IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON MACHINE LEARNING DESIGNS, IMPLEMENTATIONS AND TECHNIQUES

Received January 28, 2020, accepted February 15, 2020, date of publication February 27, 2020, date of current version March 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2976712

Interactions Between Specific Human and Omnidirectional Mobile Robot Using Deep Learning Approach: SSD-FN-KCF

CHIH-LYANG HWANG, (Senior Member, IEEE), DING-SHENG WANG, FAN-CHEN WENG, AND SHENG-LIN LAI

Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan

Corresponding author: Chih-Lyang Hwang (clhwang@mail.ntust.edu.tw)

This work was supported by the Project from the Ministry of Science and Technology at Taiwan under Grant MOST-108-2221-E-011-156.

ABSTRACT To fulfill the tasks of human-robot interaction (HRI), how to detect the specific human (SH) becomes paramount. In this paper, the deep learning approach by the integration of Single-Shot Detection, FaceNet, and Kernelized Correlation Filter (SSD-FN-KCF) is developed. From the outset, the SSD is employed to detect the human up to 8m using the RGB-D camera with 320×240 resolution. Afterward the omnidirectional mobile robot (ODMR) is driven to the neighborhood of $2.5\sim3.0m$ such that the depth image can accurately estimate the detected human's pose. Subsequently, the ODMR is commanded to the vicinity of $1.0m$ and the orientation inside $-60\sim60°$ with respect to the optical axis to identify whether he/she is the SH by the FaceNet. To reduce the computation time of the FaceNet and extend the SH's tracking, the KCF is employed to achieve the task of HRI (e.g., human following). Based on the image processing result, the required pose for searching or tracking (specific) human is accomplished by the image-based adaptive finite-time hierarchical constraint control. Finally, the experiment with the SH, who is far from and on the backside of the ODMR, validates the effectiveness and robustness of the proposed approach.

INDEX TERMS Deep learning, human detection, face recognition, visual tracking, omnidirectional mobile robot, adaptive finite-time hierarchical constraint control, human following.

I. INTRODUCTION

Human-robot interaction (or collaboration) has received increasing attention in the last decades, since robots may act as both helpers and companions for the elderly and impaired people, especially for an aging population [1], [2]. With the great progresses in robotics and rapid evolutions on computing systems, many advanced social, service, and surveillance mobile robots have been or are being developed around the world. One of the key functionality for these advanced robots is the ability to detect specific human in real time [3]–[8]. A service robot needs to be aware of human around and track a target person to provide services. A social robot should be able to pay attention to persons in the view and keep tracking the engaged persons in the interaction. A surveillance robot may be required to monitor persons in the scene and approach a suspected person for a close observation of his/her appearance and behavior. For these tasks, the most challenging scenarios are detecting and tracking multiple persons in frequently crowded and cluttered scenes in public environments. Real-time human detection and tracking has become one of the research focuses for service, social and surveillance robots in the literature due to its necessity for human-robot interaction.

Recently, at least two approaches for the object detection [9]–[12]. The 1st approach connected with deep learning is two-stage process. The 1st stage so-called ''Select Search'' finds the candidate region for the object. The convolution neural network (CNN) is employed to extract their features for prediction. In the 2nd stage, the corresponding features for different candidates are classified by support vector machine (SVM). Although the 1st approach is accurate enough, it needs much computation time. In contrast, the 2nd approach aggregates the 1st stage into the 2nd stage of the 1st approach such that the computation time is much reduced. Nevertheless, the accuracy is slightly reduced

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

but acceptable. The single shot detector (SSD) belongs to the 2nd approach [11]–[16]. In this paper, the SSD is first employed to detect the human(s), which is (are) not necessarily static. If human is not detected, the ODMR will execute the search of human by the image-based adaptive finite-time hierarchical constraint control (IB-AFTHCC) [17], [18]. After the detection of human, the ODMR approaches the detected human with an appropriate distance (e.g., 2.5∼3m). Afterward the pose between them is estimated by a depth image such that an accurate approach to a specific pose (e.g., in the vicinity of 1m and 0° with respect to optical axis) for face recognition is accomplished.

To judge whether the detected human is the specific human (SH), his/her face is recognized by the FaceNet [19], which is one kind of face recognition [20]–[23]. FaceNet directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. Only 128 bytes per face are required to achieve over 95% of robust recognition. For one SH, the calculation of FaceNet can be reduced such that the on-line searching of the SH using the IB-AFTHCC is feasible. After the identification of the SH, he/she is tracked by kernelized correlation filter (KCF) to avoid the repeated calculation of the FaceNet, extend the tracking distance, and reduce the computation time. The KCF utilizes the property of a circulant matrix and kernel to achieve the fast target tracking, and it can deal with the occlusion and scale changes in various scenes [24]–[28]. Afterwards the interactions between the SH and the ODMR (e.g., human following control) are implemented by the IB-AFTHCC. In contrast, if the face of the detected human is difficult to recognize, a searching strategy to obtain suitable pose for face recognition is progressed. If the SH is not detected, the ODMR moves forward a distance (e.g., 2.5m) to repeat the above searching procedure. To accelerate the processing, the GPU is combined with the CPU such that the on-line pose planning and control [29]–[32] and the human robot interactions [33]–[36] are more practical.

The contributions of this study are summarized as follows: (i) The learning of SSD can effectively detect the human beyond the general recognized distance of RGB-D camera system (e.g., 8m). (ii) The FaceNet is effectively learned to recognize the different faces with the recognition rate over 95% under suitable distance (0.75∼1.25m), different view angle (−60∼60°), different light angle (-80∼80°), and some occlusions. (iii) To avoid the repeated calculation of the FaceNet, extend the tracking distance, and reduce the computation time, the KCF is employed to track the SH such that human-robot interactions (e.g., human following) are achieved by the suggested IB-AFTHCC.

The outline of this study is as follows. In the next section, related work is given and discussed. In section 3, experimental setup and task description are described. In section 4, the deep learning using the integration of SSD, FaceNet, and KCF is developed. In section 5, the image-based adaptive finite-time hierarchical constraint control is employed to accomplish the required task of the human-robot interaction. In section 6, the corresponding experiments are presented to validate the effectiveness and robustness of the proposed method. Finally, the conclusions are given in section 7.

II. RELTATED WORK

At the outset, some representative papers about the human detection using the SSD are discussed. In [11], the local similarity (encoded by local descriptors) with a global context (i.e., a graph structure) of pairwise affinities among the local descriptors, embedding the query descriptors are combined into a low dimensional but discriminatory subspace. The power of Fourier transform combined with integral image to achieve superior runtime efficiency allows for testing multiple hypotheses within a reasonably short time; it is a training-free algorithm. The algorithm in [12] includes two different components that are trained ''in one shot'' at the first video frame: a detector that makes use of the generalized Hough transform with color and gradient descriptors and a probabilistic segmentation method based on global models for foreground and background color distributions. In [13], a framework integrating support vector machine based trail detection with a trail tracker is proposed to accomplish trail direction estimation and tracking at a low cost of computation and in real time. In [14], a fine-CNN with nine-layer neural network for the detailed pedestrian recognition is designed. A pedestrian in a surveillance video is segmented and fine recognized by the improved single-shot detector and several fine-CNNs, and is supported by parallel mechanisms provided by Apache Storm stream processing framework. Without post-processing other than efficient non-maximum suppression, an end-to-end trainable fast scene text detector, which is called TextBoxes++ and detects arbitrary oriented scene text with both high accuracy and efficiency in a single network forward pass, is developed [15]. In [16], the deep neural network with RGB-D image input predicts multiple grasp candidates for a single object or multiple objects, in a single shot with the real-time processing less than 0.25s.

Some representative papers for face (emotion) recognition are addressed as follows. In [19], FaceNet directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. Only 128 bytes per face are required to achieve over 95% of robust performance. In [20], a 3D face recognition method based on the fusion of shape and texture local binary patterns on a mesh is presented. It utilizes that the mesh surface preserving the full geometry doesn't require normalization, can accommodate the partial matching, and allows the early level fusion of texture and shape modalities. In [21], the CNN based face recognition using the ORL and Yale databases with gray scale images demonstrates the similar performance of the paper [23]. Nevertheless, its recognition rates are sensitive to the quantity of occlusion or pepper and salt noise. In [22], the enhanced face recognition method is proposed by utilizing local binary patterns histogram descriptors, Multi-K-nearest-neighbor, and Back Propagation Neural Network. Since the correlation method

utilized requires substantial computation time and large storage, features reduction and face representation are required. In [23], a local binary pattern histogram for the face recognition from the RGB together with suitable feature dimension of Depth images, which have a wide range of variations in head pose, illumination, facial expression, and occlusion in some cases, is developed to extract the facial features and then improve the recognition rate. In [36], two-layer fuzzy support vector regression-Takagi-Sugeno model is proposed for the emotion understanding in human-robot interaction, e.g., the scenario of drink in different emotions. However, its maximum average video-based recognition rate for different genders, provinces, and ages is 77.62%, which is not excellent.

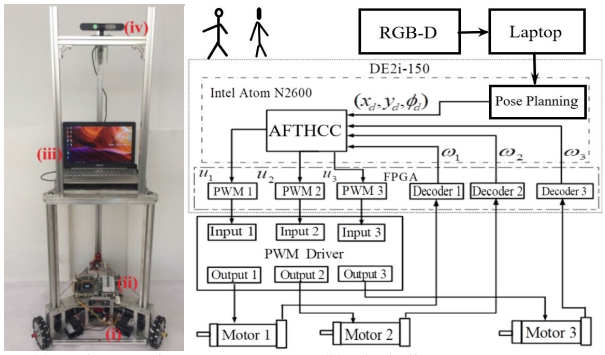
Finally, the representative papers about the tracking of the SH using kernelized correlation filter (KCF) are addressed. In [24], both KCF and dual correlation filter outperform top-ranking trackers such as structured output tracking with kernels or tracking-learning detection on a 50 videos benchmark, despite running at hundreds of frames-per-second, and being implemented in a few lines of code. It indicates that KCF is indeed a fast and effective tracker. A real-time RGB-D object tracker based on the KCF, which utilizes the property of a circulant matrix and kernel to achieve fast target tracking, is proposed to deal with occlusion and scale changes in various scenes [25]. In [26], tracking algorithm with reducing feature dimensionality and interpolating correlation score are employed to reduce the computational cost for fast tracking. Occlusion and fast motion problems can be effectively solved by the expansion of the search area with the speed of 69.5 frames per second, which is suitable for real-time application. By integrating an adaptive obstacle detection strategy within a KCF framework, a fast and robust approach for obstacle detection and tracking is developed [27]. In [28], an online learning method based on the KCF and assembles different feature channels to kernelized experts is employed to track vehicles at night. By estimating their reliabilities, the appearance model to focus on the most discriminative visual features achieves the classification.

Finally, the deep learning integrating the advantages from SSD, FaceNet, and KCF is proposed to deal with the corresponding human-robot interactions [33]–[36].

III. EXPERIMENTAL SETUP AND TASK DESCRIPTION

A. EXPERIMENTAL SETUP

The experimental setup of the proposed omnidirectional mobile robot (ODMR) in Fig. 1(a) includes the following four parts: (i) three dc servomotors, (ii) one motion control platform, (iii) a laptop for image processing, and (iv) a RGB-D camera system. Three dc servomotors are the model no. 578296 with gear ratio 66:1 from Maxon Co.; in contrast, the gear ratio between the motor and wheel is 1.2:1. The driver is the model no. ESCON-422969 with the following important specifications: (i) power: 700W, (ii) input voltage: 10∼70V, (iii) peak current: 30A, (iv) continuous current: 10A, (v) weight: 204g, (vi) dimension: 115 × 75.5 × 24mm.



(a) Photograph. (b) Block diagram.

FIGURE 1. The proposed ODMR.

The important specifications of the motion control platform DE2i-150 are as follows: (i) microprocessor: Intel® Atom™ Dual Core Processor N2600 (1M Cache, 1.6GHz); 64-bit Integrated Graphics with Base Frequency 400MHz. (ii) FPGA (Field Programmable Gate Array): 49,760 Les; 6,480 Kbits embedded memory; 8 PLLs (Phase Lock Loops); (iii) 40-pin Expansion Header (GPIO); (iv) 4GB DDR3 SODIMM. The suggested adaptive finite-time hierarchical constraint control (AFTHCC) algorithm is computed in the Intel® Atom N2600; the PWM for driving the motor and the decoder for obtaining motor velocity are executed in the FPGA (cf. Fig. 1(b)). On the other hand, the laptop is the ASUS computer with the following important specifications: (i) Intel Core i7 CPU, with six-core and 12 threads, (ii) GPU with GTX 960M, CUDA core 640.

The proposed RGB-D vision system is the model of ASUS Xtion PRO version, which provides not only depth data but also RGB colour image and audio (using a microphone array). It is installed at 145 cm and has the following important specifications: (i) Dimensions: 177.8 × 48.2 × 38.1mm. (ii) Weight: 540g, (iii) USB connection: consumption < 2.5W, (iv) Detection range: 0.8m ∼3.5m, (v) FOV: horizontal 58°, vertical 45°, (vi) Depth image: VGA (640 × 480) 30fps, QVGA (320 × 240) 60fps, (vii) Compatible with the OpenNi development framework, (viii) Supported OS: Win 32/64: XP, Vista, Win7, Win 10, Linux Ubuntu 10.10: X86, 32/64 bit, Android (on demand), (ix) Programming languages: C++/C# (Windows), C/C++ (Linux), JAVA.

B. TASK DESCRIPTION

The fundamental tool of deep learning is the Convolutional Neural Network (CNN); CNN is made up by Convolution, Pooling, and Full Connection layers. Their sizes and layers are dependent on practical applications. The Convolution layer can have different length, width and height such that the dimension, complexity, and characteristics of input image can be included into different dimension of convolution and nonlinear transform. To avoid the losing features of input image, the resolution of CNN will keep the same as that of input image. In contrast, the Pooling layer is employed to reduce the corresponding resolution. Suitable dimension also

can increase the robustness [37]. Finally, multi-dimension features are transformed into one dimension feature by the full connection layer.

In this study, the SSD is first employed to detect the human, which is not necessarily static. If human is not detected, the robot will execute the search of human. After the detection of human, the ODMR will approach the detected human with an appropriate distance (e.g., 2.5∼3m). Then using a depth image estimates the pose between them such that an accurate approach to the detected human with specific pose (e.g., 0.75∼1.25m and $-60\sim60°$) is achieved. To judge whether the detected human is the specific human (SH), his/her face is recognized by the FaceNet [19]. If the SH is identified, the SH will be tracked by the KCF [24]–[28] to avoid the repeated calculation of the FaceNet. Afterward, the corresponding interactions between the SH and the ODMR (e.g., human following control) are implemented by the image-based adaptive finite-time hierarchical constraint control (IB-AFTHCC) [17], [18]. If the face of detected human is difficult to recognize, the ODMR will be commanded to a suitable pose to execute the FaceNet algorithm. If the detected human is not the SH, the ODMR will continue the searching (e.g., moving forward a distance of 2.5m, repeating the above procedure) until the SH is detected. The corresponding flowchart is depicted in Fig. 2.

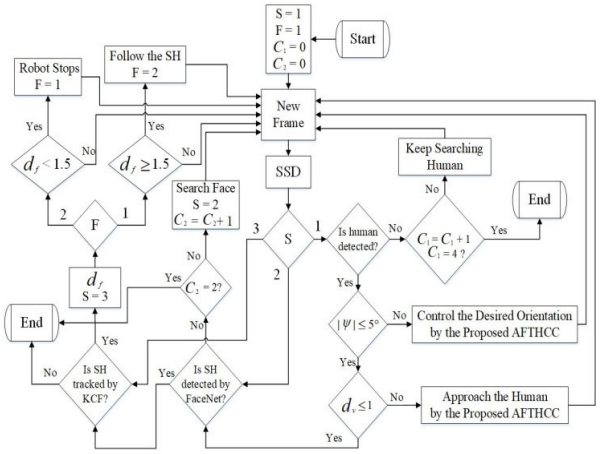


FIGURE 2. Overall flow chart of this study.

To accelerate the processing, the GPU is combined with the CPU to accomplish the processing time about 50 *ms* such that the on-line pose planning and control [29]–[32] and the human-robot interactions [33]–[36] are more practical. Moreover, the parameters of Fig. 2 are explained as follows: (i) S: S = 1, 2, 3 respectivly determines the initation of the SSD, FaceNet, and KCF functions, (ii) F: the index to determine whether the human following is executed, (iii) C_1: the counter to determine whether which region is searched, (iv) C_2: the counter to determine whether the face searching is required, (v) d_v, ψ: the vertical and orientation between human and the ODMR, (vi) d_f: the distance index to determine whether the task of human following is implemented.

IV. DEEP LEARNING APPROACH: SSD-FN-KCF

The proposed deep learning technique for the human detection, the specific human's recognition, and the tracking of specific human are tackled by the integration of Single-Shot Detection (SSD), FaceNet (FN), and Kernelized Correlation Filter (KCF).

A. HUMAN DETECTION USING SSD

Recently, at least two deep learning approaches for the object detection are developed. The 1st approach is two-stage process. The 1st stage so-called "Select Search" finds the candidate region for the object. The convolution neural network (CNN) is employed to extract their features for prediction. In the 2nd stage, the corresponding features for different candidates are classified by support vector machine (SVM). In this study, the single shot detector (SSD) belongs to the 2nd approach [11]–[16]. The corresponding human detection using the SSD is depicted in Fig. 3. Its red rectangle is the VGG16 CNN in the blue rectangle with the replacement of the original FC6 and FC7 by the Conv6 and Conv7, the removal of all connection layers, and the extra addition of CNNs.

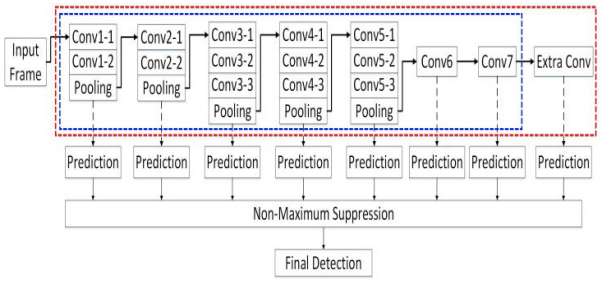


FIGURE 3. Flowchart of human detection using SSD.

The proposed approach uses the trained CNN model to predict each default box. It is also called the possibility of each class as the "score." Afterwards, non-maximum suppression (NMS) is employed to screen the best prediction. Together with the bounding box and score, the corresponding feature vector is extracted.

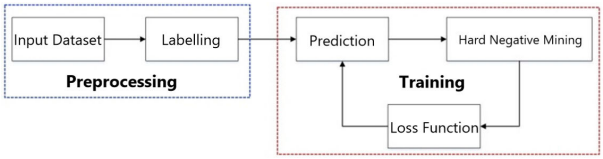


FIGURE 4. Learning process of SSD.

Subsequently, the learning process of the SSD including Preprocessing (i.e., steps 1 and 2) and Training (i.e., steps 3∼5) is introduced as follows (cf. Fig. 4).

1) INPUT DATASET

Prepare the corresponding images with the same resolutions including human and nonhuman.

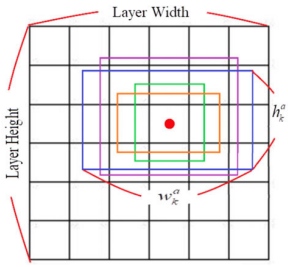


FIGURE 5. Default boxes.

2) LABELLING

The dataset are first classified with human and nonhuman. Then the coordinates of their ground truth boxes are labelled for later training.

3) PREDICTION

Since the detected humans possess different sizes caused by different human or different distance, the height and width of the CNN are different. Hence, the default boxes in Fig. 5 with respect to the feature map center include different heights and widths, which are determined by the scale and aspect ratio. At the outset, the scale is expressed as follows:

$$S_k = S_{\min} + \{(k-1)(S_{\max} + S_{\min})\}/(m-1) \tag{1}$$

where S_k is the scale of the *k-th* layer in the feature map, $S_{\max}$ denotes the scale of the highest layer (e.g., $S_{\max} = 0.95$), $S_{\min}$ denotes the scale of the lowest layer (e. g., $S_{\min} = 0.2$), and m is the number of feature map, e.g., $m = 6$.

On the other hand, the aspect ratios are assumed to be $a_r = \{1, 2, 3, 1/2, 1/3\}$. Their relations with height and width are as follows (cf. Fig. 5):

$$w_k^a = S_k\sqrt{a_r}, \quad h_k^a = S_k/\sqrt{a_r} \tag{2}$$

where w_k^a and h_k^a are the width and height of the *k-th* layer. Moreover, the scale of the default box for $a_r = 1$ is described as follows:

$$\overline{S}_k = \sqrt{S_k^2 + 1} \tag{3}$$

It is assumed that each feature map cell contains n_d default boxes (default: $n_d = 6$), each cell possesses 4 offsets, the number of classification is P (e.g., $P = 1$), the dimension of feature map is $k \times l$. Then the total number of the default box is $(P + 4) \times n_d \times k \times l$.

4) HARD NEGATIVE MINING

Using Jaccard Overlap estimates the similarity between the ground truth box B_g and the default box B_d:

$$J(B_g, B_d) = \left|B_g \cap B_d\right| / \left|B_g \cup B_d\right|. \tag{4}$$

If $\max\left\{J(B_g, B_d)\right\} > 0.5$, then they are positive boxes; otherwise, they are negative boxes. To maintain the stability of training and loss value, hard negative mining only chooses the higher score of negative boxes (i.e., far away from 0.5) such that the ratio between the positive boxes and the negative boxes is 1:3 (cf. Fig. 7(e) and 7(f)).

5) LOSS FUNCTION

The objective loss function is defined by the weighted combination of classification loss (subscript cla) and localization loss (subscript loc):

$$L_{SSD}(x, c, l, g) = \{L_{cla}(x, c) + \alpha L_{loc}(x, l, g)\}/N \tag{5}$$

where N is the total matching number between the ground truth box and the default box, in general $\alpha = 1$, and the classification loss is calculated by softmax loss [11], which purpose makes the confidence of positive and negative samples with enough robustness to recognize the ground truth box of each class. Here x is a parameter, c denotes the confidence for the detection. The classification loss is defined as follows:

$$L_{cla}(x, c) = -\sum_{j=1}^{N_g}\left[\sum_{i\in Pos}^{N_p} x_{ij}^p \log(\hat{c}_i^p) + \sum_{i\in Neg}^{N_n} \log(\hat{c}_i^n)\right] \tag{6}$$

where $\hat{c}_i^p = e^{c_i^p}/\sum_{i\in Pos}^{N_p} e^{c_i^p} \in [0, 1]$, c_i^p is the confidence of class p (positive sample) of the *i-th* default box, $\hat{c}_i^n$ is the confidence of class n (negative sample) not belonging to any class, x_{ij}^p is the matching index between the *i-th* default box and the *j-th* ground truth box in class p, and N_g, N_p, N_n are the numbers of ground truth, class p and class n, respectively. If it is matched, $x_{ij}^p = 1$; otherwise, $x_{ij}^p = 0$. In this paper, $N_g = 1$.

Before introducing the localization loss, a box in image processing is described by 4-dimensional space (cx, cy, w, h): the center in 2D, the width and height of box. Hence, $g_j^m = [g_j^{cx}, g_j^{cy}, g_j^w, g_j^h]$ and $d_i^m = [d_i^{cx}, d_i^{cy}, d_i^w, d_i^h]$ denote the *j-th* ground truth box and the *i-th* default box, respectively. Then the ground truth box regressing to offsets are defined as follows:

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w, \quad \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h \tag{7}$$

$$\hat{g}_j^w = \log(g_j^w/d_i^w), \quad \hat{g}_j^h = \log(g_j^h/d_i^h). \tag{8}$$

Similarly, the localization loss is defined as follows:

$$L_{loc}(x, l, g) = \sum_{j=1}^{N_g}\left[\sum_{i\in Pos}^{N_p}\sum_{m=1}^{4} x_{ij}^p S_{L1}(l_i^m - \hat{g}_j^m)\right] \tag{9}$$

where $m = 1, 2, 3, 4 = cx, cy, w, h$, $l_i^{1\sim4} = \left[l_i^{cx}, l_i^{cy}, l_i^w, l_i^h\right]^T$ denotes the *i-th* prediction box, and

$$S_{L1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{10}$$

The purpose of the computed loss function in (10) is to reduce the differences between prediction box and default box and between default box and ground truth box, such that the iterative prediction of the difference between prediction box and ground truth box becomes smaller. After the achievement of steady-state matching error, the predicted bounding box for the target human is satisfactorily obtained. At the outset, the resolution of RGB image is set as 320 × 240, and the total number of trained images is 512. Some representative

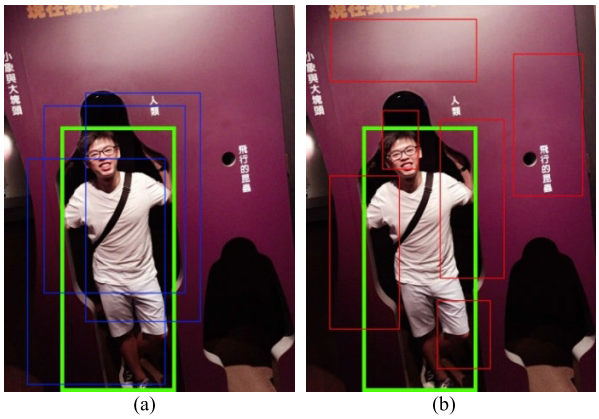


FIGURE 6. Some samples for human detection: (a) positive boxes, (b) negative boxes (ground truth box in green color).

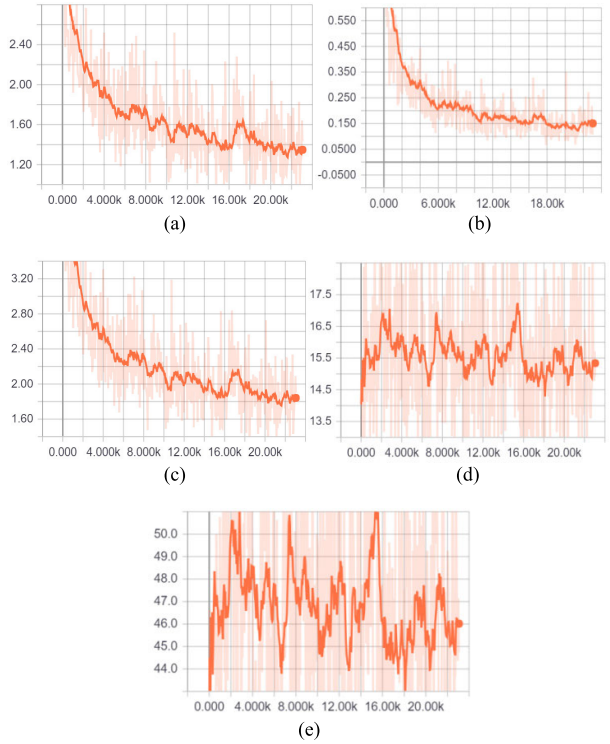


FIGURE 7. Processing of SSD for human detection: (a) Classification loss, (b) Localization loss, (c) Total loss, (d) Positive samples, (e) Negative samples.

positive and negative samples for human detection are shown in Fig. 6.

The corresponding loss function responses are presented in Fig. 7 (a), (b). After 23,000 time step to train, the total loss function converges in the neighborhood of 1.8 (Fig. 7(c)). The numbers of positive and negative samples are shown in Fig. 7(d), (e), in which the number ratio is about 1:3. Since each frame is different, the corresponding responses in Fig. 7 are fluctuated. Even though, the average responses in the solid lines are also presented.

TABLE 1. Human detection rate (HDR) using ssd uder different distances.

Distance (m)	1	2	3	4	5	6	7
HDR (%)	100	100	100	98.7	100	100	100

TABLE 2. HDR for different view angles at 2m.

Distance (m)	2						
Angle (°)	45	-45	90	-90	135	-135	180
HDR	100	100	100	99.4	100	100	100

TABLE 3. HDR for different view angles at 4m.

Distance (m)	4						
Angle (°)	45	-45	90	-90	135	-135	180
HDR (%)	100	100	98.8	100	100	99.2	99.4

6) PERFORMANCE EVALUATION

To validate the effectiveness and robustness of SSD, the (walking) human with different distances and view angles are investigated. The results are shown in Tables 1-3.

In summary, irrespective of the distance ($<8m$) and view angle ($<=180°$) between human and camera, the human detection rate above 98.6% and the frame rate above 30 fps are achieved by the resolution of 320×240. The performance evaluation can refer to URL: https://youtu.be/gUJd0ATnlXI.

B. SPECIFIC HUMAN DETECTION USING FACENET

In the previous studies about face recognition (e.g., LBP with SVM [23]), they need build the feature descriptor with suitable dimension, and then a (multiclass) support vector machine is trained to obtain a model for classification. In contrast, the approach of FaceNet directly learns Euclidean mapping transformed from the face pattern mapping. The similarity of face pattern is expressed as the Euclidean distance in the Embedding layer using 128 dimensionalities [19]. After the learned model, the compared Euclidean distance error below a specific threshold is set as a classification standard. The flowchart (or procedure) of FaceNet includes Batch (input frame), Deep Architecture, L2 Normalization, Embedding, and Triplet Loss. Since the classifier is not required, the training procedure is more fast and effective. Finally, the comparison between the trained model and the on-line calculated "Embedding values" for an image at specific time interval determines whether the detected human is the SH or not.

1) DEEP ARCHITECTURE

The CNN of FaceNet is chosen from Inception ResNet-v2 of Google. The details are described as follows [19]: (i) The 1st layer NN1 is made up by ZF-Net with 22 CNN and extra CNN, (ii) The 2nd layer NN2 is made up by many Inceptions, (iii) NN3, NN4, NNS1, and NNS2 are employed to reduce the resolution or scale.

2) NORMALIZATION

To map the image $x \in \Re^n$ into hypersphere, L2 (or Euclidean distance) normalization $x' \in \Re^n$ is considered.

$$x_i' = x_i / \|x\|_2, \quad i = 1, 2, \ldots, n \tag{11}$$

In general, the feature vector in practical image is discontinuous or discrete. Hence, one-shot encoding (i.e., one feature uses one code) can transform this discrete feature into another feature in Euclidean space. The advantages of one-shot encoding are less computation and strong feature description.

3) EMBEDDING

Even though the advantages of one-shot encoding, the storage will be largely increasing as the number of feature increases. To improve this drawback, the Embedding layer in the FaceNet is reduced to a 128 dimensional byte vector, satisfying the requirement of face recognition.

4) TRIPLET LOSS

The loss function of the FaceNet is triplet loss [19]. There have three features for the modeling: (i) the desired feature is denoted as ''Anchor'', (ii) the feature slightly deviating from Anchor is denoted as ''Positive'', and (iii) the features dominantly deviating from Anchor is called as ''Negative.'' The purpose of triplet loss is to minimize the Euclidean distance between Anchor and Positive, and simultaneously maximize the Euclidean distance between Anchor and Negative (cf. Fig. 8).

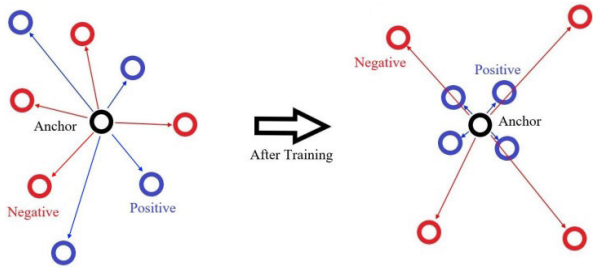


FIGURE 8. Triplet loss after training.

It is assumed that Anchor and all Positives are similar, and that Anchor and all Negatives are dissimilar. In brief, the following loss function is defined to be minimized.

$$L_{Tr} = \sum_{i=1}^{N} \left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \beta \tag{12}$$

where $\beta > 0$, x_i^a, x_i^p and x_i^n are Anchor, Positive and Negative, respectively. Moreover, $f(x_i^a), f(x_i^p), f(x_i^n) \in \Re^d$, and $\left\| f(x_i^j) \right\| = 1$, for $j = a, p, n$.

Based on the distances among Anchor, Positive and Negative, triplets are categorized as (i) Easy triplets with $L_{Tr} < 0$, (ii) Hard triplets with $L_{Tr} > \beta$,(iii) Semi-hard triplets with $0 \le L_{Tr} \le \beta$. If Negative belongs to Easy triplets, the Triple loss equals zero. If Negative belongs to Hard triplets, the similarity between them is large such that it is difficult to recognize. In contrast, if Negative belongs to Semi-hard triplets, these datasets are learned to maximize the face recognition rate.

5) COMPARISON

After the effective training model of the FaceNet, the output of Embedding layer for the real-time image will be compared with that of the trained model. If the Euclidean distance between them is smaller than a specific threshold, then the recognition of the specific human (SH) is achieved. Otherwise, it is not the SH. In summary, the end-to-end training both simplifies the setup and shows that directly optimizing a loss relevant to the task at hand improves performance [19].

6) PERFORMANCE EVALUATION

Two important factors to affect the face recognition rate (FRR): one is the parameter setting in the FaceNet (e.g., β, the resolution of camera, the number of training samples), the other is the environmental change (e.g., the distance or view angle between camera and robot, the occlusion of human, the lighting condition).The SH in this paper is shown in Fig. 9(a). On the other hand, non-SH is presented in Fig. 9(b).

(a) SH.

(b) Non-SH.

FIGURE 9. Training samples for the SH and Non-SH.

At the outset, $\beta = 0.6$, the resolution 320×240, 300 training samples, and the height of camera at 145 *cm* are assigned. The face recognition rate (FRR) for different distances, view

TABLE 4. FRR for different distances.

Distance (m)	0.5	0.75	1	1.25	1.5
FRR (%)	100	100	100	100	67.9

TABLE 5. FRR for different view angles at 1m.

Distance (m)	1							
Angle (°)	30	-30	45	-45	60	-60	75	-75
FRR (%)	100	100	100	100	100	99	75.3	71.8

TABLE 6. FRR for different view angles at 1.25m.

Distance (m)	1.25							
Angle (°)	30	-30	45	-45	60	-60°	75	-75
FRR (%)	100	100	100	100	100	98.2	61.1	59.9

TABLE 7. FRR for different half face at 1m.

Lighting Angle (°)	45	-45	90	-90
FRR (%)	100	100	98.9	94.8

TABLE 8. FRR for different lighting angles at 1m.

Half Face	Upper	Lower	Right	Left
FRR (%)	**17.3**	100	100	100

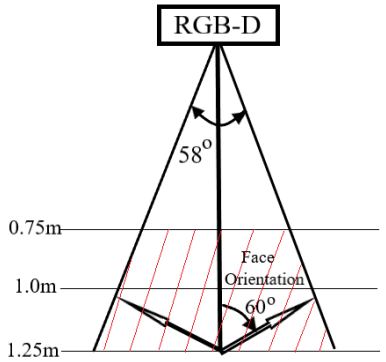


FIGURE 10. Effective recognition regions of the FaceNet using RGB-D camera.

angles at $1m$ and $1.5m$, half faces, and lighting angles are investigated and listed in Table 4, 5, 6, 7, and 8 respectively.

Based on the results of Tables 4-8 and the FOV of camera with the horizontal 58° and vertical 45° view angles, the important observations are concluded as follows: (i) The red-cross region in Fig. 10 (i.e., between $0.75m$ and $1.25m$ with the orientation less than 60°) at the height of $1.45m$ is called as the effective face recognition with the FRR over 95% using the FaceNet with RGB-D camera. (ii) Human in the front of camera with one meter, without the occlusion of two eyes and the occlusion not over 50% possesses the FRR over 95%. (iii) Light orientation not over 90° in the dark environment also has the FRR over 95%. (iv) The occlusion of the upper half (i.e., two eyes and nose) only has FRR 17.3. It is reasonable since two eyes are one of key sub-region for the face recognition. (v) The processing frequency is about 20 fps, which is 50% smaller than that of the SSD. (vi) The performance evaluation can refer to the attached URL.

C. SPECIFIC HUMAN TRACKING USING KCF

Since the SSD and FaceNet are only for *a specific frame* to judge whether the specific human (SH) is recognized, the corresponding result will disappear in the next frame. Hence, if the SH is detected, then the Kernelized Correlation Filter (KCF) can maintain to track the SH. The KCF is identified as a tracking approach, i.e., the simultaneous SH tracking and training prediction.

1) PRELIMINARIES

To consider the practical situation, a nonlinear mapping for sample x_i around the (blue) rectangle detected by the FaceNet is denoted as $\varphi(x_i)$. Its kernelizing regression function becomes $y_i = f(x_i) = w^T\varphi(x_i)$. Define the following circulant matrix [24]:

$$X = C(x) = \begin{bmatrix} x_1 & x_2 & \cdots & x_{n-1} & x_n \\ x_n & x_1 & x_2 & \cdots & x_{n-1} \\ \vdots & x_n & \ddots & \ddots & \vdots \\ x_3 & \vdots & \ddots & x_1 & x_2 \\ x_2 & x_3 & \cdots & x_n & x_1 \end{bmatrix} \tag{13}$$

where the 1st row of $C(x)$ is $x = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$. The regularized least square weight estimation is to minimize the following cost function:

$$\varepsilon(w) = \left\{ \left\| f(x) - w^T\phi(x) \right\|^2 + \lambda \left\| w \right\|^2 \right\} \tag{14}$$

where $\phi(x) \in \Re^n$, $w = \alpha^T\phi(x), f(x) = \alpha^T\phi^T(x)\phi(x)$. Then the minimization of (14) yields

$$\alpha = [K + \lambda I_n]^{-1} y \tag{15}$$

where $K_{ij} = \phi^T(x_i)\phi(x_j) = k(x_i, x_j)$ is the row i and column j of K.

2) DIAGONALIZATION OF CIRCULANT MATRIX USING DFT

$$X = F\ diag\left\{\hat{x}\right\} F^H, \quad X^H = F\ diag\left\{\hat{x}^*\right\} F^H \tag{16}$$

where $\hat{x}$ is the DFT of x, $X^H = (X^*)^T$, and

$$F = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega & \cdots & \omega^{n-1} \\ \vdots & & & \vdots \\ 1 & \omega^{n-1} & \cdots & \omega^{2n+1} \end{bmatrix} \Big/ \sqrt{n}, \quad \omega = e^{-2\pi i/n}.$$

From (15), (16), and $FF^H = I_n$, the following result is achieved.

$$\alpha = \left[C(\mathbf{k}^{xx}) + \lambda I_n\right]^{-1} y = \left[F\ diag(\hat{\mathbf{k}}^{xx})F^H + \lambda I_n\right]^{-1} y$$
$$= F\left[diag(\hat{\mathbf{k}}^{xx} + \lambda)\right]^{-1} F^H y. \tag{17a}$$

Or, equivalently,

$$F^H\alpha = \left[diag(\hat{\mathbf{k}}^{xx} + \lambda)\right]^{-1} F^H y. \tag{17b}$$

Using $\hat{z}^* = F^H z$ yields

$$\hat{\alpha}^* = \left[diag(\hat{\mathbf{k}}^{xx} + \lambda)\right]^{-1} \hat{y}^* = \hat{y}^*/(\hat{\mathbf{k}}^{xx} + \lambda). \quad (18)$$

The regression function or all candidate patches with z is computed as follows:

$$f(z) = \left(C(\mathbf{k}^{xz})\right)^T \alpha = \left(F\ diag(\hat{\mathbf{k}}^{xz})F^H\right)^T \alpha$$
$$= F^H\ diag(\hat{\mathbf{k}}^{xz})F\alpha. \quad (19)$$

Here $f(z)$ is a vector, containing the output for all cyclic shifts of z, i.e., the full detection response. From (19),

$$Ff(z) = diag(\hat{\mathbf{k}}^{xz})F\alpha.$$

Similarly,

$$\hat{f}(z) = \hat{\mathbf{k}}^{xz} \odot \hat{\alpha}. \quad (20)$$

Here $\odot$ denotes the pointwise operator. It indicates that the output of each input vector is the pointwise multiplication of $\hat{\mathbf{k}}^{xz}$ and $\hat{\alpha}$. The maximum output is the maximum likelihood of the next position of target.

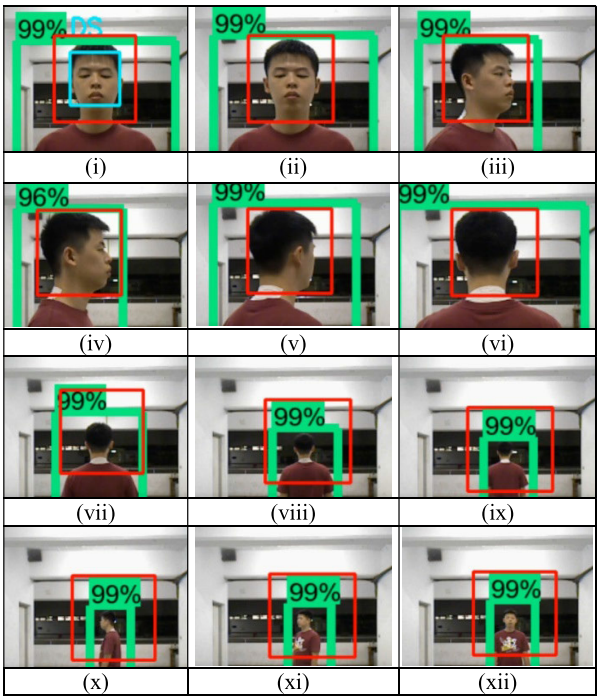


FIGURE 11. Performance evaluation of KCF with simultaneous FaceNet and SSD operations. (i) SH is respectively detected, recognized and tracked by SSD, FaceNet and KCF, (ii) FaceNet does not work, (iii)–(vi) different orientations of human are still successfully detected and tracked by SSD and KCF, (vii)–(ix) different distances of human are still successfully detected and tracked by SSD and KCF, (x)–(xii) different orientations at 8m between them are still successfully detected and tracked by SSD and KCF.

3) PERFOMANCE EVALUATION

The representative experimental result is shown in Fig. 11. The green, blue and red rectangles in Fig. 11 are the human detected by the SSD, the SH recognized by the FaceNet, and tracked by the KCF, respectively. After the Fig. 11 (v), the FaceNet fails to recognize the SH but the KCF still continues to track the SH. The reasons for the failure are that Fig. 11 (iv), 11(v) and 11(vi) have too large viewing angle (cf. Tables 5 and 6), that Fig. 11(vii)-11(xii) have too large distance between them (cf. Table 4). It confirms the effectiveness of the KCF.

TABLE 9. Relation between the width of the green rectangle and the vertical distance between robot and human.

$d_v\,(m)$	3	4	5	6	7	8
w (pixel)	57.16	46.53	36.06	34.07	29.65	28.08

V. IMAGE-BASED POSE TRACKING USING ADAPTIVE FINITE-TIME HIERARCHICAL CONSTRAINT CONTROL

A. IMAGE-BASED DESIRED POSE

The width of the green rectangle for the human detection using the SSD is employed to estimate the vertical and horizontal distances between the ODMR and the detected human. As distance is smaller than $3.5m$, the depth image can be accurately estimated. Hence, only $3 \sim 8m$ between them are presented in Table 9.

Since the detected human is too near the vision system, e.g., $4m$, the whole ROI for the human is infeasible. Only the green rectangle's width is used to estimate the 2D pose between ODMR and detected human. Based on the result of Table 9, the estimated vertical distance between the ODMR and the rectangle's width is achieved as follows:

$$d_v = 0.0079w^2 - 0.83w + 24.725. \quad (21)$$

Moreover, the horizontal distance $d_{h,j}$ at distance j meter is computed by the central pixel c_p at $d_v = 3, 4, 5, 6, 7m$:

$$d_{h,3} = 0.01173c_p - 1.8502, \quad d_{h,4} = 0.01486c_p - 2.3721$$
$$d_{h,5} = 0.01867c_p - 2.9233, \quad d_{h,6} = 0.02199c_p - 3.5463$$
$$d_{h,7} = 0.02613c_p - 4.0813. \quad (22)$$

The d_h between $d_v = 3$ and $7m$ is achieved as follows:

$$d_h = d_{h,i} + (d_{h,i+1} - d_{h,i})(d_v - d_{v,i})/(d_{v,i+1} - d_{v,i}) \quad (23)$$

where $d_{h,i} < d_h < d_{h,i+1}, d_{v,i} < d_v < d_{v,i+1}, i = 3, 4, 5, 6$. The desired 2D pose for the AFTHCC in the next subsection becomes

$$x_d = d_h, y_d = d_v, \quad \phi_d = \psi = \tan^{-1}(d_h/d_v). \quad (24)$$

B. AFTHCC

From the outset, the system variables and parameters of ODMR are listed in Table 10 and Table 11, respectively. These values in Table 11 are obtained from the manual of motor, the dimension of ODMR, the knowledge of friction.

Based on the information of image processing, the required pose tracking for searching or tracking (specific) human is achieved by the ODMR with AFTHCC [17], [18]. At first, the kinematic relation of the ODMR is depicted in Fig. 12.

Its dynamics using the states $z_1(t) = x(t), z_2(t) = y(t), z_3(t) = \phi(t), z_4(t) = v_x(t), z_5(t) = v_y(t), z_6(t) =$

TABLE 10. Variables of ODMR.

Variable	Description
(v_x, v_y, ω)	the velocity and the orientation of the ODMR with respect to the robot coordinate $X_m - Y_m$
$v_j(t),\ \omega_j(t)$	the linear and angular velocity of wheel j
$f_x(t), f_y(t)$ $f_{f_x}(t), f_{f_y}(t)$	the force and friction force in X_m- and Y_m- directions at the mass center of ODMR
$\tau(t), \tau_f(t)$	the torque and friction torque at the mass center of ODMR
$f_j(t), \tau_j(t)$	the force and torque of wheel j
$f_{f_j}(t), \tau_{f_j}(t)$	the friction force and torque of wheel j
(x, y, ϕ)	the pose of the ODMR with respect to the world coordinate $X-Y$
$f_{sl}(v_j), f_{st}(f_j)$	the slipping and stiction forces of wheel j
$\tau_{sl}(\omega_j), \tau_{st}(\tau_j)$	the slipping and stiction torques of wheel j
$i_j(t), u_j(t)$	the current and voltage of motor j

TABLE 11. Parameters of ODMR.

Parameter	Description	Value
M_r	the total mass of ODMR	$42.35kg$
I_r	the total moment of inertia of ODMR	$2.98kgm^2$
l_w	the length between the mass center and wheel	$0.36\,m$
I_w	the moment inertia of wheel	$0.182kgm^2$
r_w	the radius of wheel	$0.076m$
R_j	the resistances of motor j	0.103Ω
L_j	the inductances of motor j	$0.072H$
K_{b_j}	the back electromotive force constant of motor j	$0.0385V/rad/s$
K_{t_j}	the torque constant of motor j	$0.0385Nm/A$
N_j	the total gear ratio of wheel j	66:1
N_j^*	the gear ratio from motor j to its wheel	1:2.67
s_θ, c_θ	$s_\theta=\sin(\theta)$, $c_\theta=\cos(\theta)$	$\theta=\pi/6$
f_s^+, f_s^-,	Coulomb friction force	$8.76, -8.62N$
τ_s^+, τ_s^-	Coulomb friction torque	$8.26, -8.09Nm$
$\delta f_s^+, \delta f_s^-$	Stribeck friction force	$4.24, -4.19N$
$\delta\tau_s^+, \delta\tau_s^-$,	Stribeck friction torque	$3.84, -3.73Nm$
v^+, v^-	Stribeck velocity	$0.20, 0.19m/s$
ω^+, ω^-	Stribeck angular velocity	$0.22, 0.21rad/s$
C_f^+, C_f^-	Viscous friction force coefficient	$0.25, 0.25\,Ns/m$
C_τ^+, C_τ^-	Viscous friction torque coefficient	$0.25, 0.25\,Ns/rad$
δ_f, δ_τ	Stiction (angular) velocity	$0.25m/s, 0.25rad/s$

$\omega(t)$, $z_7(t) = i_1(t)$, $z_8(t) = i_2(t)$, $z_9(t) = i_3(t)$, and constraint input is as follows:

$$\dot{Z}_1(t) = A_1(Z,t), \dot{Z}_2(t) = A_2(Z,t) + B(Z,t)C(U) \tag{25}$$

$$Y_1(t) = C_1\left[Z_1(t) + \Delta Z_1(t)\right], \quad Y_2(t) = C_2\left[Z_2(t) + \Delta Z_2(t)\right] \tag{26}$$

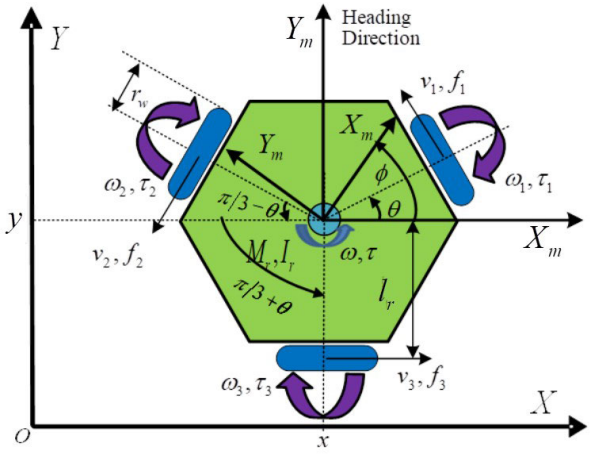


FIGURE 12. Schematic description of the ODMR.

where $Z_1^T(t) = \left[z_1(t)\ z_2(t)\ z_3(t)\ z_4(t)\ z_5(t)\ z_6(t)\right]$ and $Z_2^T(t) = \left[z_7(t)\ z_8(t)\ z_9(t)\right]$ are the indirect and direct states, respectively; $Y_1(t)$ and $Y_2(t) \in \Re^3$ are the indirect and direct outputs, respectively; $C(U)$ denotes the constraint of the control input $U(t) = \left[u_1(t)\ u_2(t)\ u_3(t)\right]^{\mathrm{T}} \in \Re^3$; $A_i(Z,t)$, $i = 1, 2$ and $B(Z,t)$ are the true system vector functions; $C_1 = \left[I_3\ \mathbf{0}_{3\times 3}\right]$ and $C_2 = I_3$ are the output gain matrices; $\Delta Z_1^T(t) = \left[\Delta z_1(t)\ \Delta z_2(t) \cdots \Delta z_6(t)\right]$, $\Delta Z_2^T(t) = \left[\Delta z_7(t)\ \Delta z_8(t)\ \Delta z_9(t)\right]$ are the output disturbances. The components of the nominal system vector functions in (25) are expressed as follows:

$$\begin{aligned}
\bar{a}_{11}(Z_1) &= z_4(t)c_{z_3} - z_5(t)s_{z_3} \\
\bar{a}_{12}(Z_1) &= z_4(t)s_{z_3} + z_5(t)c_{z_3}, \bar{a}_{13}(Z_1) = z_6(t) \\
\bar{a}_{14}(Z) &= \left[f_x(Z_2) - \bar{f}_{f_x}(Z_1)\right]/M_r \\
\bar{a}_{15}(Z) &= \left[f_y(Z_2) - \bar{f}_{f_y}(Z_1)\right]/M_r \\
\bar{a}_{16}(Z) &= \left[\tau(Z_2) - \bar{\tau}_f(Z_1)\right]/I_r \\
\bar{a}_{2j}(Z) &= \left[-R_j z_{6+j}(t) - K_{b_j}N_j^*\omega_j(t)\right]/L_j \\
\bar{B} &= diag\left\{1/L_j\right\}, \quad j = 1, 2, 3.
\end{aligned} \tag{27}$$

Furthermore, the constraint input is defined as follows:

$$C(U) = \left[c(u_1) \quad c(u_2) \quad c(u_3)\right]^T \tag{28}$$

where $c(u_i) = u_i(t)$, if $|u_i(t)| \le u_c$; $c(u_i) = u_c sign(u_i)$, otherwise. In contrast, the nominal friction force and friction torque of wheel j, integrating viscous, Coulomb, and Stribeck frictions, are modeled as follows:

$$\bar{f}_{f_j}(v_j) = f_{sl}(v_j)\lambda_f(v_j) + f_{st}(f_j)\left[1 - \lambda_f(v_j)\right] \tag{29}$$

$$\bar{\tau}_{f_j}(\omega_j) = \tau_{sl}(\omega_j)\lambda_\tau(\omega_j) + \tau_{st}(\tau_j)\left[1 - \lambda_\tau(\omega_j)\right] \tag{30}$$

where $\lambda_f(v_j) = 1$ as $\left|v_j(t)\right| > \delta_f$ and $\lambda_f(v_j) = 0$, otherwise; $\lambda_\tau(\omega_j) = 1$ as $\left|\omega_j(t)\right| > \delta_\tau$ and $\lambda_\tau(\omega_j) = 0$, otherwise. Here,

$$f_{st}(f_j) = \begin{cases} f_s^+, & \text{as } f_j(t) > f_s^+ > 0 \\ f_j(t), & \text{as } f_s^- \le f_j(t) \le f_s^+ \\ f_s^-, & \text{as } f_j(t) < f_s^- < 0, \end{cases}$$

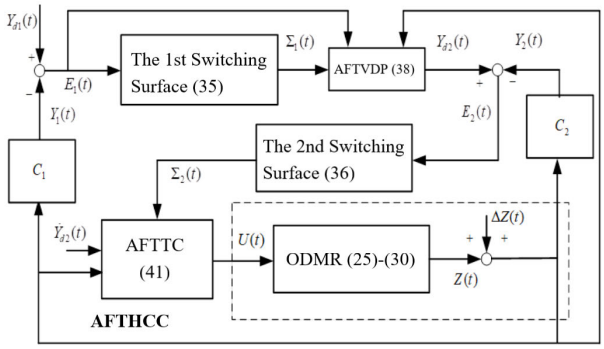


FIGURE 13. Overall control diagram of AFTHCC.

$$\tau_{st}(\tau_j) = \begin{cases} \tau_s^+, & \text{as } \tau_j(t) > \tau_s^+ > 0 \\ \tau_j(t), & \text{as } \tau_s^- \le \tau_j(t) \le \tau_s^+ \\ \tau_s^-, & \text{as } \tau_j(t) < \tau_s^- < 0 \end{cases}$$

$$f_{sl}(v_j) = \begin{cases} f_s^+ - \delta f^+ \left[1 - e^{-|v_j(t)/v^+|}\right] \\ +C_f^+ v_j(t), & \text{as } v_j(t) > 0, \\ f_s^- - \delta f^- \left[1 - e^{-|v_j(t)/v^-|}\right] \\ +C_f^- v_j(t), & \text{as } v_j(t) \le 0, \end{cases}$$

$$\tau_{sl}(\omega_j) = \begin{cases} \tau_s^+ - \delta\tau^+ \left[1 - e^{-|\omega_j(t)/\omega^+|}\right] \\ +C_\tau^+ \omega_j(t), & \text{as } \omega_j(t) > 0, \\ \tau_s^- - \delta\tau^- \left[1 - e^{-|\omega_j(t)/\omega^-|}\right] \\ +C_\tau^- \omega_j(t), & \text{as } \omega_j(t) \le 0. \end{cases}$$

The overall control block diagram of the proposed AFTHCC is shown in Fig. 13 including adaptive finite-time virtual desired pose (AFTVDP) and adaptive finite-time tracking control (AFTTC).

The uncertainties of the indirect mode are given as follows:

$$\begin{aligned} \mathrm{P}_1(Z_1, Y_{d2,eq}, t) = &-\left[D_p + \alpha_1 D_n sign(E_1)^{\alpha_1 - 1}\right] \Delta A_{123}(Z_1, t) \\ &- D_d \left[\Delta G(Z_1, t) + \overline{H}(Z_1) \Delta Y_{d2,eq}(t)\right] \\ &+ D_n \left\|E_1(t)\right\|^{\alpha_1} \dot{E}_{1,0}(t) \end{aligned} \tag{31a}$$

where $\dot{E}_{1,0}(t) = dt[sign(E_1)]/dt|_{E_1=0}$ is the uncertainty caused by the time-derivative of the sign function at zero; $\Delta Y_{d2,eq}(t)$ is the uncertainty of the $Y_{d2,eq}(t)$ in (38b); $\Delta A_{123}(Z_1, t)$ is the uncertainty of the $\overline{A}_{123}(Z_1)$; $\Delta G(Z_1, t)$ is the uncertainty of the $\overline{G}(Z_1)$ in (39a); $\overline{H}(Z_1)$ is given in (39b); and $0 < \beta_1 < 1$. Since $\left\|\dot{E}_{1,0}(t)\right\| \le 2/\Delta t$, where Δt is the sampling time, is bounded, $\|E_1(t)\|^{\alpha_1} \dot{E}_{1,0}(t)$ becomes smaller as $\|E_1(t)\| \approx 0$. Hence, the upper bound of (31a) is assumed as follows:

$$\left\|\mathrm{P}_1(Z_1, Y_{d2,eq}, t)\right\| \le \overline{\rho}_{11} \left\|\Sigma_1(t)\right\|^{\beta_1} + \overline{\rho}_{12} \left\|\Sigma_1(t)\right\|, \quad \forall t \tag{31b}$$

where $0 \le \overline{\rho}_{11}, \overline{\rho}_{12}$ are bounded but unknown and learned by the adaptive law (37). Similarly, the uncertainties of the direct mode and their upper bound are described as follows:

$$\begin{aligned} \mathrm{P}_2(Z, U, t) = &-H_p \left[\Delta A_2(Z, t) + \Delta B(Z_2, t) U_{eq}(t) \right. \\ &\left. + B(Z_2, t) \left(C(U) - U(t)\right) + \Delta \dot{Z}_2(t)\right] \end{aligned} \tag{32a}$$

$$\left\|\mathrm{P}_2(Z, U, t)\right\| \le \overline{\rho}_{21} \left\|\Sigma_2(t)\right\|^{\beta_2} + \overline{\rho}_{22} \left\|\Sigma_2(t)\right\|, \quad \forall t. \tag{32b}$$

Here $0 < \beta_2 < 1, 0 \le \overline{\rho}_{21}, \overline{\rho}_{22}$ are bounded but unknown, and learned by the adaptive law (37). In addition, the uncertain switching gain of the AFTVDP satisfies the inequality:

$$\left\|Y_{d2,sw}^{\dagger}(t) \Delta Y_{d2,sw}(t)\right\| < \mu_1 < 1, \quad \forall t. \tag{33}$$

Here $Y_{d2,sw}^{\dagger}(t) = \left[Y_{d2,sw}^T(t) Y_{d2,sw}(t)\right]^{-1} Y_{d2,sw}^T(t)$ is the pseudo-inverse of the $Y_{d2,sw}(t)$ in (38c). Similarly, the uncertain gain of AFTTC satisfies the following inequality:

$$\left\|H_p \Delta B(Z, t) \left(H_p \overline{B}\right)^{-1}\right\| \le \mu_2 < 1, \quad \forall t. \tag{34}$$

The first switching surface for the indirect mode is as

$$\Sigma_1(t) = D_d \dot{E}_1(t) + D_p E_1(t) + D_n sign(E_1)^{\alpha_1}. \tag{35}$$

Here $\dot{E}_1(t) = \dot{Y}_{d1}(t) - \overline{A}_{123}(Z_1)$, $\overline{A}_{123}(Z_1) = [\overline{a}_{11}(Z_1)\ \overline{a}_{12}(Z_1)\ \overline{a}_{13}(Z_1)]^T$, $\Sigma_1(t) \in \Re^3$, $D_d, D_p, D_n > 0 \in \Re^{3\times 3}$ are constant diagonal matrices, D_d is nonsingular, and $0 < \alpha_1 < 1$. Similarly, the second switching surface for direct mode is as follows:

$$\Sigma_2(t) = H_p E_2(t) + H_i \int_{t_0}^{t} E_2(\tau) d\tau + H_n \int_{t_0}^{t} sign(E_2)^{\alpha_2} d\tau. \tag{36}$$

Here $\Sigma_2(t) \in \Re^3, H_p, H_i, H_n > 0 \in \Re^{3\times 3}$ are constant diagonal matrices, H_p is nonsingular, and $0 < \alpha_2 < 1$.

The adaptive laws for two unknown coefficients of the upper bounds of the uncertainties in indirect and direct modes (i.e., $\overline{\rho}_{ij}, i, j = 1, 2$ in (31b) and (32b)) are designed as follows:

$$\dot{\hat{\rho}}_{ij}(t) = \begin{cases} \hat{\rho}_{ij}(t) + \rho_{ij,l}(t), & \text{if } \rho_{ij}(t) < 0, \\ \hat{\rho}_{ij}(t), & \text{if } 0 \le \rho_{ij}(t) \le \overline{\rho}_{ij,M}, \\ \hat{\rho}_{ij}(t) + \rho_{ij,u}(t), & \text{if } \overline{\rho}_{ij,M} < \rho_{ij}(t). \end{cases} \tag{37}$$

Here $\hat{\rho}_{ij}(t) = \lambda_{ij} \left\|\Sigma_i(t)\right\|^{j+(2-j)\beta_i} \left[1 - \delta_{ij} \rho_{ij}(t)\right]$, $\overline{\rho}_{ij,M} = \max\{\overline{\rho}_{ij}\}$, $i, j = 1, 2$ are assumed to be known; $\lambda_{ij}, \delta_{ij} > 0$ are denoted as the learning rate and e-modification rate, respectively; $\rho_{ij,l}(t) = \lambda_{ij} \left\|\Sigma_i(t)\right\|^{j+(2-j)\beta_i} \rho_{ij}(t)$, $\rho_{ij,u}(t) = \lambda_{ij} \left\|\Sigma_i(t)\right\|^{j+(2-j)\beta_i} \left(\rho_{ij}(t) - \overline{\rho}_{ij,M}\right)$; and $\Sigma_i(t)$ denotes the i-th switching surface.

Then the AFTVDP is designed as follows:

$$Y_{d2}(t) = Y_{d2,eq}(t) + Y_{d2,sw}(t) \tag{38a}$$

$$\begin{aligned} Y_{d2,eq}(t) = &\left[D_d \overline{H}(Z_1)\right]^{-1} \left\{D_d \left[\ddot{Y}_{d1}(t) - \overline{G}(Z_1)\right] \right. \\ &+ \left[D_p + \alpha_1 D_n sign(E_1)^{\alpha_1 - 1}\right] \dot{E}_1(t) \\ &\left. + \left[\rho_{11}(t) \left\|\Sigma_1(t)\right\|^{\beta_1 - 1} + \rho_{12}(t)\right] \Sigma_1(t)\right\} \end{aligned} \tag{38b}$$

$$Y_{d2,sw}(t) = \left[D_d\overline{H}(Z_1)\right]^{-1}\{\mathrm{K}_{11}(\|\Sigma_1\|)\Sigma_1(t) + \mathrm{K}_{12}(\|\Sigma_1\|) \cdot\Sigma_1(t)\|\Sigma_1(t)\|^{\beta_1-1}/\left(\|\Sigma_1(t)\|^{1+\beta_1}+\varepsilon_1\right)\}/(1-\mu_1) \tag{38c}$$

where ε_1 is a small positive constant, the components of $\overline{G}(Z_1) = \left[\overline{g}_i(Z_1)\right]\big|_{i=1,2,3}$, and $\overline{H}(Z_1) = \left[\overline{h}_{ij}(Z_1)\right]\big|_{i,j=1,2,3}$ are given as follows:

$$\begin{aligned}
\overline{g}_1(Z_1) &= -z_4(t)z_6(t)s_{z_3} - z_5(t)z_6(t)c_{z_3} \\
&\quad -\left[c_{z_3}\overline{f}_{f_x}(Z_1) - s_{z_3}\overline{f}_{f_y}(Z_1)\right]/M_r \\
\overline{g}_2(Z_1) &= z_4(t)z_6(t)c_{z_3} - z_5(t)z_6(t)s_{z_3} \\
&\quad -\left[s_{z_3}\overline{f}_{f_x}(Z_1) + c_{z_3}\overline{f}_{f_y}(Z_1)\right]/M_r \\
\overline{g}_3(Z_1) &= -\overline{\tau}_f(Z_1)/I_r \qquad (39a) \\
\overline{h}_{1j} &= N_jK_{t_j}\left(c_{z_3}s_j - s_{z_3}c_j\right)/(M_r r_w), \\
\overline{h}_{2j} &= N_jK_{t_j}\left(s_{z_3}s_j + c_{z_3}c_j\right)/(M_r r_w), \\
\overline{h}_{3j} &= l_wN_jK_{t_j}/(I_r r_w), \quad j = 1, 2, 3. \qquad (39b)
\end{aligned}$$

To achieve the finite-time to zero switching surface and then tracking error, the nonlinear switching gains [38], [39] are designed as follows:

$$\mathrm{K}_{11}(\|\Sigma_1\|) = \left[\kappa_{111} + \kappa_{112}\|\Sigma_1(t)\|^{\beta_1-1}\right]F_{11} > 0 \tag{40a}$$

$$\mathrm{K}_{12}(\|\Sigma_1\|) = \left[\kappa_{121} + \kappa_{122}\|\Sigma_1(t)\|^{2\beta_1}\right]F_{12} > 0 \tag{40b}$$

where $F_{1j} \geq I_3, \kappa_{1ij} > 0, i, j = 1, 2$. In addition, $sign(E_1)^{\alpha_1-1}$ is a diagonal matrix. The AFTTC for the system (25)-(26) is designed as follows:

$$U(t) = U_{eq}(t) + U_{sw}(t) \tag{41a}$$

$$\begin{aligned}
U_{eq}(t) &= \left(H_p\overline{B}\right)^{-1}\left\{H_p\dot{Y}_{d2}(t) + H_iE_2(t) + H_nsign(E_2)^{\alpha_2}\right. \\
&\quad \left.+\left[\rho_{21}(t)\|\Sigma_2(t)\|^{\beta_2-1} + \rho_{22}(t)\right]\Sigma_2(t)\right\} \qquad (41b) \\
U_{sw}(t) &= \left(H_p\overline{B}\right)^{-1}\{\mathrm{K}_{21}(\|\Sigma_2\|)\Sigma_2(t) + \mathrm{K}_{22}(\|\Sigma_2\|) \\
&\quad \cdot\Sigma_2(t)\|\Sigma_2(t)\|^{\beta_2-1}/\left(\|\Sigma_2(t)\|^{1+\beta_2}+\varepsilon_2\right)\}/(1-\mu_2)\,. \qquad (41c)
\end{aligned}$$

Similarly, ε_2 is a small positive constant and the diagonal nonlinear switching gains are designed as follows:

$$\mathrm{K}_{21}(\|\Sigma_2\|) = \left[\kappa_{211} + \kappa_{212}\|\Sigma_2(t)\|^{\beta_2-1}\right]F_{21} > 0 \tag{42a}$$

$$\mathrm{K}_{22}(\|\Sigma_2\|) = \left[\kappa_{221} + \kappa_{222}\|\Sigma_2(t)\|^{2\beta_2}\right]F_{22} > 0 \tag{42b}$$

where $F_{2j} \geq I_3, \kappa_{2ij} > 0, i, j = 1, 2$. The stability analysis can refer to [38], [39].

VI. INTERACTIONS BETWEEN SPECIFIC HUAMN AND ODMR

At the outset, the specific human is at the pose $(7m, 1m, 180°)$, i.e., the human face is back to the ODMR. The corresponding human robot interactions with the control parameters in Table 12 are depicted by the important snapshots from the ODMR as shown in Fig. 14.

TABLE 12. Control parametrs of the AFTHCC.

Parameter	Value
$\alpha_1, \beta_1,\ D_d, D_p, D_n$	$0.85, 0.15,\ diag\{0.5, 0.5, 1.2\},\ 140I_3, 6.25\times10^{-3}I_3$
$\alpha_2, \beta_2,\ H_p, H_i, H_n$	$0.85, 0.45, I_3,\ 0.25I_3,\ 2.5\times10^{-3}I_3$
$\mathrm{K}_{11}(\|\Sigma_1\|), \mathrm{K}_{12}(\|\Sigma_1\|)$	$\left[46+6\times10^{-5}\|\Sigma_1(t)\|^{\beta_1-1}\right]I_3, \left[0.05+0.005\|\Sigma_1(t)\|^{2\beta_1}\right]I_3$
$\mathrm{K}_{21}(\|\Sigma_2\|), \mathrm{K}_{22}(\|\Sigma_2\|)$	$\left[0.5+0.005\|\Sigma_2(t)\|^{\beta_2-1}\right]I_3, \left[0.2+0.005\|\Sigma_2(t)\|^{2\beta_2}\right]I_3$
$\varepsilon_1, \varepsilon_2, \mu_1, \mu_2$	0.02, 0.02, 0.1, 0.1
$\lambda_{11}, \lambda_{12}, \delta_{11}, \delta_{12}, \overline{\rho}_{11,M}, \overline{\rho}_{12,M}$	$4.41\times10^{-4}, 4.41\times10^{-4}$, 0.084, 0.096, 2, 2
$\lambda_{21}, \lambda_{22}, \delta_{21}, \delta_{22}, \overline{\rho}_{21,M}, \overline{\rho}_{22,M}$	$4.41\times10^{-4}, 4.41\times10^{-4}$, 0.084, 0.096, 5, 5

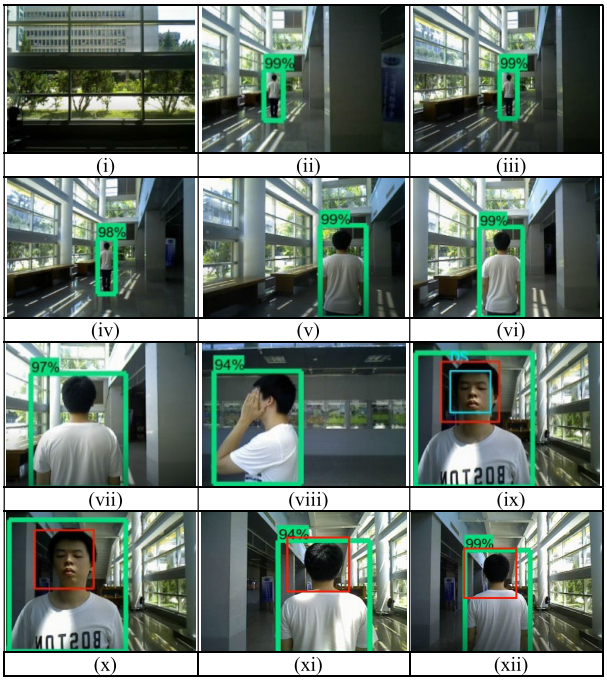


FIGURE 14. Important snapshots from the ODMR. (i) no human is detected; ODMR turns 90°, (ii) human is detected by SSD, (iii) the human's pose is computed, (iv) ODMR is controlled to the desired orientation ϕ_d, (v) ODMR is controlled to the 2.5∼3m between them, (vi) ODMR is controlled to the desired orientation ϕ_d, (vii) ODMR is controlled the pose of 1*m* and 0°; he is not SH, (viii) ODMR turns to another pose 1*m* and 90°; the SH raises his right hand to occlude the recognition, (ix) he is the SH; the KCF tracker initiates, (x) the KCF works to track the SH; the FaceNet does not work, (xi) the SH turns 180° to execute the human following, (xii) the ODMR tacks the SH until less than 1.5*m* and then stops.

Furthermore, the control response using the proposed AFTHCC is presented in Fig. 15, including (a) trajectory tracking in XY plane, (b) 2D pose, (c) control input, (d) switching surfaces of the indirect and direct mode, (e) tracking errors of the indirect and direct modes, (f) the estimated coefficients for the upper bounds of the indirect and direct modes' uncertainties [18].

The responses are illustrative as follows: (i) At the very inception, the SSD detects the human in the first field of view (FOV). Since no human is detected, the ODMR turns 90°. (ii) Even the background lighting is not uniform, the human over $7m$ is detected by the SSD. Then the 2D

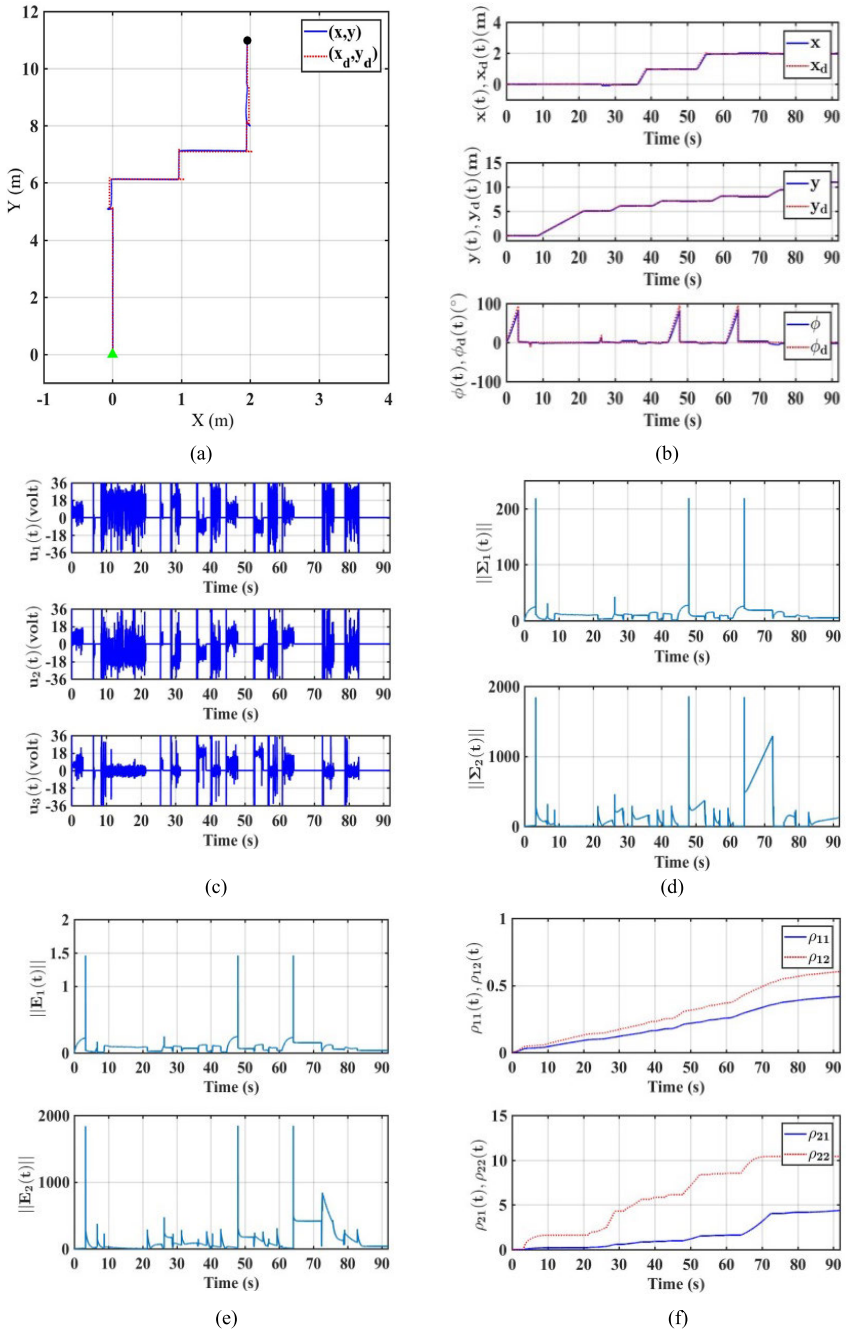


FIGURE 15. Response of the motion control signals. (a) XY position, (b) 2D pose, (c) control input, (d) switching surface norm, (e) tracking error norm, (f) learning coefficient.

pose between them (i.e., (24)) is computed. (iii) Based on the IB-AFTHCC, the ODMR is controlled to the desired orientation $\phi_d(t)$. Human is at the central position of FOV. (iv) The 2D pose (x_d, y_d, ϕ_d) between them is computed. (v) The ODMR is controlled to the desired position (x_d, y_d) i.e., $2.5\sim3m$ between them, by the IB-AFTHCC. (vi) The 2D pose between them is computed by the depth image. The ODMR is controlled to the desired orientation $\phi_d(t)$ by the IB-AFTHCC such that human is at the central position of FOV. (vii) The vertical position of the ODMR (about $1m$ and $0°$ between them) is also controlled by the IB-AFTHCC. Since the SH is not recognized, the ODMR is controlled to another desired pose (about $1m$ and $90°$ between them). (viii) The FaceNet is applied to recognize the SH with his right hand to occlude the recognition because the SH is probably recognized at $90°$ orientation (cf. Table 5). Since the SH is not recognized, the ODMR is controlled to another desired pose (about $1m$ and $180°$ between them). (ix) Since he is the SH, the KCF tracker initiates (i.e., the red rectangle denotes the tracking of the SH). (x) Since the KCF works to track the SH, the FaceNet doesn't work; hence, the blue rectangle in the snapshot (ix) of Fig. 14 disappears. (xi) The SH turns $180°$ to execute the task

of human following; the green and red rectangles still remain unchanged. (xii) As the distance larger than 1.5m, the ODMR tacks the SH until less than 1.5m, and then stops.

The maximum position and orientation errors are respectively about 4 cm and 5° which are excellent for the motion control task. Finally, the planned human-robot interactions are successfully accomplished. The corresponding experimental video can refer to the URL: https://youtu.be/FF-cf7nv5Uo.

VII. CONCLUSION

The deep learning approach using the SSD-FN-KCF is developed such that a specific human (SH) is identified and tracked to execute the required interactions. The green, blue, and red rectangles are the outputs of the SSD, FN, and KCF, respectively. Besides the image-based adaptive finite-time hierarchical constraint control (IB-AFTHCC) executes the planned poses, three techniques are integrated to enhance each method's advantages and avoid their drawbacks. The SSD using the RGB-D camera with the resolution of 320×240 can detect humans up to 8m. It is a favorable result as compared to other state-of-the-art methods [3]–[8]. Due to the low resolution of RGB for FaceNet, only up to 1.25m can successfully recognize the specific face (or human). Even though, the larger pose variations including the occlusion of human and the large change of lighting orientation still confirm the robustness of face recognition. To reduce the repeated face recognition, the KCF not only accelerates the processing time but also extends the tracking distance of the detected SH to achieve the satisfactory task of HRI. Furthermore, the advantages of ODMR (i.e., simultaneous translation and rotation) and the IB-AFTHCC fulfill the HRIs, e.g., search, detect, track the (specific) human, and human following. One of our future studies is to extended dynamic face emotion recognition result up to 3.5m using stereo camera and very deep CNN method [40], [41].

REFERENCES

[1] W. Wang, R. Li, Y. Chen, Z. M. Diekel, and Y. Jia, "Facilitating human–robot collaborative tasks by Teaching-Learning-Collaboration from human demonstrations," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 2, pp. 640–653, Apr. 2019.

[2] Z. Li, C. Deng, and K. Zhao, "Human-cooperative control of a wearable walking exoskeleton for enhancing climbing stair activities," *IEEE Trans. Ind. Electron.*, vol. 67, no. 4, pp. 3086–3095, Apr. 2020.

[3] W. Chung, H. Kim, Y. Yoo, C.-B. Moon, and J. Park, "The detection and following of human legs through inductive approaches for a mobile robot with a single laser range finder," *IEEE Trans. Ind. Electron.*, vol. 59, no. 8, pp. 3156–3166, Aug. 2012.

[4] L. Li, S. Yan, X. Yu, Y. Kee Tan, and H. Li, "Robust multiperson detection and tracking for mobile service and social robots," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 42, no. 5, pp. 1398–1412, Oct. 2012.

[5] S. Wu, S. Wang, R. Laganiere, C. Liu, H.-S. Wong, and Y. Xu, "Exploiting target data to learn deep convolutional networks for scene-adapted human detection," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1418–1432, Mar. 2018.

[6] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 973–986, Apr. 2018.

[7] J. Nie, L. Huang, W. Zhang, G. Wei, and Z. Wei, "Deep feature ranking for person re-identification," *IEEE Access*, vol. 7, pp. 15007–15017, 2019.

[8] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[9] L. Chen, M. Li, W. Su, M. Wu, K. Hirota, and W. Pedrycz, "Adaptive feature selection-based AdaBoost-KNN with direct optimization for dynamic emotion recognition in human–robot interaction," *IEEE Trans. Emerg. Topics Comput. Intell.*, to be published.

[10] J. Li, Z. Li, Y. Feng, Y. Liu, and G. Shi, "Development of a human–robot hybrid intelligent system based on brain teleoperation and deep learning SLAM," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 4, pp. 1664–1674, Oct. 2019.

[11] S. K. Biswas and P. Milanfar, "One shot detection with laplacian object and fast matrix cosine similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 546–562, Mar. 2016.

[12] S. Duffner and C. Garcia, "Fast pixelwise adaptive visual tracking of non-rigid objects," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2368–2380, May 2017.

[13] Y. Liu, Q. Wang, Y. Zhuang, and H. Hu, "A novel trail detection and scene understanding framework for a quadrotor UAV with monocular vision," *IEEE Sensors J.*, vol. 17, no. 20, pp. 6778–6787, Oct. 2017.

[14] W. Zhang, Z. Wang, X. Liu, H. Sun, J. Zhou, Y. Liu, and W. Gong, "Deep learning-based real-time fine-grained pedestrian recognition using stream processing," *IET Intell. Transp. Syst.*, vol. 12, no. 7, pp. 602–609, Sep. 2018.

[15] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.

[16] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3355–3362, Oct. 2018.

[17] C.-L. Hwang and Y. Lee, "Tracking design of an omni-direction autonomous ground vehicle by hierarchical enhancement using fuzzy second-order variable structure control," *J. Dyn. Syst., Meas., Control*, vol. 140, no. 9, pp. 1–11, Apr. 2018.

[18] C.-L. Hwang, W.-H. Hung, and Y. Lee, "Tracking design of omnidirectional drive service robot using hierarchical adaptive finite-time control," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Washington, DC, USA, Oct. 2018, pp. 5680–5685.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[20] N. Werghi, C. Tortorici, S. Berretti, and A. Del Bimbo, "Boosting 3D LBP-based face recognition by fusing shape and texture descriptors on the mesh," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 5, pp. 964–979, May 2016.

[21] A. Rikhtegar, M. Pooyan, and M. T. Manzuri-Shalmani, "Genetic algorithm-optimised structure of convolutional neural network for face recognition applications," *IET Comput. Vis.*, vol. 10, no. 6, pp. 559–566, Sep. 2016.

[22] M. A. Abuzneid and A. Mahmood, "Enhanced human face recognition using LBPH descriptor, multi-KNN, and back-propagation neural network," *IEEE Access*, vol. 6, pp. 20641–20651, 2018.

[23] H. B. Abebe and C.-L. Hwang, "RGB-D face recognition using LBP with suitable feature dimension of depth image," *IET Cyber-Phys. Syst., Theory Appl.*, vol. 4, no. 3, pp. 189–197, Sep. 2019.

[24] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[25] J. Leng and Y. Liu, "Real-time RGB-D visual tracking with scale estimation and occlusion handling," *IEEE Access*, vol. 6, pp. 24256–24263, 2018.

[26] Q. Liu, G. Hu, and M. M. Islam, "Fast visual tracking with robustifying kernelized correlation filters," *IEEE Access*, vol. 6, pp. 43302–43314, 2018.

[27] S. P. Bharati, Y. Wu, Y. Sui, C. Padgett, and G. Wang, "Real-time obstacle detection and tracking for Sense-and-Avoid mechanism in UAVs," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 2, pp. 185–197, Jun. 2018.

[28] W. Tian, L. Chen, K. Zou, and M. Lauer, "Vehicle tracking at nighttime by kernelized experts with channel-wise and temporal reliability estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 10, pp. 3159–3169, Oct. 2018.

[29] F. Fabrizio and A. De Luca, "Real-time computation of distance to dynamic obstacles with multiple depth sensors," *IEEE Robot. Autom. Lett.*, vol. 2, no. 1, pp. 56–63, Jan. 2017.

[30] P. Long, W. Liu, and J. Pan, "Deep-learned collision avoidance policy for distributed multiagent navigation," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 656–663, Apr. 2017.
[31] C.-L. Hwang and H.-H. Huang, "Experimental validation of a car-like automated guided vehicle with trajectory tracking, obstacle avoidance, and target approach," in *Proc. 43rd Annu. Conf. IEEE Ind. Electron. Soc (IECON)*, Beijing China, Oct. 2017, pp. 2858–2863.
[32] M. Graf Plessen, D. Bernardini, H. Esen, and A. Bemporad, "Spatial-based predictive control and geometric corridor planning for adaptive cruise control coupled with obstacle avoidance," *IEEE Trans. Control Syst. Technol.*, vol. 26, no. 1, pp. 38–50, Jan. 2018.
[33] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1208–1221, Sep. 2004.
[34] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele, "Monocular visual scene understanding: Understanding multi-object traffic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 882–897, Apr. 2013.
[35] A. S. Rao, J. Gubbi, S. Marusic, and M. Palaniswami, "Crowd event detection on optical flow manifolds," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1524–1537, Jul. 2016.
[36] L. Chen, M. Wu, M. Zhou, Z. Liu, J. She, and K. Hirota, "Dynamic emotion understanding in human–robot interaction based on two-layer fuzzy SVR-TS model," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 50, no. 2, pp. 490–501, Feb. 2020.
[37] J. Li, X. Mei, D. Prokhorov, and D. Tao, "Deep neural network for structural prediction and lane detection in traffic scene," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 690–703, Mar. 2017.
[38] C.-L. Hwang and Y.-H. Chen, "Fuzzy fixed-time learning control with saturated input, nonlinear switching surface and switching gain to achieve null tracking error," *IEEE Trans. Fuzzy Syst.*, to be published.
[39] C.-L. Hwang and B.-S. Chen, "Adaptive finite-time saturated tracking control for a class of partially known robots," *IEEE Trans. Syst., Man, Cybern. Syst.*, to be published.
[40] C.-K. Lee, C.-L. Hwang, and F. Wu, "Very deep CNN stereo camera based dynamic face emotion recognition and its application," in *Proc. IEEE WCCI*, Glasgow, Scotland, 2020, pp. 1–6.
[41] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2263–2276, Dec. 2016.

CHIH-LYANG HWANG (Senior Member, IEEE) received the B.E. degree in aeronautical engineering from Tamkang University, Taipei, Taiwan, in 1981, and the M.E. and Ph.D. degrees in mechanical engineering from the Tatung Institute of Technology, Taipei, in 1986 and 1990, respectively.

From 1990 to 2006, he was with the Department of Mechanical Engineering, Tatung Institute of Technology, where he was involved in teaching and research in the area of servo, control, and control of manufacturing systems and robotic systems, and was also a Professor of mechanical engineering, from 1996 to 2006. From 1998 to 1999, he was a Research Scholar at the George W. Woodruf School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA. From 2006 to 2011, he was a Professor with the Department of Electrical Engineering, Tamkang University. Since 2011, he has been a Professor with the Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei. From August 2016 to July 2017, he was a Visiting Scholar at the Electrical and Computer Engineering Department, Auburn University, Auburn AL, USA. He is the author or coauthor of approximately 166 journal and conference papers in the related fields. His current research interests include robotics, fuzzy neural modeling, classification and control, finite-time control, (distributed) visual or wireless localization or navigation systems, nonlinear multiagent systems, remote control of UAV, face and speech emotion recognition, and human–robot interaction.

Dr. Hwang was a recipient of the Excellent and Outstanding Research Awards from the National Taiwan University of Science and Technology, in 2018 and 2019, respectively.

DING-SHENG WANG received the B.E. and M.E. degrees in electrical engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2015 and 2019, respectively. He currently works at Taiwan Semiconductor Manufacturing Company. His research interests include robotics and computation intelligence.

FAN-CHEN WENG received the B.E. and M.E. degrees in electrical engineering from the National Taiwan Ocean University, Keelung, Taiwan, in 2017, and the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2019. He is currently in a military service at Taiwan. His research interests include robotics, voice recognition, and computation intelligence.

SHENG-LIN LAI received the B.E. degree from the National Taipei University of Technology, in 2016, and the M.E. degree in electrical engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2019. He is currently in a military service at Taiwan. His research interests include face expression recognition, robotics, and computation intelligence.