Received: 17 March 2020 | Revised: 9 September 2020 | Accepted: 2 October 2020

DOI: 10.1049/ipr2.12062

IET Image Processing

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

A robust tracking algorithm for a human-following mobile robot

Tsung-Han Tsai | Chia-Hsiang Yao

Department of Electrical Engineering, National Central University, Taoyuan, Taiwan

Correspondence
Tsung-Han Tsai, Department of Electrical Engineering, National Central University, Taoyuan City, Taiwan.
Email: han@ee.ncu.edu.tw

Abstract

The capability of recognizing and tracking a specific human is considered a key technique for serving mobile robots. This paper proposes a real-time tracker that uses a stereo camera to track specific people in different environments. A tracker has been designed to detect the selected target from multiple people by use of block matching algorithm, human detection, and colour histogram comparison. Considering the similar colour of the objects, the predictor of the Kalman filter determines the position of the object. Finally, the mobile robot is moved according to the tracking results with a relative distance between the robot and the target. The effectiveness of the proposed method is demonstrated by experiments on many video sequences and real environments.

1 | INTRODUCTION

With high speed and complex developments in the industry infrastructure, the work environment has become more complex and dangerous. So, the robots are replacing humans for dangerous or difficult tasks, such as moving cargo in warehouses, helping humans fight fires and chemical disasters [1]. The research in the robotic area is becoming more and more vigorous. In recent years, robots are gradually merging into people's daily life from the industry. In daily life, mobile robots can help with picking up heavy objects in many places, such as a supermarket or hospital. The development of the human-following vehicles that are able to automatically follow the user has become more of a trend [2, 3].

There have been many papers in the past using various sensors to track specific people [4–6]. The main issues in human tracking are detection of the target, calculation of the distance between human and robot and following the target person. Different sensors will induce different design approaches. Recently, more and more papers can be found on vision-based object detection and tracking [7–9]. Mostly the related techniques are on object tracking. Object tracking is mainly used to track objects in image sequences. It is commonly used by object features such as texture, edge, optical flow, colour and so on. The results have been widely used in object identification, automatic monitoring, human-computer interaction, and other applications.

Target detection and tracking is a prerequisite in human tracking algorithms. To recognize and follow the target is the main task for the human-tracking robots. First, the human detection module is realized by computing a histogram of oriented gradient (HOG) features and classifying the HOG features by the support vector machine (SVM). Furthermore, to avoid the influence of the background colour, the background colour information is removed by a simple foreground extractor before calculating the colour histogram. Then the tracking target recognition is performed by comparing the colour histogram of the people which is detected by the human detection module.

Once the colour feature of the people in one frame is similar, a predictor based on the Kalman filter is used to estimate the target location. Although the human detection module based on the HOG and SVM classifiers has achieved a high detection accuracy, it spends too much execution time and is hard to implement on a real-time system. To make the system more effective, the block matching algorithm (BMA) that takes less execution time has cooperated.

As an integration concern, a video sensor is also the key point to construct a robot system. There are many three-dimensional (3-D) vision sensors in the market today. The 3-D sensing has more depth detection ability than the 2-D. It can capture the depth image of the object and obtain the spatial stereoscopic information.

This paper constructs a human-following mobile robot. It is designed based on the proposed high-performance tracking algorithm. We just use a binocular camera to track the target people in indoor and outdoor environments. To detect and track the target well, we apply many vision-based technologies to facilitate it. The tracking results and depth information of the target are combined to control the mobile robot. The results show that our proposed method can achieve a good performance in accuracy and execution speed.

The remainder of this paper is structured as follows. We provide some background works and analysed the pros and cons in Section 2. Section 3 then describe the proposed system. The experimental results based on the embedded system are shown in Section 4. Section 5 includes the conclusion.

2 | BACKGROUND

The main issue for human tracking is the distance of the tracking target and obstacle. Many studies have adopted various sensors to obtain the depth information to implement human-following applications, such as [10]: they combined the laser range finder location information with the visual features of the camera to achieve human tracking. However, there are still some methods that do not needed depth information, such as [11]. In [11], their mobile robot is equipped with a radio-frequency identification (RFID) sensor. Because the antenna provides the direction of the RFID source, the robot can track the person carrying the RFID tag. There are several important differences between these two kinds of studies. The advantage of the RFID method is that it has a relatively low cost, but users must wear special devices.

Moreover, the RFID method cannot be used in complex environments, especially when there are many obstacles.

To achieve the task by vision-based methods, the distance of the tracking target and obstacle with the depth sensor is needed. The depth camera is mainly divided into two kinds, one kind is to use an IR sensor to calculate the time of flight (TOF). The other is stereo vision, where the distance of the object is obtained by triangulation and other operations through two camera modules. TOF method, for using infrared, easily affected by illumination, which results in low accuracy in an outdoor environment. However, the stereo vision camera has a more complex operation and needs to consume more resources for computation.

In recent years, many research works on the human-following system are vision based. Verma et al. [12] presented to develop a robust vision-based tracking algorithm using speeded-up robust features (SURF). Pang et al. [13] presented a tracker which is using the kernelized correlation filter (KCF) to detect the target in multiple scales. It is designed with a combination of HOG, colour naming, and local depth pattern features. Chi et al. [14] have proposed a service robot based on the gait recognition method to perform the human-following task. Sun et al. [15] presented a human recognition method for the human-following robot based on soft biometrics. They combined two kinds of soft biometric traits that are clothes colour and body size as the features of the human. Furthermore, a particle filter has been used for human following in several years [16, 17]. Iswanto et al. [18] applied a mean-shift and particle-Kalman filter to perform the human tracking algorithm.

In addition, different from the above-mentioned traditional methods, the deep learning-based methods [19, 20] replaced the hand-crafted feature with the deep feature. With the development of the graphics processing unit (GPU) accelerator, these methods are becoming more and more popular. Chen et al. [21] have used a convolutional neural network (CNN) tracker for a person-following robot. Zhu et al. [22] designed a wheeled mobile robot system with a monocular pan-tilt camera to follow humans, where they exploited the Siamese network and optical flow information. Koide et al. [23] used the skeleton information and convolution channel feature to detect humans and identify the following target. Due to the prerequisite of training the model, the deep-learning-based methods can identify the specific people in advance and then pre-train the model. At the same time, they also must rely on a powerful GPU accelerator. As compared to them, the traditional algorithms are able to select the specific following person without any training state. Thus, it can be used in a more complex background and can perform without GPU acceleration on the embedded system. Table 1 shows the comparison of relative work specifications.

Object tracking plays a key important role in the vision-based human-following system. To detect and track moving objects in the image, many types of feature information such as shape, colour, speed or direction, etc. are used. There are many different algorithms for object tracking. Object tracking can be divided into three categories, including point, kernel, and silhouette, as shown in Figure 1 [24]. In [26], Fang et al. proposed a visual tracking method based on multi-cue information, they get a rough estimation of the target with optical flow, then adopted the part-based structure to optimize the result. In [25], Wang et al. also utilized the part-based structure on their tracker with the geometry constraint and attention selection, then propose a two-stage motion model. In [27], Kim and Park proposed a tracking scheme by combining the region projection of the region of interest (ROI) and probabilistic pixel classification, which is based on colour and intensity information of the neighboring regions. In [28], Alwar and Bajic constructed an efficient tracking engine by combining the hybrid tracker, which takes the motion information from the compressed video stream and a general-purpose semantic object detector. In [29], based on the tracker, which is separating an object from the background in a given image, Kim and Park proposed a novel target detection method to suppress the background clutter issue.

3 | PROPOSED ALGORITHM

Figure 2. shows the block diagram and the flowchart of the proposed visual human tracking system. The overview of the whole system is described in Section 3.1. Three key modules are discussed in Section 3.2 to Section 3.5 in detail. We explain how we apply the BMA to our system in Section 3.2, the optimizations we made on the HOG-based human detection method, and the reason for applications of the Kalman filter and colour histogram.

TABLE 1 Comparison of relative work specifications

	Method of measuring distance	Implement platform	Tracking algorithm
[10]	Laser sensors	CPU only	Face Detection + leg detection
[11]	Infrared sensor	N/A	RFID
[12]	Infrared sensor	CPU only	SURF
[13]	stereo matching	CPU only	KCF
[14]	Infrared sensor	N/A	Gait recognition
[15]	Infrared sensor	CPU only	Human Detection + Matching
[18]	Infrared sensor	N/A	Particle filter
[21]	stereo matching	CPU + GPU	CNN Tracker
[22]	N/A	CPU + GPU	CNN Tracker
[23]	N/A	Jetson TX2	CNN Tracker
Proposed work	stereo matching	Jetson TX1	BMA + Human Detection + Matching

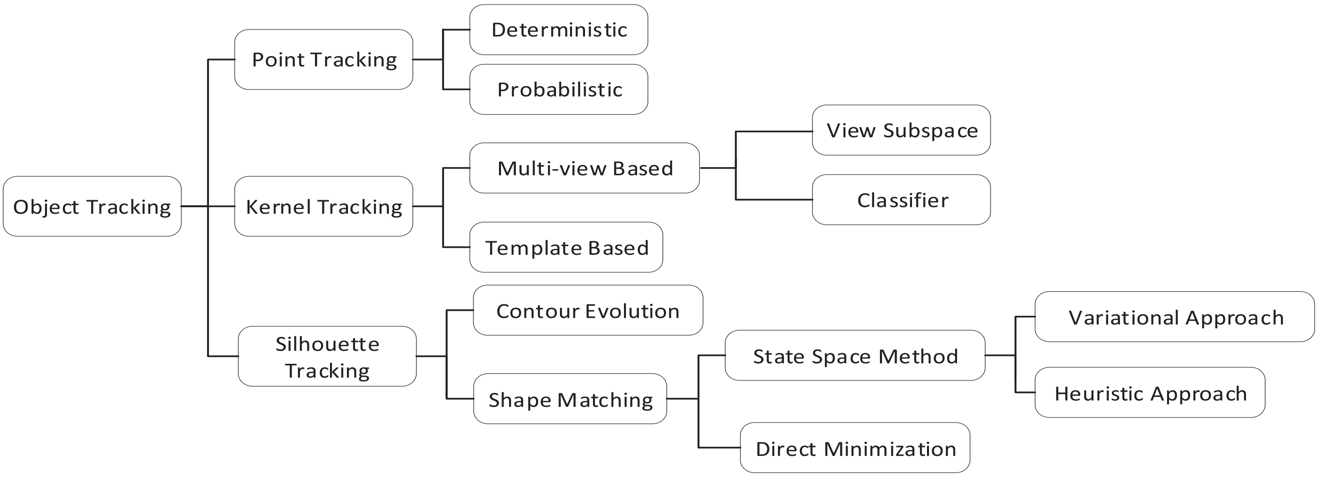


FIGURE 1 The categories of object tracking methods

3.1 | Whole system overview

In the initialization stage, we need to track several sequent frames of video to construct the target model by the selected target. Next, before performing the human detection, we will first perform block matching. If the target is found at this stage, we will save the execution time required for pedestrian detection. Once the target is not found, human detection will be initiated. Once the object is detected by the human detection module, the colour histogram information of each detected human is compared with the colour histogram information previously recorded by the target. The most similar object is considered as the tracking target. If the colour similarity between two peoples is too close, the location of the target is determined by a predictor based on the tracking result. Finally, we output the control signals to the motor control module according to the bounding box position and depth information of tracking results. The flowchart of motor control signals processing is shown in Figure 3 and the mechanism of movement decision is shown in Figure 4.

3.2 | Block matching algorithm

The BMA is applied to search for the best matching block from the reference frame. The full search algorithm is the most intuitive strategy, but a full search algorithm has a high calculation time. As a result, fast search algorithms have been developed, such as the three-step search (TSS), the new three-step search [29], four-step search, the diamond search (DS) [30], and the novel hexagon-based search [31]. An analysis of the fast BMAs is provided in [32] and [33]. The survey found that HEXS requires a minimum number of search points while the accuracy is still high. In [34], they present a cross-hexagon search (CHEXS) algorithm using two cross-shaped search patterns as the first two initial steps. This method can effectively save the search times and achieves almost the same PSNR value as other search algorithms.

In this module, we used colour histogram as the matching criterion for object tracking. The search method used in this system is the CHEXS algorithm. The CHEXS algorithm consists of two patterns: cross-based pattern and hexagon-based

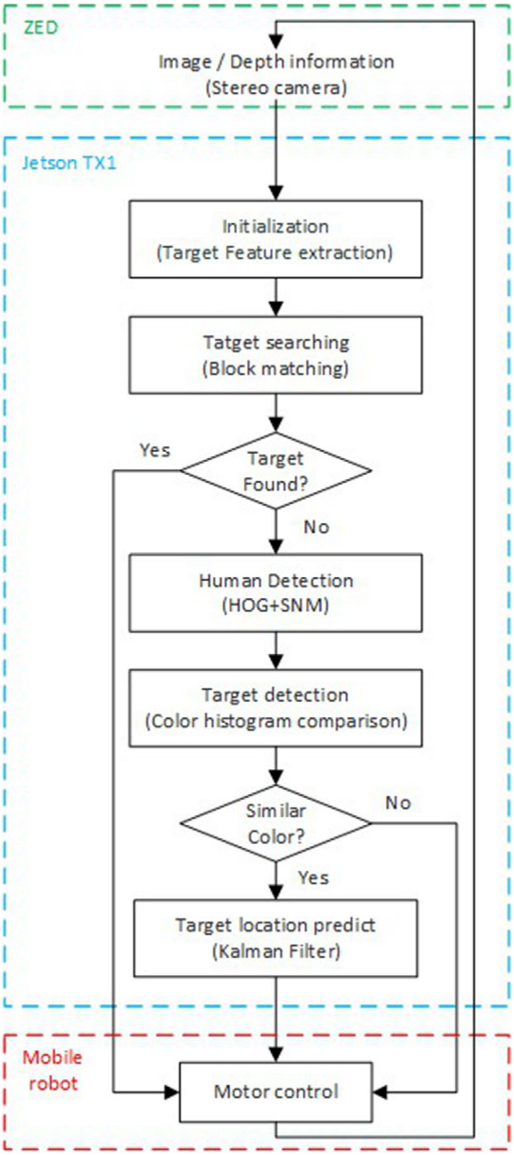


FIGURE 2 Flowchart of the human-following system

pattern, as shown in Figure 5. The steps of the CHEXS algorithm is divided into five steps.

Step 1 [Small-cross-shaped pattern (SCSP)]: Search for the most similar block in the five search points in the small cross mode. If the most similar block is at the centre of the small cross mode, then the search stops and the target is stationary. Otherwise, go to step 2.

Step 2 (SCSP): A new SCSP is formatted by the vertex in the first SCSP as the centre. If the most similar block occurs at the centre, then the search stops. Otherwise, go to step 3.

Step 3 [Large-hexagon-shaped pattern (LHSP)]: Search the big cross mode for three unsearched points. The most similar block is the centre of the next search.

Step 4 (LHSP): A new LHSP is formed by repositioning the most similar block found in the previous step as the cen-

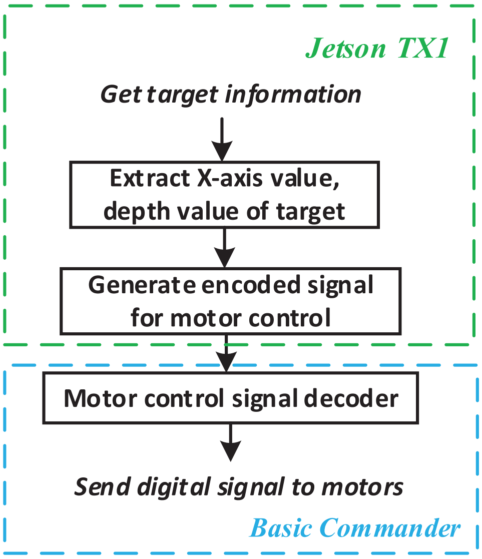


FIGURE 3 Flowchart of motor control signals processing

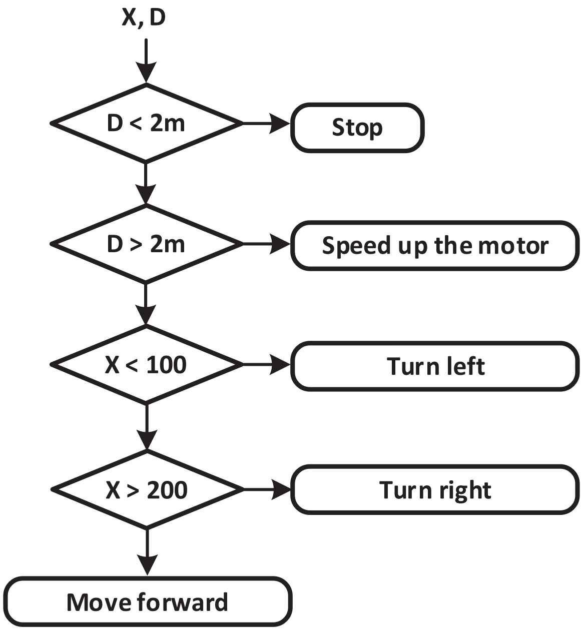


FIGURE 4 Mechanism of movement decision

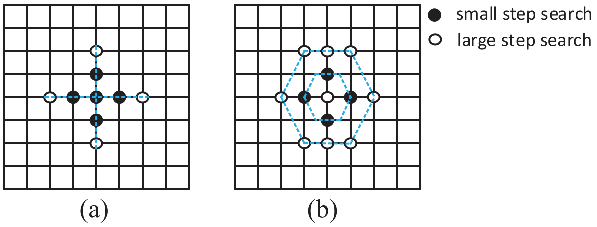


FIGURE 5 Search patterns in this module. (a) Cross-shaped pattern. (b). Hexagon-shaped pattern

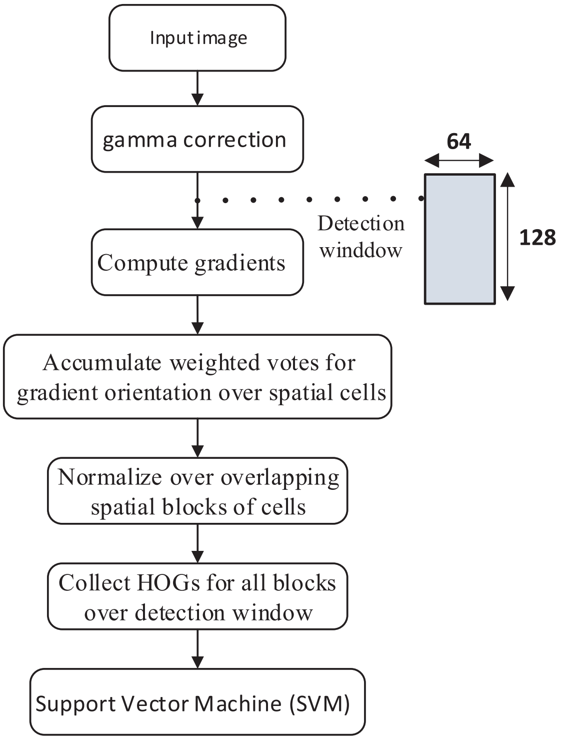

Input image
gamma correction
Compute gradients
Detection winddow
64
128
Accumulate weighted votes for gradient orientation over spatial cells
Normalize over overlapping spatial blocks of cells
Collect HOGs for all blocks over detection window
Support Vector Machine (SVM)

FIGURE 6 Flowchart of the human detection

tre of LHSP. If the most similar block occurs at the centre of the newly formed LHEXSP, then go to step 5. Otherwise, this step is repeated.

Step 5 (SHSP): An SHSP is formed by the results of step 4. A new most similar block is found from these five points, which is the target position.

In this module, we search for the target in our defined search areas through the CHEXS algorithm. BMA can take advantage of the faster computation time. However, it is more sensitive to object rotation and the changes of light, so if the target is not found with BMA, we will implement the human detection module to find the target. Through the cooperation of the two methods, we can focus the target promptly, and also the accuracy is still maintained.

3.3 | Human detection

The HOG feature extraction is widely used in embedded systems for the detection of objects such as pedestrians. It uses the intensity distribution of the gradient or the direction of the contours to describe the local appearance and shape of the object. In [35], Belloulata et al. proposed a pedestrian detection method with the combination of HOG features and SVM classifiers. They found that this method is able to maintain the balance between speed and effect. Therefore, we apply the HOG feature and the SVM classifier to detect the position of the people.

The flowchart of HOG feature extraction and detection is shown in Figure 6. First, the input image is converted to grayscale and standardize the image using gamma correction. The aim is to adjust the contrast of the image, reduce the influence of local shadow and illumination changes, and suppress the noise interference. Next, we extract the descriptor of the HOG feature from the detection window of the image. Finally, the SVM classifier is used to classify the people in the image through the descriptor of the HOG feature. The training dataset we used is the "INRIA person dataset [35]." There are many types of human posture and background in this dataset.

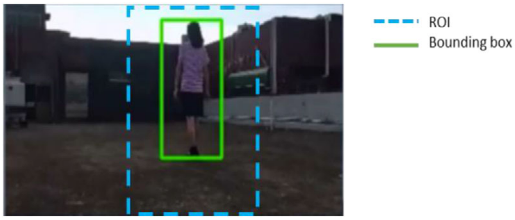


FIGURE 7 The region of interest (ROI) region

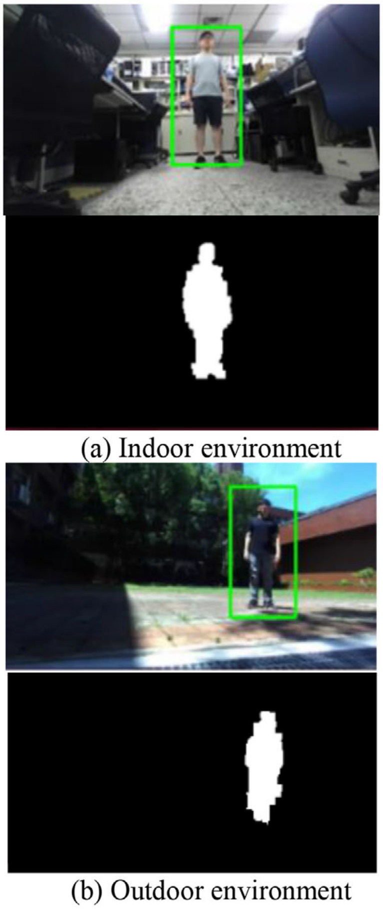
Since the location of the target between each frame is close, the position of the tracking target in the last frame can be conductive to predict the target on the current frame. It is not necessary to search the whole frame once the possible location of the target is identified. The computation time of human detection can be reduced by shrinking the ROI. This paper redefined the width and height of the ROI as shown in Figure 7. The time reduction depends on the bounding box size in the last frame, which is equal to $\frac{W_f-2*W_R}{W_f}$, where W_f is the width of the frame image and W_R is the width of the bounding box of the target in the last frame.

3.4 | Target detection from colour histogram

Colour histograms are the most widely used method for representing the distribution of the image. Their advantages include the fact that they are not affected by image rotation and translational changes. Further normalization can be used without affecting the image scale. The colour histogram can be built by any kind of colour space. Three-dimensional spaces such as RGB or HSV are usually used. Although most of the images are RGB colour space, the RGB spatial structure does not conform to people's subjective judgment of colour similarity. HSV is the most commonly used colour space. It offers improved perceptual uniformity. Furthermore, it is easier to compensate for many artifacts and colour distortions.

HSV colour space contains three components representing hue, saturation, and value. We use that to analyse the human detection results. In consideration of illumination variation, the colour histogram of the target is updated at intervals of a few frames.

After human detection, we will extract the colour features within each bounding box. Then we calculate the similarity between the candidate object and the target's colour information. To compute the similarity of two histograms H_1 and H_2, we have to select a metric $d(H_1, H_2)$ to indicate how close the


(a) Indoor environment

(b) Outdoor environment

FIGURE 8 Segmentation mask for target human. (a) Indoor environment. (b) Outdoor environment

two histograms match. The metric is given by

$$d\ (H_1, H_2) = \sqrt{1 - \frac{1}{\sqrt{\overline{H_1} \cdot \overline{H_2} \cdot N^2}} \sum_I \sqrt{H_1\,(I) \cdot H_2\,(I)}} \qquad (1)$$

where

$$\overline{H_k} = \frac{1}{N} \sum_J H_k\,(J) \qquad (2)$$

and N is the total number of histogram bins.

Figure 8 shows the bounding box of human detection. It includes the background colour information and, thus, results in the worse accuracy of the colour histogram comparison. To solve this, we propose a simple foreground extractor with a stereo camera. We apply the depth result to mark the pixels in the bounding box with a depth difference greater than 1 m from the centre as background and other pixels as unknown. The unknown pixels belong to the background which is decided by the neighbouring pixels, which have been marked as a background. To extract the foreground precisely, there are lots of existing methods that are performing well, such as ViBe and frame difference. Because the system is designed for execution in real-time, and the computational complexity of ViBe is low, so we chose ViBe [36] as our foreground extraction method for our system.

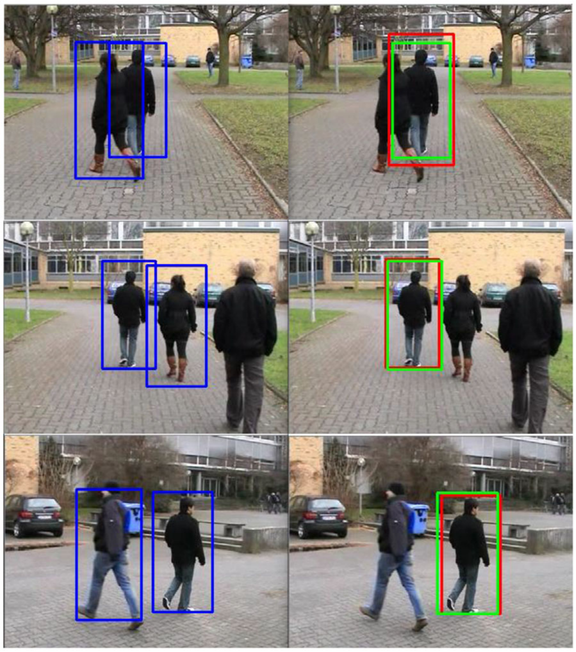


FIGURE 9 Schematic diagram of the Kalman filter

To distinguish the background, we calculate the Euclidean distance between the pixel value and each random background model sample 20 times. We collect the samples with Euclidean distance less than the threshold R and count the number of it as N. When n is greater than the threshold we defined (the threshold is usually set to 20), then the x point is considered to be the background point. Specifically, we represent the pixel value at the x point as $V\,(x)$. M $(x) = \{V_1, V_2, \ldots, V_n\}$, which is the background sample set at x (the sample set size is N), where n is the total number of the samples which is set to 20. The region with x as the centre and R as the radius is expressed as $S_R(V\,(x))$. If M$(x)[S_R \cap \{V_1, V_2, \ldots, V_n\}]$ is greater than the threshold, then x is considered to be the background point.

3.5 | Kalman filter

Since we only use colour features to find the target, it is easy to make mistakes on the other similar targets. To address this problem, the target location is extracted by a predictor based on the Kalman filter which predicts the position based on the tracking result. First, we get the position of the target that is predicted in this frame through the Kalman filter. Then we calculated the bounding box overlap rate of these candidates and the bounding box predicted with the Kalman filter. Finally, the object with the highest overlap rate is the target that is going to be tracked.

As shown in Figure 9, the left is the result of pedestrian detection, the green box on the right is the tracking result, and the red box is the result predicted by the Kalman filter. When the colour

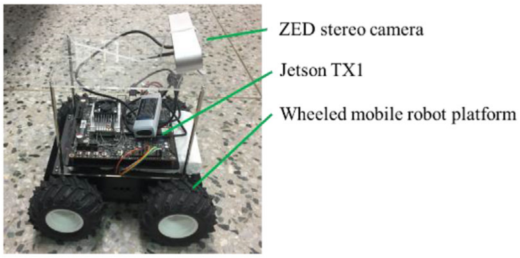


FIGURE 10 The human-following mobile robot used for implementing the proposed human tracker

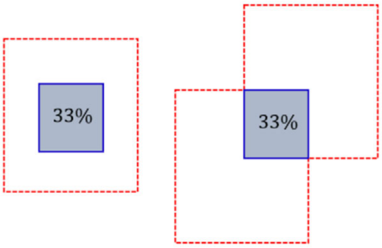


FIGURE 11 A schematic diagram of how the overlap for bounding boxes is calculated in the BoBoT dataset

features are very close, we will compare the overlap rate of the pedestrian detection box with the red box.

4 | EXPERIMENTAL RESULTS

The experimental setup used for system implementation is shown in Figure 10. We validated our proposed tracking algorithm on the BOBOT dataset [37] and showed the simulation result of the human-following system in the real-world environment. The mobile robot is equipped with a Jetson TX1 (Quad ARM A57/2 MB L2), a ZED stereo camera, and the wheeled mobile robot platform. We choose TX1 as our experiment platform because we can get the target's depth information through the ZED stereo cameras. ZED stereo camera as the vision sensor is used and the robot is kept at a fixed distance from the target through depth measurement. However, CUDA support is necessary for ZED SDK, so TX1 is selected as it can support CUDA. Although the proposed work is using CUDA, the algorithm does not need any GPU acceleration. It can be executed at 20 fps on raspberry-pi4 and also maintain a high performance of accuracy.

Next, we evaluate the tracking algorithms using the public benchmark. The benchmark we used is Bonn benchmark on tracking (BoBoT). The BoBoT is a benchmark for testing and comparing different properties of visual tracking algorithms. Ground truth annotations are also given for each video sequence. They correspond to the coordinates of the target's bounding box and its size. As shown in Figure 11, this benchmark's criterion for successfully tracking the target is more than one-third overlap with the bounding box of ground truth. The effectiveness of the proposed system is verified by two evaluation values: precision and recall.

TABLE 2 The tracking performance results with different BMA methods in a different similarity threshold value

Methods	BMA threshold	FPS	Recall	Precision
CHEXS	0.6	31.4	0.90	0.94
	0.7	22.6	0.85	0.98
	0.8	18.1	0.77	0.96
	0.9	18.3	0.78	0.98
DS	0.6	30.4	0.76	0.81
	0.7	22.1	0.85	0.96
	0.8	17.3	0.76	0.81
	0.9	17.2	0.81	0.98
TSS	0.6	31.5	0.81	0.98
	0.7	23.2	0.84	0.95
	0.8	18.6	0.78	0.98
	0.9	18.6	0.78	0.98
Without BMA	N/A	15.9	0.78	0.97

4.1 | Experimental results with joining BMA

To reduce the execution time of tracking, the defined search window is found by block matching before human detection. Through the search of adjacent areas, the target can be successfully detected when the HOG feature is not obvious enough so that the target is not detected as a pedestrian. Meanwhile, compared with human detection, BMA takes a significantly less execution time, which is very important for the tracking algorithm that requires high execution speed. In this section, we will show that adding BMA can effectively reduce the execution time of tracking and the accuracy will not reduce significantly.

Some famous fast algorithms on BMA are DS and TSS. We evaluate these three methods used in our system. According to the experimental results and the analysis, it makes the same except that CHEXS can take fewer search times since our system is designed with low execution time.

In addition, different BMA similarity threshold had different effects on the system. For example, if the threshold is too small, the system can easily enter the BMA module to get lower accuracy. On the contrary, it will cause the system hard to enter the BMA module and cause high computation time. Here, we evaluate it on a different threshold on the BOBOT dataset. As Table. 2 shows, as the system is processed without going to the BMA method, the execution speed is 15.9 fps. When 0.6 is used as the similarity threshold, the execution speed is reached to 31.4 fps on TX1. It means the execution speed is greatly increased. The experimental results also show that the precision and recall values are higher than 90%. In addition, according to the algorithmic complexity theory, the algorithmic complexity of the

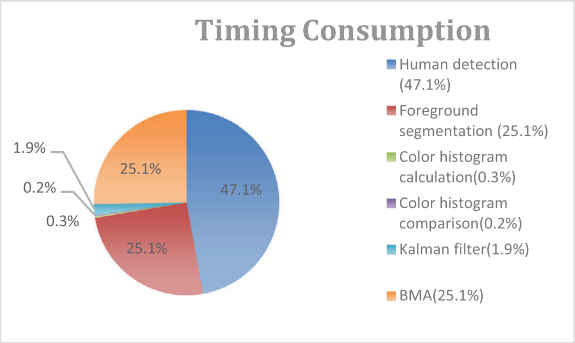


FIGURE 12 Profiling of the complexity of the whole system

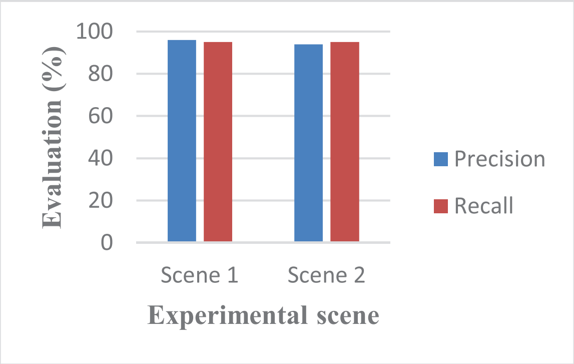


FIGURE 13 Evaluation of results in indoor experiments

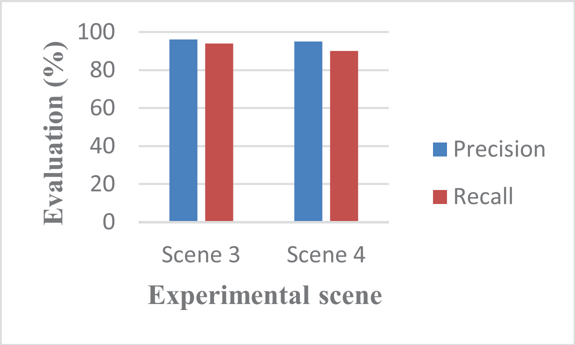


FIGURE 14 Evaluation of results in outdoor experiments

TABLE 3 Summary of attributes for the dataset used for evaluating the performance of the proposed algorithm

	Seq. D	*Seq. E*	*Seq. F*	*Seq. I*	*Seq. J*
Total number of Frames	947	305	453	1017	388
Tp	943	292	384	971	351
Tn	0	0	35	12	5
Fp	4	4	2	14	15
Fn	0	9	32	20	17
Precision	99.5 %	98.6 %	99.4 %	98.5 %	95.9 %
Recall	100 %	97.0 %	92.3 %	97.9 %	95.3 %

HOG descriptor is O(n^2). Our proposed approach is to effectively reduce the overall complexity of the system by adding a block matching method which is O(n) complexity. In order to prove it, we also analyze the time consumption of each module, as shown in Figure 12. Referring to the fast search algorithm, it obviously can help to reduce the block search times compared with the exhausting search on the full search method. Now, this module occupies about 25% complexity of the whole system.

4.2 | Target tracking experiment in an indoor and outdoor environment

We test the tracking algorithm through some videos. There are four test films in total, two for indoor and two for outdoor environments. The indoor environment video sequence is mainly shot in the corridor, and the outdoor environment video sequence is mainly shot in the campus. Through a more comprehensive evaluation of this system, it can be verified that the system has good performance in indoor and outdoor scenarios.

The results of the target tracking experiment in an indoor environment are shown in Figure 13. The accuracy and recall values of the experimental results are both good. The results of the target tracking experiments in an outdoor environment are shown in Figure 14. High precision values are available with each recall value above 90%. These experimental results show that the robot can track the target in both indoor and outdoor environments.

We evaluate our tracking algorithm by five video sequences in BoBoT (*Seq. D* to *J*). Table 3 shows the result of the evaluation. The main challenges of the *Seq. D* are moving camera and rotation. The main challenges of the *Seq. E* and *Seq. F* are moving camera and partial occlusion.

The main challenges of the *Seq. I* are that the sequence is captured in a moving camera, with moving target, rotation, similar distractors, full occlusion, and outdoor environment. There are more complex light variations in the outdoor environment, which is a big challenge for tracking algorithm. There are also a lot of obscurations in this video sequence, and some people have clothing colours that are close to the target. *Seq. I* can effectively evaluate the tracking effect in the case of occlusion and similar distractors. Table 4 shows the tracking results when the person is occluding and clothes are similar in colour. The main challenges of *Seq. J* are moving camera, moving target, and full occlusion. This video sequence shows two people walking in an indoor environment and two people occluding each other multiple times. Occlusion has always been a big test in tracking algorithms. *Seq. J* can effectively evaluate the tracking effect in the case of occlusion. Table 5 shows the tracking results when the person is occluding.

In Table 6, the average performance for each sequence is presented for several methods. We also compare them with the results of the proposed method. In [16], they proposed a

TABLE 4 The tracking results when the person is occluding and clothes are similar in colour

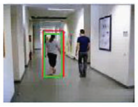
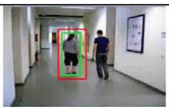
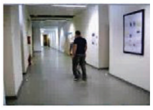
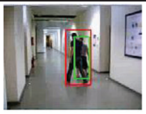
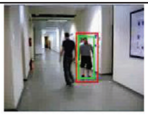
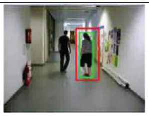
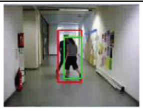
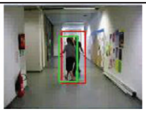
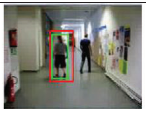
Frame.190	Frame.210	Frame.250
		
Frame.270	Frame.290	Frame.550
		
Frame.570	Frame.590	Frame.630
		

TABLE 5 The tracking results when the person is occluding and clothes are similar in colour

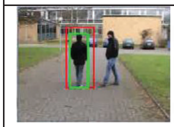
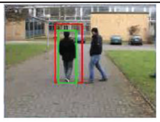
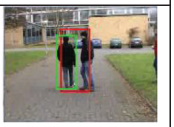
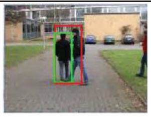
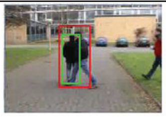
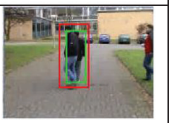
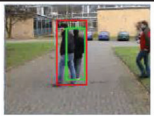
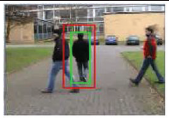
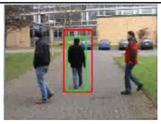
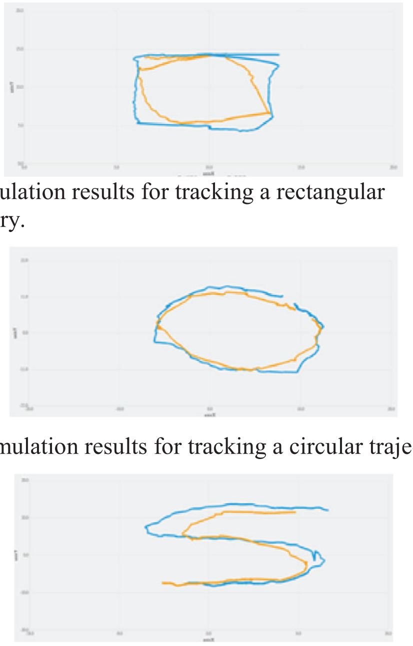
Frame.1007	Frame.1014	Frame.1028
		
Frame.1035	Frame.1042	Frame.1056
		
Frame.1063	Frame.1070	Frame.1084
		

TABLE 6 Results on BoBot dataset with different approaches

		accuracy [%]			
Seq.	**[16]**	**[38]**	**[18]**	**[39]**	**Proposed method**
D	85.9	98.0	N/A	81.1	99.6
E	67.9	99.0	N/A	93.8	95.7
F	51.5	52.0	91.0	67.2	92.4
I	33.4	80.0	95.25	67.8	96.7
J	N/A	73.0	100	N/A	91.8



(a) Simulation results for tracking a rectangular trajectory.

(b) Simulation results for tracking a circular trajectory.

(c)Simulation results for tracking a s-shaped trajectory.

FIGURE 15 The simulation result for tracking trajectory. (a) Simulation results for tracking a rectangular trajectory. (b) Simulation results for tracking a circular trajectory. (c) Simulation results for tracking an s-shaped trajectory

tracking framework by particle filter based on the condensation algorithm and built a reliable appearance model for the problem of object tracking in videos. In [18], they presented a tracking algorithm that combines the mean-shift and particle-Kalman filter. In [38], they introduce an approach that makes human detection by the fast version of an implicit shape model detector, which is trained on people. In [39], they proposed the tracker based on the point matching method; the point-based features will be extracted first and be corrected by the DBScan algorithm, and then construct an object model as the tracker. With the same video test sequences, the proposed design takes advantages of high accuracy.

4.3 | Simulation results in real environment condition

In addition to evaluate the tracking algorithm, we test three paths and recorded the trajectory of the human-following mobile robot. One commercial product used an ultra-wideband (UWB) sensor for the applications of positioning and navigation of moving objects. It is because the advantages of UWB are strong penetrating power, low power consumption, high security, and low system complexity. We use this technology to track and locate the location of people and the robot as the benchmark. We perform three paths: rectangular, circular and s-shaped. We take the tracking trajectory as shown in Figure 15. Meanwhile, the blue line is the result of a pedestrian and the origin line is the result of our motor tracker.

5 | CONCLUSION

We develop a robot that can follow a specific person using the results of tracking through the motor control module. First, we use the depth sensor to obtain the distance of tracking targets and obstacles. We use the BMA to search the target. This effectively solves the slow execution speed of human detection. Through the experimental results, it was verified that the system could track target people no matter in the indoor and outdoor environments. It shows that the computation complexity of the whole system has been reduced; however, the precision and recall values are higher than 85%. We further implement it on Raspberry-pi4 for real demonstration. Our system can still execute at 20 FPS and also maintain the high performance of accuracy.

REFERENCES

1. Graham, J., Shillcutt, K.: Robot tracking of human subjects in field environments. In: Proceedings of the International Symposium on Artificial Intelligence, Robotics, and Automation in. Space, pp. 1–7, (2003)
2. Dang, Q.K., Suh, Y.S.: Human-following robot using infrared camera. In: 2011 11th International Conference on Control, Automation and Systems, pp. 1054–1058 (2011)
3. Yoshimi, T., et al.: Development of a person following robot with vision based target detection. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, pp. 5286–5291 (2006)
4. Lee, J., et al.: People tracking using a robot in motion with laser range finder. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2936–2942 (2006)
5. Mattos, L., Grant, E.: Passive sonar applications: Target tracking and navigation of an autonomous robot. In: IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, vol. 5, pp. 4265–4270 (2004)
6. Feng, G., et al.: A compressed infrared motion sensing system for human-following robots. In 11th IEEE International Conference on Control & Automation (ICCA), pp. 649–654 (2014)
7. Hu, C., et al.: A robust person tracking and following approach for mobile robot. In: 2007 International Conference on Mechatronics and Automation, pp. 3571–3576 (2007)
8. Ess, A., et al.: A mobile vision system for robust multi-person tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
9. Tsai, T.-H., Yao, C.-H.: A real-time tracking algorithm for human following mobile robot. In: International SOC Design Conference (ISOCC), pp. 78–79, (2018)
10. Bellotto, N., Hu, H.: Multisensor-based human detection and tracking for mobile service robots. IEEE Trans. Syst. Man, Cybern. Part B 39(1), 167–181, (2009).
11. Kim, M., et al.: RFID-enabled target tracking and following with a mobile robot using direction finding antennas. In: 2007 IEEE International Conference on Automation Science and Engineering, pp. 1014–1019 (2007)
12. Verma, N.K., et al.: Vision based object follower automated guided vehicle using compressive tracking and stereo-vision. In: 2015 IEEE Bombay Section Symposium (IBSS), pp. 1–6 (2015)
13. Pang, L., et al.: A human-following approach using binocular camera. In: 2017 IEEE International Conference on Mechatronics and Automation (ICMA), pp. 1487–1492 (2017)
14. Chi, W., et al.: A gait recognition method for human following in service robots. IEEE Trans. Syst. Man, Cybern. Syst. 48(9), 1429–1440, (2018).
15. Sun, S., et al.: Human recognition for following robots with a Kinect sensor. In: 2016 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 1331–1336 (2016)
16. Lee, S., Horio, K.: Human tracking using particle filter with reliable appearance model. The SICE Annual Conference 2013, Nagoya, Japan, pp. 1418–1424 (2013)
17. Hoshino, F., Morioka, K.: Human following robot based on control of particle distribution with integrated range sensors. 2011 IEEE/SICE International Symposium on System Integration (SII), Kyoto, pp. 212–217 (2011)
18. Iswanto, I.: Visual object tracking based on mean-shift and particle-Kalman filter. Procedia Comput. Science A., Li, B. 116, 587–595 (2017)
19. Lee, C., et al.: Real-time embedded system for human detection and tracking. Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), pp. 147–148 (2019)
20. Lee, B., et al.: Robust human following by deep Bayesian trajectory prediction for home service robots. 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, pp. 7189–7195 (2018)
21. Chen, B.X., Sahdev, R., Tsotsos, J.K., et al.: Integrating stereo vision with a CNN tracker for a person-following robot. . Lect. Notes Comput. Sci., 10528, 300–313 (2017)
22. Zhu, Z., et al.: Human following for wheeled robot with monocular pan-tilt camera. arXiv:1909.06087, 2019
23. Koide, K., et al.: Monocular person tracking and identification with on-line deep feature selection for person following robots. Rob. Autom. Syst. 124, 103348
24. Ramasubramanian, M., et al.: A survey study on detecting and tracking objective methods. In: 2014 IEEE National Conference on Emerging Trends In New & Renewable Energy Sources And Energy Management (NCET NRES EM), pp. 159–164 (2014)
25. Wang, Q., et al.: Multi-cue based tracking. Neurocomputing 131, 227–236, (2014)
26. Fang, J., et al.: Part-based online tracking with geometry constraint and attention selection. IEEE Trans. Circuits Syst. Video Technol. 24(5), 854–864, (2014)
27. Kim, B.G., Park, D.J.: Unsupervised video object segmentation and tracking based on new edge features. Pattern Recogn. Lett. 25(15), 1731–1742, (2004)
28. Alvar, S.R., Bajic, I.V.: MV-YOLO: Motion vector-aided tracking by semantic object detection. IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–5, (2018)
29. Kim, B.G., Park, D.J.: Novel target segmentation and tracking based on fuzzy membership distribution for vision-based target tracking system. Image Vision Comput. (24), 1319–1331 (2006)
30. Zhu, S., Ma, K.-K.: A new diamond search algorithm for fast block matching motion estimation. In: Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. No. 97TH8237), vol. 1, pp. 292–296 (1997)
31. Zhu, C., et al.: Enhanced hexagonal search for fast block motion estimation. IEEE Trans. Circuits Syst. Video Technol. 14(10), 1210–1214, (2004)
32. Baraskar, T., et al.: Survey on block based pattern search techniques for motion estimation. In: 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 513–518 (2015)
33. Choudhury, H.A., Saikia, M.: Survey on block matching algorithms for motion estimation. In: Proceeding of 2014 International Conference on Communications and Signal Processing (ICCSP), pp. 513–518, (2014)
34. Belloulata, K., et al.: A novel cross-hexagon search algorithm for fast block motion estimation. In: International Workshop on Systems, Signal Processing and Their Applications, WOSSPA, pp. 1–4 (2011)
35. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 886–893 (2005)
36. Barnich, O., VanDroogenbroeck, M.: ViBe: A universal background subtraction algorithm for video sequences. IEEE Trans. Image Process. 20(6), 1709–1724, (2011)
37. Klein, D.A., et al.: Adaptive real-time video-tracking for arbitrary objects. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), 20(6), 1712–1715 (2010)

38. Königs, Schulz, D.: Fast visual people tracking using a feature-based people detector. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, 3614–3619 (2011)
39. Guler, Z., et al.: A new object tracking framework for interest point based feature extraction algorithms. Elektronika Ir Elektrotechnika 26(1), 63–71 (2020)

How to cite this article: Tsai T-H, Yao C-H. A robust tracking algorithm for a human-following mobile robot. *IET Image Process*. 2021;15:786–796.
https://doi.org/10.1049/ipr2.12062