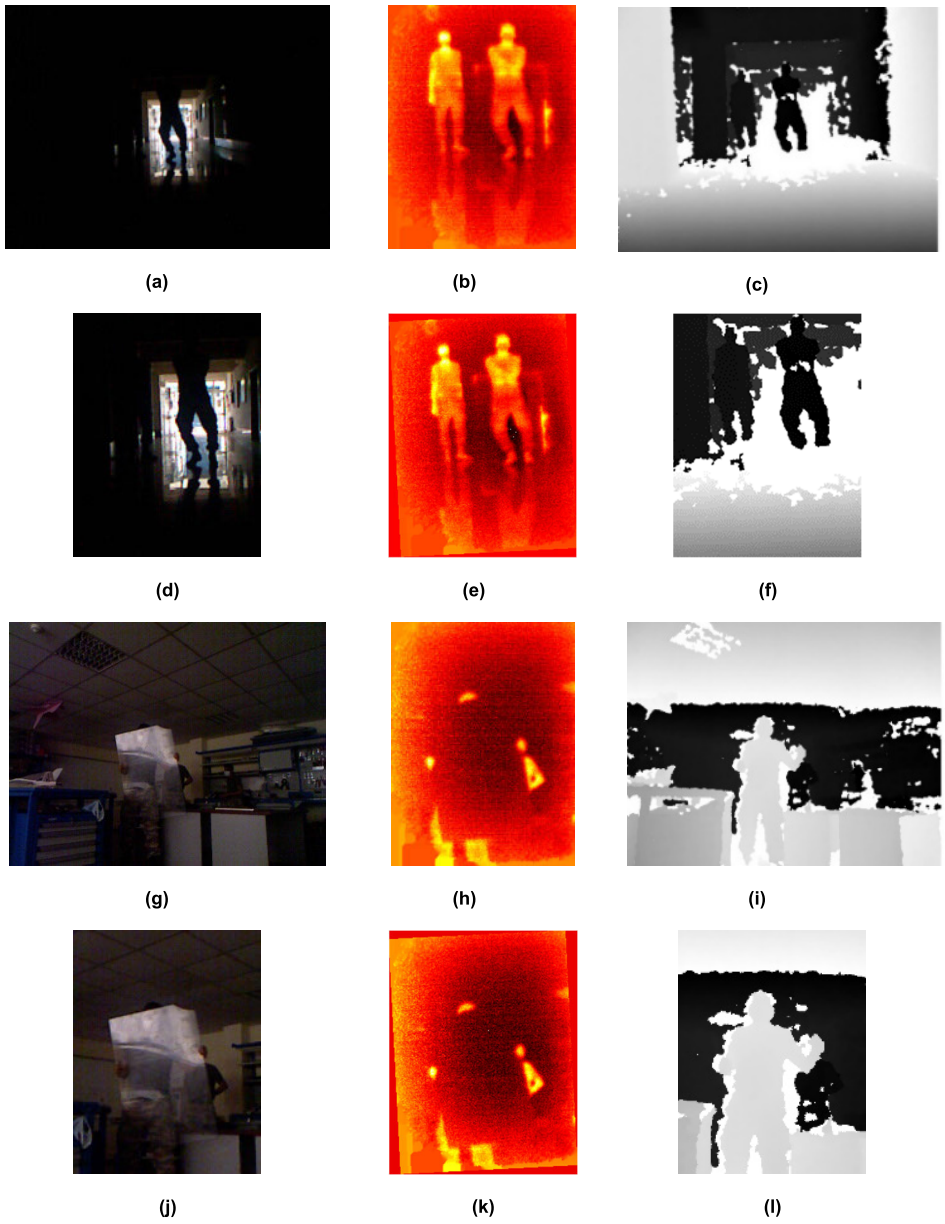

FIGURE 4. Sample data from sensors and their ROI images (a) optical image (b) thermal image (c) depth image (d) ROI of optical image (e) ROI of thermal image (f) ROI of depth image (g) optical image (h) thermal image (i) depth image (j) ROI of optical image (k) ROI of thermal image (l) ROI of depth image.

tagged with a "LabelImg" annotation tool [48]. The markings and boundaries of the objects are then cross-checked with thermal and optical images.

The pixel intensity values of the ROI images taken from the three sensors are unified in different proportions to test the object detection accuracy of the alternative methods. Eleven alternative unification and grayscale data are produced from three sensors. In the first unification, a three-channel (RGB) image is produced by taking the grayscale values of the images obtained from each of the three sensors where each sensor forms a channel. In the second method, a 3-channel image was created with thermal and color data by leaving the depth data channel values empty. For the other unifications, the thermal and optical images are merged using the OpenCV addWeighted method. For these unifications, the pixel intensity values obtained from both images are multiplied by the desired weights and added together to calculate the pixel intensity values of the new image. Using this method, the thermal image pixel values are added to the optical image pixel values by using 0.1 to 0.9 weights by 0.1 step size.

The raw data gathered from the three sensors are then stored in 26 different forms (11 colored unification, 11 grayscale unification, three raw sensor images, and one grayscale thermal image) to determine the best one for the

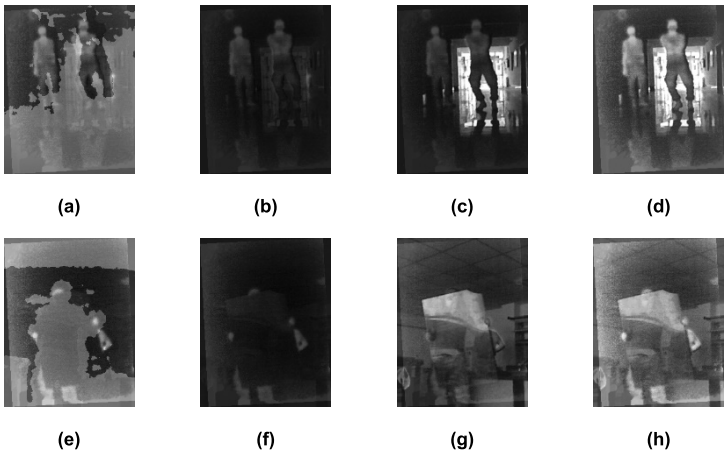

FIGURE 5. Some samples from different unifications (a),(e) 3 channel image from RGB, Depth and Thermal (b),(f) 3 channel image from RGB and Thermal (c),(g) RGB with 0.3 weighted Thermal (d),(h) RGB with 0.8 weighted thermal.

accurate detection of objects. Some samples from different unifications are shown in Fig. 5. In the samples, the improvement in unification is proved by the increase in the number of true detections of the objects.

V. PERFORMANCE EVALUATION

To test the performance of human detection based on sensor fusion, the accuracy of detection and classification are measured using a pre-trained convolutional neural network. Subsequently, the detection network is trained with the newly generated dataset. The detection performance after training and comparison with the pre-trained network are shown below.

A. COMPARING THE ACCURACY RATES OF REGISTERED IMAGES

After collecting data from various places and processing them in several ways, the accuracy rates of raw data and unified images are compared with each other. A pre-trained YOLOv4 [51] network on the COCO [50] dataset is used for comparison. Since the test environment targeted humans, only the ''person'' class is evaluated. The results are shown in Table 4 where the accuracy of each raw data and combining method are given in detail. The number of true, false and missed detections, average and mean average precision, average intersection over unit (IOU), precision, recall, and F1 score values are calculated. Under poor light conditions, the accuracy increased significantly for the RGB sensor data, whereas it varies under adequate lighting.

For all samples, the accuracy rate of the thermal data increased after unification. Even though all accuracy metrics are evaluated, the F1 score is the focal point because it can be used for model selection on datasets that are not evenly distributed. The optimal values are highlighted in bold in Table 4 Looking at F1 scores, the 0.1, 0.2 and 0.3 weighted unifications stand out. Among these values, the 0.3 weighted unification performs slightly better. Therefore, 0.3 is used in other tests as a basis.

Human detection in thermal and RGB data can be seen in Fig. 6 for two sample scenarios. Where it is possible to detect a human behind a transparent obstacle as shown in Fig. 6(a), this human cannot be detected using thermal data as seen in Fig. 6(c). In addition, while the second human in the field of view can be detected in optical data, it is misclassified in thermal data. Considering the combined data in Fig. 7(a), it is shown that the detection accuracy of the human behind the obstacles increases. Even if the detection accuracy for the second person in the field of view decreases when compared to optical data, the classification is correct. In the second scenario, the data obtained from a corridor with low light can be seen in Fig. 6(b) and Fig. 6(d). In both data types, 1 out of 3 people in the field of view cannot be detected. In the combined data given in Fig. 7(b), it is seen that all 3 people are detected more accurately compared to the data before the fusion.

According to the F-measure results, the accuracy rate of the thermal data increased by 37%, while optical data accuracy increased by only 5%. The metrics calculated using 3310 images show that unification methods significantly improve object detection accuracy. This improvement is more significant under poor light conditions, whereas the accuracy of the optical sensor data decreases substantially.

B. TRAINING AND TESTING OF NEWLY GENERATED DATASET

In this part, the accuracy of the newly generated dataset is tested by using best unification value. To train the dataset, the unified grayscale images with 0.3 weighted thermal data are divided into two. Eighty percent of the images are used for training, whereas twenty percent of the images are used for testing. Training is performed for 10,000 iterations for approximately 120 epochs. The test data is evaluated with the

TABLE 4. Detection results on dataset with different image types and unification methods.

Image Type / Metric	AP	TP	FP	FN	Average IoU	Recall	Precision	F1-Score
Depth	9.11	102	**52**	5608	0.42	0.01786	0.66234	0.03479
RGB	77.68	4145	1064	1565	0.66	0.72592	0.79574	0.75923
Thermal	42.39	1227	165	4483	**0.72**	0.21489	0.88147	0.34554
Thermal (Grayscale)	68.65	2541	439	3169	0.69	0.44501	0.85268	0.58481
RGB-Depth-Thermal	37.73	926	118	4784	0.68	0.16217	**0.88697**	0.27421
RGB-Depth-Thermal (Grayscale)	51.19	1804	358	3906	0.65	0.31594	0.83441	0.45833
RGB-Thermal	73.13	3118	550	2592	0.69	0.54606	0.85005	0.66496
RGB-Thermal (Grayscale)	84.63	4206	771	1504	0.69	0.73660	0.84509	0.78712
0.1*Thermal+RGB	83.14	4190	978	1520	0.65	0.73380	0.81076	0.77036
0.1*Thermal+RGB (Grayscale)	85.09	4331	937	1379	0.68	0.75849	0.82213	0.78903
0.2*Thermal+RGB	81.64	4075	881	1635	0.70	0.71366	0.82224	0.76411
0.2*Thermal+RGB (Grayscale)	86.14	4377	904	1333	0.68	0.76655	0.82882	0.79647
0.3*Thermal+RGB	80.94	3963	859	1747	0.68	0.69405	0.82186	0.75256
0.3*Thermal+RGB (Grayscale)	**86.22**	**4399**	909	**1311**	0.68	**0.77040**	0.82875	**0.79851**
0.4*Thermal+RGB	80.02	3881	781	1829	0.69	0.67968	0.83248	0.74836
0.4*Thermal+RGB (Grayscale)	86.08	4387	902	1326	0.68	0.76790	0.82946	0.79749
0.5*Thermal+RGB	78.86	3764	744	1946	0.69	0.65919	0.83496	0.73674
0.5*Thermal+RGB (Grayscale)	85.80	4343	878	1367	0.69	0.76060	0.83183	0.79462
0.6*Thermal+RGB	77.21	3621	662	2089	0.70	0.63415	0.84544	0.72471
0.6*Thermal+RGB (Grayscale)	85.40	4306	846	1404	0.69	0.75412	0.83579	0.79286
0.7*Thermal+RGB	75.34	3423	588	2287	0.70	0.59947	0.85340	0.70425
0.7*Thermal+RGB (Grayscale)	85.06	4270	1270	1440	0.69	0.74781	0.77076	0.75911
0.8*Thermal+RGB	83.60	3845	648	1865	0.70	0.67338	0.85578	0.75370
0.8*Thermal+RGB (Grayscale)	81.23	3609	542	2101	0.71	0.63205	0.86943	0.73197
0.9*Thermal+RGB	71.03	2881	498	2829	0.70	0.50455	0.85262	0.63395
0.9*Thermal+RGB (Grayscale)	84.13	4148	746	1562	0.70	0.72644	0.84757	0.78235

best weight values acquired from the training dataset and the performance metrics are given in Table 5.

The detection samples after training can be seen in Fig. 7(c) and Fig. 7(d). As shown in the figures, data fusion increases

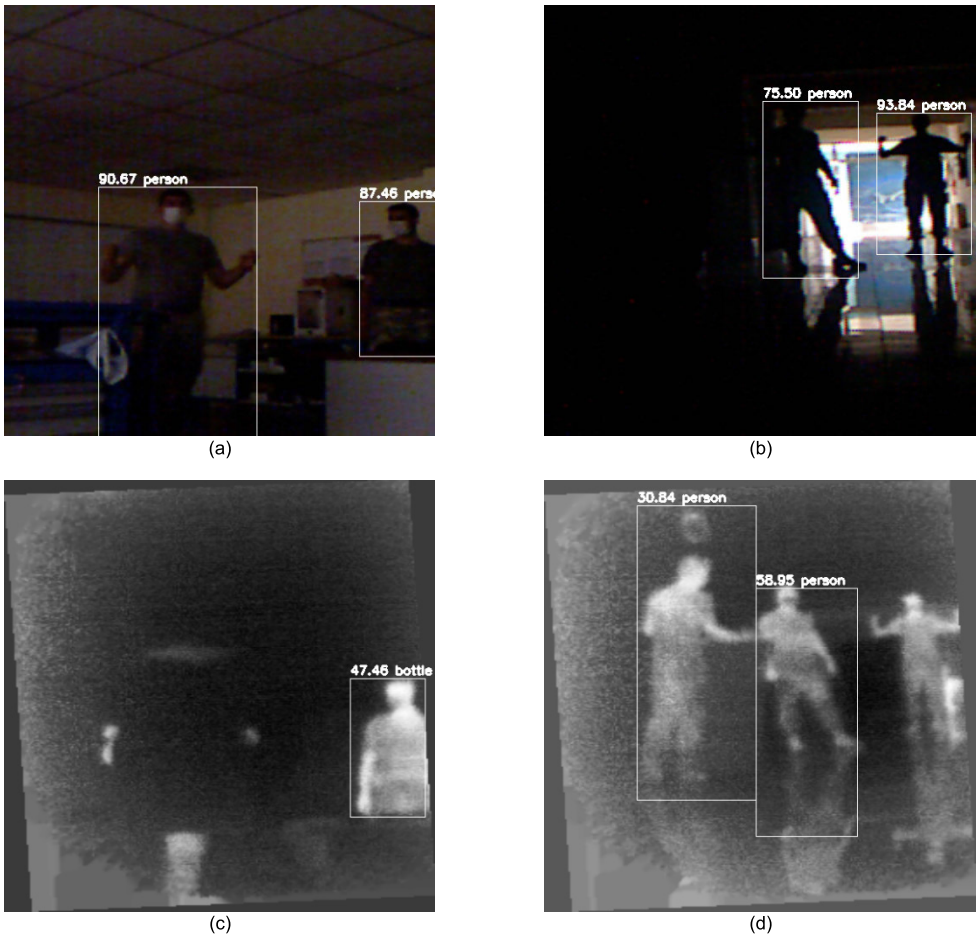

FIGURE 6. Detection results on sample images from the dataset. (a),(b) results on a RGB image, (c),(d) results on a Thermal image.

TABLE 5. Detection results after training of the newly generated dataset.

AP	TP	FP	FN	Average IoU	Recall	Precision	F1-Score
0.92	1120	98	37	0.73	0.97	0.92	0.94

TABLE 6. Performance enhancement by the fusion method after training of the newly generated dataset.

	Performance Increase by percentage according to		
	RGB Image	Thermal Image	Unified Image*
Average IoU	12 %	7 %	8 %
F1-score	24 %	61 %	18 %
Recall	33 %	118 %	26 %
Precision	16 %	8 %	11 %

*Unified image is acquired with the best unification method (Grayscale of 0.3 weighted Thermal and RGB Image).

TABLE 7. Performance comparison of the fusion method with existing studies.

Reference	Image Number in Dataset	Avg. Fusion Time per Frame (s)	F1-score
Yang et al. [31]	0	1.66	No Object Detection
Cao et al. [34]	0	0.02	No Object Detection
John et al. [39]	1640	0.06	0.899
Ours	3310	0.02	0.94

detection accuracy, while testing the challenging data in the newly generated dataset further increases the performance.

Performance metrics enhancement by the fusion method after training with the newly generated dataset is presented in Table 6. When the table is evaluated, it is observed that the

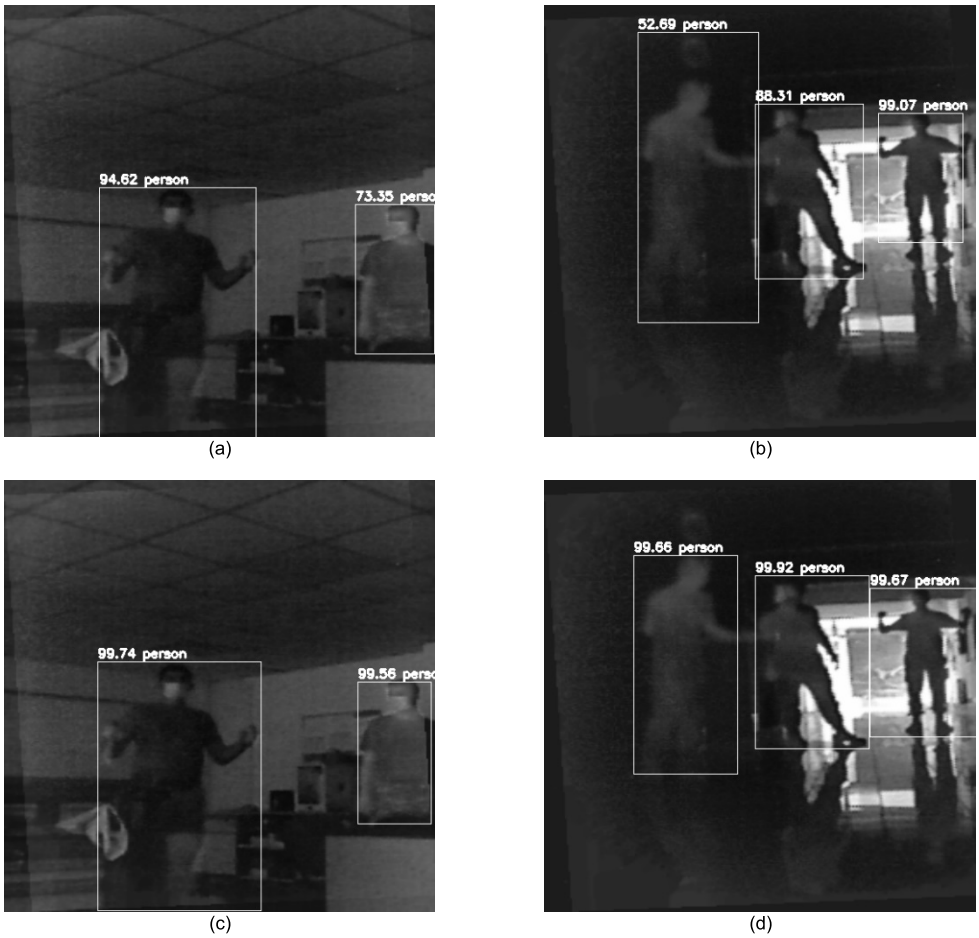

FIGURE 7. Detection results on sample images from the dataset. (a),(b) results on a combined image from RGB and Thermal data, (c),(d) results on combined a image from RGB and Thermal data after training of the newly generated dataset.

performance obtained from the unified data increases even more. Finally, based on the F1 scores, it can be said that the human detection accuracy of the optic sensor increased by 24%, while the detection accuracy of the thermal sensor increased by 61% after the sensor fusion process. The performance improvement is also compared with other existing fusion methods. As listed in Table 7, most studies show a dataset generation and object detection. In [39], 1640 frames were used for the study and the F1score is measured as 0.899, which are both smaller than our study.

VI. CONCLUSION

A simple and efficient sensor fusion approach, which uses thermal and RGB-D sensors together is proposed in this study to improve human detection accuracy under different environmental conditions. In order to benefit from the diverse features of multiple sensors, they are combined using an adaptive method to comply with varying environmental conditions. The proposed method defines the flexible fusion relationship between the thermal and RGB-D sensors. With this method, sensor data containing different features are successfully combined with each other in real time using limited processor capabilities. Therefore, it can be applied to real-time problems such as surveillance or rescue systems. It is also easily applicable to various types of sensors with unknown specifications because the method is free of the sensor model. Therefore, it can be used in missions in different environments or conditions.

Using a dataset that has challenging conditions, such as low-light and transparent obstacles, results in lower accuracy for human detection. However, sensor fusion increases the accuracy of detection when used properly. In this study, to demonstrate the success of hybrid data based on object detection methodology, a human is selected as a sample object, while it is also easy to implement the same method for different objects such as animals and vehicles. The newly generated dataset is prepared using the same method with different pixel intensity values. The success of the method is measured by well-known metrics such as precision, recall, and F1-score. The accuracy of the method is evaluated by the COCO dataset with YOLOv4 neural network weights. As a result, the accuracy rate of the thermal data increased by 37%, whereas the optical data detection accuracy increased by 5%. After training the neural network with the hybrid

dataset instead of using default weights, the accuracy of this network is also compared with that of the pre-trained network and sole sensor data. After training, when using a thermal sensor and an optical sensor together as a hybrid solution, the detection accuracy increased by 18% compared to using only one optical sensor, and by 61% compared to using only one thermal sensor. It is observed that the fusion of sensors increased the human detection accuracy, and this accuracy increased even more with the training of the newly generated hybrid dataset. The fusion method's performance is also compared with other fusion methods by fusion time and detection accuracy. The results show that our method provides better results.

Based on the results of this study, the fusion method can be used in real-time problems such as human surveillance and disaster rescue operations to obtain a more precise detection system. Because the proposed fusion method is free of sensor manufacturer and brand and requires low computational cost, it can be easily implemented in robotic systems such as UAVs to achieve higher accuracy rates. As a future study, to obtain better results for detection accuracy, a specialized network can be used in the detection method instead of YOLO, and more sensor data can be implemented to achieve different unique features. Furthermore, the number of sensors could be doubled to obtain a wider view of the environment. Detected humans may also be classified using segmentation methods by which the temperature and distance of the objects can be determined precisely. In addition, the success of the method can be tested in a disaster zone or surveillance location where mobile robots are operating.

REFERENCES

[1] R. Zheng, R. Yang, K. Lu, and S. Zhang, "A search and rescue system for maritime personnel in disaster carried on unmanned aerial vehicle," in *Proc. 18th Int. Symp. Distrib. Comput. Appl. Bus. Eng. Sci. (DCABES)*, Nov. 2019, pp. 43–47.

[2] J. Dong, K. Ota, and M. Dong, "UAV-based real-time survivor detection system in post-disaster search and rescue operations," *IEEE J. Miniaturization Air Space Syst.*, vol. 2, no. 4, pp. 209–219, Dec. 2021.

[3] M. Sneha, G. A. Aravindakshan, S. S. V. Vardini, D. A. Rajeshwari, V. A. R. R. Sastika, J. T. Selvi, and M. Sathiyanarayanan, "An effective drone surveillance system using thermal imaging," in *Proc. Int. Conf. Smart Technol. Comput., Electr. Electron. (ICSTCEE)*, Oct. 2020, pp. 477–482.

[4] Ö. Alparslan and Ö. Çetin, "Comparison of object detection and classification methods for mobile robots," *Sakarya Univ. J. Sci.*, vol. 25, no. 3, pp. 764–778, Jun. 2021.

[5] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.

[6] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3496–3504.

[7] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. M. López, "Pedestrian detection at day/night time with visible and FIR cameras: A comparison," *Sensors*, vol. 16, no. 6, p. 820, Jun. 2016.

[8] S. Tzafestas, "Sensor integration and fusion techniques in robotic applications," *J. Intell. Robotic Syst.*, vol. 43, no. 1, pp. 5–6, May 2005.

[9] M. S. Aman, M. A. Mahmud, H. Jiang, A. Abdelgawad, and K. Yelamarthi, "A sensor fusion methodology for obstacle avoidance robot," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (EIT)*, May 2016, pp. 458–463.

[10] Y. Xi, "Improved intelligent water droplet navigation method for mobile robot based on multi-sensor fusion," in *Proc. IEEE Int. Conf. Power, Intell. Comput. Syst. (ICPICS)*, Jul. 2019, pp. 417–420.

[11] K. Senthil, G. Kavitha, R. Subramanian, and G. Ramesh, "Visual and thermal image fusion for UAV based target tracking," in *MATLAB—A Ubiquitous Tool for the Practical Engineer*. London, U.K.: IntechOpen, Oct. 2011.

[12] B. Dai, Y. He, L. Yang, Y. Su, Y. Yue, and W. Xu, "SIMSF: A scale insensitive multi-sensor fusion framework for unmanned aerial vehicles based on graph optimization," *IEEE Access*, vol. 8, pp. 118273–118284, 2020.

[13] B. Cook and K. Cohen, "Multi-source sensor fusion for small unmanned aircraft systems using fuzzy logic," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2017, pp. 1–6.

[14] G. C. Nandi and D. Mitra, "Development of a sensor integration strategy for robotic application based on geometric optimization," *Proc. SPIE*, vol. 4385, pp. 282–291, Mar. 2001.

[15] Y. Dobrev, S. Flores, and M. Vossiek, "Multi-modal sensor fusion for indoor mobile robot pose estimation," in *Proc. IEEE/ION Position, Location Navigat. Symp. (PLANS)*, Apr. 2016, pp. 553–556.

[16] F. Matía and A. Jiménez, "Multisensor fusion: An autonomous mobile robot," *J. Intell. Robot. Syst.*, vol. 22, no. 2, pp. 129–141, Jun. 1998.

[17] W. Liu, J. Hu, and W. Wang, "A novel camera fusion method based on switching scheme and occlusion-aware object detection for real-time robotic grasping," *J. Intell. Robot. Syst.*, vol. 100, nos. 3–4, pp. 791–808, Jul. 2020.

[18] T. Dieterle, F. Particke, L. Patino-Studencki, and J. Thielecke, "Sensor data fusion of LiDAR with stereo RGB-D camera for object tracking," in *Proc. IEEE SENSORS*, Oct. 2017, pp. 1–3.

[19] Y. Zhang, X. Wei, and X. Zhou, "Dynamic obstacle avoidance based on multi-sensor fusion and *Q*-learning algorithm," in *Proc. IEEE 3rd Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, Mar. 2019, pp. 1569–1573.

[20] T. Sankey, J. Donager, J. McVay, and J. B. Sankey, "UAV LiDAR and hyperspectral fusion for forest monitoring in the southwestern USA," *Remote Sens. Environ.*, vol. 195, pp. 30–43, Jun. 2017.

[21] P. Saha and S. Mukhopadhyay, "Multispectral information fusion with reinforcement learning for object tracking in IoT edge devices," *IEEE Sensors J.*, vol. 20, no. 8, pp. 4333–4344, Apr. 2020.

[22] W. Kong, J. Hong, M. Jia, J. Yao, W. Cong, H. Hu, and H. Zhang, "YOLOv3-DPFIN: A dual-path feature fusion neural network for robust real-time sonar target detection," *IEEE Sensors J.*, vol. 20, no. 7, pp. 3745–3756, Apr. 2020.

[23] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, Jan. 2021, Art. no. 104042.

[24] D. Han and J. Huh, "Thermal data fusion for building insulation," in *Proc. Int. Conf. Syst. Sci. Eng. (ICSSE)*, Jul. 2019, pp. 368–371.

[25] W. Treible, P. Saponaro, S. Sorensen, A. Kolagunda, M. O'Neal, B. Phelan, K. Sherbondy, and C. Kambhamettu, "CATS: A color and thermal stereo benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2961–2969.

[26] B. Tippetts, D. J. Lee, K. Lillywhite, and J. Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems," *J. Real-Time Image Process.*, vol. 11, no. 1, pp. 5–25, Jan. 2013.

[27] K. Ambrosch, C. Zinner, and H. Leopold, "A miniature embedded stereo vision system for automotive applications," in *Proc. IEEE 26th Conv. Elect. Electron. Eng. Isr.*, Nov. 2010, pp. 786–789.

[28] R. Ben-Ari and N. Sochen, "Stereo matching with Mumford–Shah regularization and occlusion handling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2071–2084, Nov. 2010.

[29] M. Bleyer and M. Gelautz, "A layered stereo matching algorithm using image segmentation and global visibility constraints," *ISPRS J. Photogramm. Remote Sens.*, vol. 59, no. 3, pp. 128–150, May 2005.

[30] S. Gehrig, N. Schneider, R. Stalder, and U. Franke, "Stereo vision during adverse weather—Using priors to increase robustness in real-time stereo vision," *Image Vis. Comput.*, vol. 68, pp. 28–39, Dec. 2017.

[31] M.-D. Yang, T.-C. Su, and H.-Y. Lin, "Fusion of infrared thermal image and visible image for 3D thermal model reconstruction using smartphone sensors," *Sensors*, vol. 18, no. 7, p. 2003, Jun. 2018.

[32] X. Chen, G. Tian, J. Wu, C. Tang, and K. Li, "Feature-based registration for 3D eddy current pulsed thermography," *IEEE Sensors J.*, vol. 19, no. 16, pp. 6998–7004, Aug. 2019.

[33] S. Vidas, P. Moghadam, and S. Sridharan, "Real-time mobile 3D temperature mapping," *IEEE Sensors J.*, vol. 15, no. 2, pp. 1145–1152, Feb. 2015.

[34] Y. Cao, B. Xu, Z. Ye, J. Yang, Y. Cao, C.-L. Tisse, and X. Li, "Depth and thermal sensor fusion to enhance 3D thermographic reconstruction," *Opt. Exp.*, vol. 26, no. 7, p. 8179, Mar. 2018.
[35] M. Correa, G. Hermosilla, R. Verschae, and J. Ruiz-del-Solar, "Human detection and identification by robots using thermal and visual information in domestic environments," *J. Intell. Robot. Syst.*, vol. 66, nos. 1–2, pp. 223–243, Jul. 2011.
[36] A. Carrio, Y. Lin, S. Saripalli, and P. Campoy, "Obstacle detection system for small UAVs using ADS-B and thermal imaging," *J. Intell. Robot. Syst.*, vol. 88, nos. 2–4, pp. 583–595, Mar. 2017.
[37] Y. Bastanlar, A. Temizel, and Y. Yardimci, "Automatic point matching and robust fundamental matrix estimation for hybrid camera scenarios," in *Proc. IEEE 17th Signal Process. Commun. Appl. Conf.*, Apr. 2009, pp. 177–180.
[38] G. Ben-Artzi, T. Halperin, M. Werman, and S. Peleg, "Epipolar geometry based on line similarity," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1864–1869.
[39] V. John, S. Tsuchizawa, Z. Liu, and S. Mita, "Fusion of thermal and visible cameras for the application of pedestrian detection," *Signal, Image Video Process.*, vol. 11, no. 3, pp. 517–524, Oct. 2016.
[40] V. P. M. Gonealves, L. P. Silva, F. L. S. Nunes, J. E. Ferreira, and L. V. Araujo, "Evaluation of traditional and deep learning human detection techniques applied to surveillance: A performance comparison at distinct object sizes," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2021, pp. 1–5.
[41] K. Boudjit and N. Ramzan, "Human detection based on deep learning YOLO-v2 for real-time UAV applications," *J. Exp. Theor. Artif. Intell.*, vol. 34, no. 3, pp. 527–544, 2021.
[42] W. Rahmaniar and A. Hernawan, "Real-time human detection using deep learning on embedded platforms: A review," *J. Robot. Control*, vol. 2, no. 6, pp. 462–468, 2021.
[43] Y. Xue, Z. Ju, Y. Li, and W. Zhang, "MAF-YOLO: Multi-modal attention fusion based YOLO for pedestrian detection," *Infr. Phys. Technol.*, vol. 118, Nov. 2021, Art. no. 103906.
[44] B. C. Ko, M. Jeong, and J. Y. Nam, "Fast human detection for intelligent monitoring using surveillance visible sensors," *Sensors*, vol. 14, no. 11, pp. 21247–21257, Nov. 2014.
[45] A. Haider, F. Shaukat, and J. Mir, "Human detection in aerial thermal imaging using a fully convolutional regression network," *Infr. Phys. Technol.*, vol. 116, Aug. 2021, Art. no. 103796.
[46] V. Dolinay, L. Pivnickova, and V. Vasek, "Objectivization of traditional otoneurological examinations based on kinect sensor Hautant's test based on kinect," in *Proc. 15th Int. Carpathian Control Conf. (ICCC)*, May 2014, pp. 91–94.
[47] *Thermal Cameras for Your Smartphone—Seek Thermal | Affordable Infrared Thermal Imaging Cameras*," Accessed: May 3, 2022. [Online]. Available: https://www.thermal.com/compact-series.html
[48] *Imaging Information—OpenKinect*. Accessed: May 3, 2022. [Online]. Available: https://openkinect.org/wiki/Imagin_Information
[49] *GitHub—Tzutalin/LabelImg: LabelImg is a Graphical Image Annotation Tool and Label Object Bounding Boxes in Images*. Accessed: May 3, 2022. [Online]. Available: https://github.com/tzutalin/labelImg
[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 740–755.
[51] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

AHMET OZCAN received the B.Sc. and M.Sc. degrees in computer engineering from Erciyes University, Kayseri, Turkey, in 2011 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Hezârfen Aeronautics and Space Technologies Institute, National Defence University. His research interests include computer vision, machine learning, and robotic systems.

OMER CETIN received the B.Sc. degree in computer engineering from the Turkish Air Force Academy, Istanbul, Turkey, in 2003, and the M.Sc. degree in software engineering and the Ph.D. degree in computer engineering program from the Aeronautics and Space Technologies Institute (ASTIN), Istanbul, in 2008 and 2015, respectively. He is currently acting as an Assistant Professor with National Defence University (NDU), Turkey, where he is also currently acting as the Director of the Smart and Autonomous System Laboratory and besides lecturing and researching related with deep learning and autonomous systems.