IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 October 2022, accepted 10 November 2022, date of publication 12 December 2022, date of current version 19 April 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3228825

APPLIED RESEARCH

Human–Robot Collaboration Using Sequential-Recurrent-Convolution-Network-Based Dynamic Face Emotion and Wireless Speech Command Recognitions

CHIH-LYANG HWANG, (Senior Member, IEEE), YU-CHEN DENG, AND SHIH-EN PU
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 Projects from the Ministry of Science and Technology at Taiwan under Grant NSTC-111-2622-E-011-017 and Grant NSTC-111-2221-E-007-127.

ABSTRACT The proposed sequential recurrent convolution network (SRCN) includes two parts: one convolution neural network (CNN) and a sequence of long short-term memory (LSTM) models. The CNN is to achieve the feature vector of face emotion or speech command. Then, a sequence of LSTM models with the shared weight reflects a sequence of inputs provided by a (pre-trained) CNN with a sequence of input sub-images or spectrograms corresponding to face emotion and speech command, respectively. Simply put, one SRCN for dynamic face emotion recognition (SRCN-DFER) and another SRCN for wireless speech command recognition (SRCN-WSCR) are developed. The proposed approach not only effectively tackles the recognitions of dynamic mapping of face emotion and speech command with average generalized recognition rate of 98% and 96.7% but also prevents the overfitting problem in a noisy environment. The comparisons among mono and stereo visions, Deep CNN, and ResNet50 confirm the superiority of the proposed SRCN-DFER. The comparisons among SRCN-WSCR with noise-free data, SRCN-WSCR with noisy data, and multiclass support vector machine validate its robustness. Finally, the human-robot collaboration (HRC) using our developed omnidirectional service robot, including human and face detections, trajectory tracking by the previously designed adaptive stratified finite-time saturated control, face emotion and speech command recognitions, and music play, validates the effectiveness, feasibility, and robustness of the proposed method.

INDEX TERMS Human–robot collaboration, CNN, LSTM, human and face detection, dynamic face emotion recognition, wireless speech command recognition, omnidirectional service robot, visual searching and tracking, adaptive stratified finite-time saturated control.

I. INTRODUCTION

Recently, different kinds of robots have been developed to fulfill human-robot collaborations. Some representative and outstanding works are reviewed as follows. A robot that can understand and express emotions in voice, gesture, and gait by a controller trained only on voice is developed such that the robot can recognize happiness, sadness, and fear in a completely different modality [1]. A humanoid robot's visual imitation of 3-D motion of a human is developed by a neural-network-based inverse kinematics [2] or support vector machine for the classification of 11 low-body postures [3]. In [4], the suggested robot Mortimer including social behaviors can increase engagement and social presence; in addition, the effect of extending weekly collocated musical improvisation sessions is investigated by making Mortimer an active member of the participant's virtual social network. In [5], specific human following through machine learning of SSD-FN-KCF is developed. In [6], musical robots are designed to control those dynamics, articulation, and tempo to give

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

the audience an experience as compared with listening to a professional human musician. Wolfe et al. [7] develop a singing robot platform that could interact with surrounding humans by communicating through song, musical, nonlinguistic utterances to evoke emotional responses in humans. An auxiliary online diagnosor using the Bayesian decision theory provides not only a collision identification for human-collaborative robots but also a confidence index to represent their reliability [8]. Additionally, many researchers have been committed to the robots with the ability to detect and identify human emotions, and then apply this information to guide their own behaviors, which are called affective intelligent robots [9], [10], [11], [12], [13]. Since face emotion plays the most important role, the issue of making affective intelligent robots with accurate and real-time face emotion recognition becomes a challenging task. Many related articles examine the classification of facial expression image into several typical classes: angry, disgusted, fearful, happy, surprised, and sad [14].

It is known that the detection of human emotions from facial emotions is crucial for social interaction. The proposed sequential recurrent convolution network (SRCN) improves the corresponding drawbacks and simultaneously enhances its performance, and then applies to human-robot collaboration (HRC) task. From the outset, the stereo camera on the omnidirectional service robot (ODSR) is planned to search and detect the human by Faster R-CNN [15]. If a face candidate exists, the detected face using Haar Cascade feature descriptor is cropped as a suitable size to recognize his/her face emotions. If not, the strategy to approach the above pose region is achieved by the stereo vision based localization [16], [17] and an stratified finite-time saturated control [18], [19]. Subsequently, the facial emotion is recognized by the SRCN-DFER. Stereo camera not only estimates the 3D position up to 20*m,* but also improves the recognition rate since the use of left and right cameras increases FOV to achieve a better recognition [16]. A dynamic recognition rate for video to indicate the stabilized facial emotion with a specific time interval is also defined to meet the requirement of HRC.

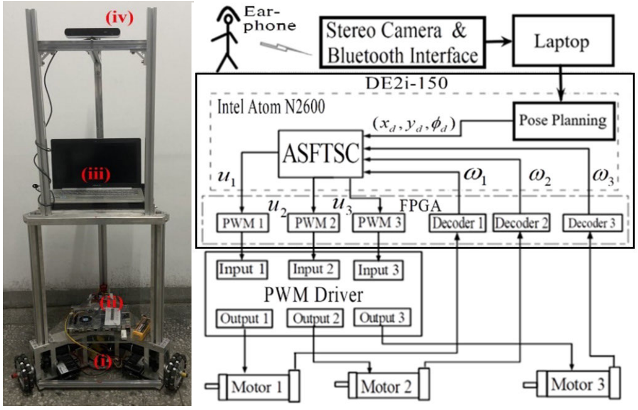
In contrast, another SRCN for wireless speech command recognition (SRCN-WSCR) is developed to deal with more complex HRC task. With this, eight speech commands from Google Speech Commands Dataset are employed to train and verify the SRCN-WSCR. At the outset, the sampled speech command is transferred into the frequency domain signal by Short-Time Fourier Transform (STFT) with suitable window length and hop length. Multiplying the power spectrum of STFT signal by Mel filter matrix obtains the logarithm of Mel-Spectrogram [19], which is set as the input signal of SRCN-WSCR. Since the video sequence for facial emotion is limited, the large amount of facial emotion images is applied to train the CNN, which is the first part of SRCN. After that, these weights in CNN without fully and softmax layers are assigned as the partial initial weights in the SRCN-DFER. The other initial weight for a stack of LSTMs is set as a small random number. In contrast, eight designed speech commands have many dynamic files, its pre-trained CNN is not required. In summary, the proposed approaches not only effectively tackles the recognitions of dynamic mapping of face emotion and speech command, but also prevents the overfitting problem in the presence of noises. Finally, the HRC by omnidirectional service robot, including human and face detections, trajectory tracking using adaptive stratified finite-time saturated control, face emotion and speech command recognitions, and music play, validates the effectiveness and robustness of our method.

The salient contributions of this article are summarized as follows. (i) Two novel sequential recurrent convolution networks (SRCNs): one for the dynamic face emotion recognition (SRCN-DFER) in Algorithm 1 and another for the wireless speech command recognition (SRCN-WSCR) in Algorithm 2, are developed. (ii) The average recognition rate of SRCN-DFER is 98%. It is superior to some previous studies (e.g., [11], [12], [13], [20], [21], [22], [23], [24], [25], [26], [27], [28]). Additionally, the comparisons among SRCN-DFER, DCNN, and ResNet50 validate the superiority of the proposed SRCN-DFER. The comparisons among SRCN-DFER, LRCN [29], and 3D-CNN [30] further confirm the state-of-the-art performance. (iii) After the preprocessing of the speech command, SRCN-WSCR is similar to SRCN-DFER. Furthermore, the confusion matrices with background noises and without background noise confirm its superiority in comparison to some previous research (e.g., multiclass SVM [19]). (iv) The implementation of human-robot collaboration confirms the practicality and feasibility of the proposed method.

II. RELATED WORK

As a pattern recognition task, there are plenty of classification methods that can be adopted to classify facial emotions [11], [12], [13], [20], [21], [22], [23], [24], [25], [26], [27], [28]. In [11] and [12], three- and two-layer fuzzy support vector regression-Takagi-Sugeno models are suggested for the emotion understanding in human-robot-interaction (HRI) task, e.g., the drink reflecting different emotions, human following [5]. Their average video-based recognition rates for different genders, provinces, and ages are ordinary. The aims of [13] is to make good use of the CNN's potential performance in avoiding local optima and speeding up the convergence by the hybrid genetic algorithm with optimal initial population, in such a way that it realizes deep and global emotion understanding in HRI. Nonetheless, its average video-based recognition rate is usual. In [20], multi-modal recurrent attention networks learn spatiotemporal attention volumes to robustly recognize the facial expression. Besides the sequent RGB images, the depth and thermal sequences are also required. In [21], a deep learning framework based on the hybrid of 3D conditional generative adversarial network and two-level attention bidirectional long short-term memory network has been proposed for robust driver drowsiness recognition. However, the averaging recognition rate in different situations is only acceptable. In [22], a 3D-CNN is first designed to capture subtle spatiotemporal changes that may occur on the face, and a Conv-LSTM network is then designed to learn semantic information by taking into account longer spatiotemporal dependencies. Although the recognized result is acceptable, the proposed scheme is complex. A two-branch disentangled generative adversarial network disentangles expressional information from other unrelated facial attributes [23]. Although the average image-based recognition rate for the datasets of CK+, TFEID, and RaFd is excellent, its generalized recognition is poor. In [24], a correlation-based graph convolutional network for automatic emotion recognition (ARE) is developed, which can comprehensively consider the correlation of the intra-class and inter-class videos for feature learning and information fusion. However, the average recognition rate of ARE is normal and there are large variations for different datasets. In [25], modeling pose variations in facial images to boost the performance of face emotion recognition is achieved by an end-to-end weakly supervised approach. However, its average recognition rate and generalization are not excellent. In [26], the Learnable Graph Inception Network, that jointly learns to recognize emotion and identify the underlying graph structure in the dynamic data, is developed. It possesses satisfactory average recognition for RML, eNTERFACE, and RAVDESS datasets. In [27], event-cameras can capture motion at millisecond-rates, work under challenging conditions like low illumination and understand human reactions by only observing facial expressions. Even a combination of CNN and Bi-LSTM for dealing with face emotion recognition [28] failed to accomplish a satisfactory performance due to a lack of effective data sets for training. Besides the above researches, a multimodal fusion framework for noncontact heart rate (HR) estimation, including the feature representation maps from facial visible-light and thermal infrared videos, a temporal information-aware HR feature extraction network for encoding discriminative spatiotemporal information is accomplished [31]. It indicates that face recognition can be adopted for different applications [9], [10], [11], [12], [13].

Recently, the dynamic mapping for machine learning, e.g., the combination of CNN with LSTM [20], [22], [28], [29], [31], [32], [33], [34], [35], [36], is effective for the task with sequential relationship: face emotion recognition [20], [22], [28], visual recognition and description [29], human activity recognition [32], emotion expression with fact transfer for video description [33], real-time health monitoring for machine [34], solar irradiance forecasting [35], welding defect areas localization [36]. Besides video captioning with emotion expression [33], a cognitive load estimation from speech commands focused on human-robot interaction to simulated aircraft is considered by Vukovic et al. [37]. By using voice assistants, users are able to control smart homes via speech commands [38]. To boot robust recognition of speech command, the noisy training data [25] is employed to train and validate the SRCN-WSCR.



(a) Photograph. (b) Block diagram.

FIGURE 1. The proposed ODSR.

III. EXPERIMENTAL SETUP AND PROBLEM DESCRIPTION

A. EXPERIMENTAL SETUP

The experimental setup of the omnidirectional service robot (ODSR) in Fig. 1(a) consists of the following five parts [19]: (i) three dc servomotors, (ii) one motion control module, (iii) a laptop for image processing, (iv) a stereo camera system, and (v) a Bluetooth earphone. The adaptive stratified finite-time saturated control (ASFTSC) $u_i, i = 1, 2, 3$ [18], [19] is computed in the Intel® Atom N2600. Afterwards, the signal is transformed into the pulse width modulation (PWM) using Field Programming Gate Array (FPGA) to drive the servomotors. Then the motor velocity $\omega_i, i = 1, 2, 3$ is achieved by the encoder using FPGA (cf. Fig. 1(b)). Furthermore, the specifications of Intel Zed stereo camera system are as follows: (i) resolution and sampling rate: WVGA(1334×376), 720p(2560×720),1080p (3840×1080),2.2k(4416×1242) :100, 60, 30, 15fps; (ii) field of view in H-V-D planes: 90°, 60°, and 110°; (iii) depth: $0.5\sim20m$; (iv) power via USB: $5V/380mA$; (v) size: $175 \times 30 \times 33mm$;(vi) weight: 159g. The laptop for image processing is the MSI-GF63 computer: (i) Intel Core i7-10750H, 2.6G, (ii) GPU with NVIDIA GTX 1650 Ti, 2GB. On the other hand, the important specification of SONY WI-C300 Bluetooth earphone is given as follows: (i) Bluetooth version: 4.2, (ii) range: $10m$, (iii) sampling rate: 16 bits, 16000 Hz, (iv) battery life: 8hrs. Finally, the block diagram of the ODSR system is depicted in Fig. 1(b).

B. PROBLEM DESCRIPTION

At first, the ODSR searches and detects the human through Faster R-CNN [15]. If the face is in the orientation of $-45° \sim 45°$ with respect to the optical axis and the position is less than $3.5m$, the Haar Cascade feature descriptor is employed to crop a suitable face for recognizing his/her face emotion (e.g., angry, disgusted, fearful, happy, surprised, and sad). If not, the strategy to approach the above pose region is achieved by the stereo vision based localization and an adaptive stratified finite-time saturated control (ASFTSC) [19]. Subsequently,

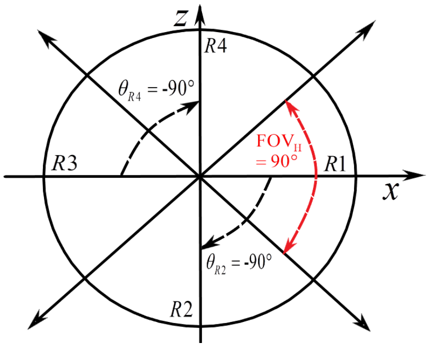


FIGURE 2. Searching strategy for the face of human candidate.

the face emotion is recognized by the proposed SRCN-DFER. Based on the FOV, face candidate in the R1, R2, R3, or R4 depicted in Fig. 2 is searched by the ODSR. It indicates that the initial optical axis of the FOV in region R1 is 0°. If a human is not detected through the Faster R-CNN, then the optical axis of FOV rotates $\theta_{R2} = -90°$ to detect the human candidate in region R2. If a human is still not detected, the optical axis rotates $\theta_{R3} = -90°$ to detect the human candidate in region R3. Likewise, if human is not detected, the optical axis rotates $\theta_{R4} = -90°$ to detect the human candidate in region R4. If the ODSR can't find the human from the above searching strategy, it moves forward a specific distance (e.g., 5*m*) to execute the same procedure. If a human is detected, the center point of the bounding box for detected human is estimated by stereo vision system to achieve the 2D pose between the detected human and ODSR. The overall flowchart of HRC using the proposed SRCN-DFER and SRCN-WSCR is also depicted in Fig. 3.

The eight speech commands, i.e., "Forward", "Backward", "Left", "Right", "Stop", "Follow", "Yes", and "No", are designed for the collaboration of human and ODSR. Likewise, SRCN-WSCR is trained by the speech data files from Google Speech Commands Dataset v0.02 such that 8 human's speech commands through Bluetooth are recognized. Two background noises, i.e., crashing noise by metal and chopstick, and hand clapping noise, are considered to validate its robustness. Finally, the human-robot collaborations, including human and face detections, trajectory tracking control [39], face emotion and speech command recognitions, and music play, are presented.

IV. SRCN-DFER AND SRCN-WSCR

A. SRCN-DFER

The proposed architecture of SRCN-DFER is depicted in Fig. 4, which has the upper part for the architecture (e.g., Convolution, Max-pooling) and the lower part for the output. The designed concepts of SRCN-DFER are described as follows: (i) Five pairs of the conv-pooling with appropriate size are made up of the main part of CNN [40] such that the classification of face emotion is improved. (ii) The max-pooling is often applied to reduce the unnecessary calculation. Nevertheless, the size of max-pooling should be not too large to avoid information loss. The size of 2 × 2 is suitable. Multiscale of convolution kernel (i.e., 7 × 7, 5 × 5, 3 × 3) for face emotion recognition are suitable [16]. (iii) The zero padding of feature maps can better utilize their border information, which is beneficial for the final performance. (iv) From Table 1, the weight of the fully connected layer and LSTM layer possesses the main number of total weight. Nevertheless, the total number is still smaller than that of DCNN [16] (cf. Table 5) or the ResNet50 in [25]. (v) The symbol $\Psi > 1$ denotes the number of LSTMs to tackle the dynamic mapping problem since the each LSTM contains feedback loop [20], [22], [28], [35], [41]. Moreover, these Ψ LSTMs have common weight. (vi) With the online preprocessing mechanism, i.e., Faster R-CNN combined with Haar Cascade feature descriptor, the proposed method is more practical in comparison to some studies [23], [25], which must have the suitable faces cropped in advance.

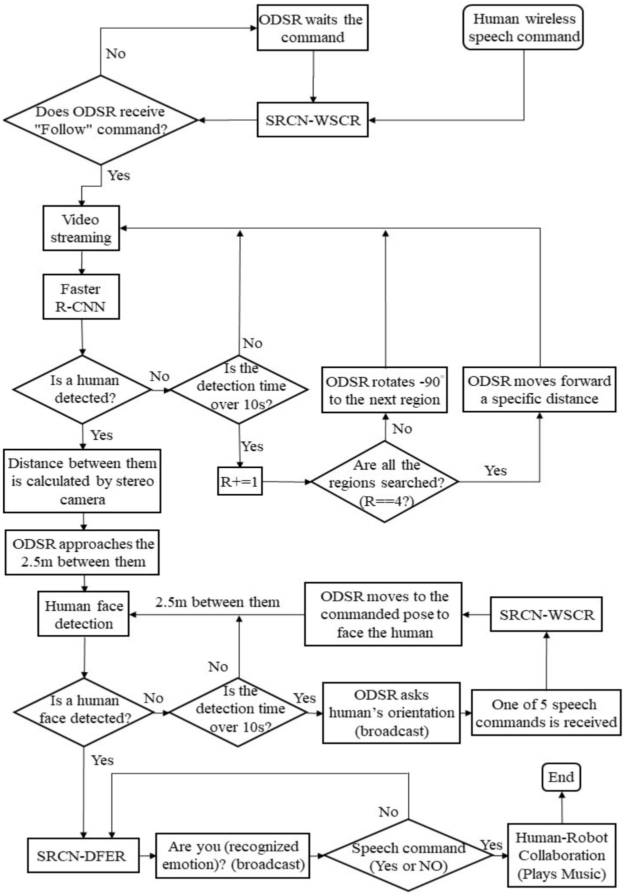

ODSR waits the command
Human wireless speech command
No
Does ODSR receive "Follow" command?
SRCN-WSCR
Yes
Video streaming
Faster R-CNN
No
Is a human detected?
No
Is the detection time over 10s?
ODSR rotates -90° to the next region
ODSR moves forward a specific distance
Yes
Yes
No
Distance between them is calculated by stereo camera
R+=1
Are all the regions searched? (R==4?)
Yes
ODSR approaches the 2.5m between them
Human face detection
2.5m between them
ODSR moves to the commanded pose to face the human
SRCN-WSCR
No
Is a human face detected?
No
Is the detection time over 10s?
Yes
ODSR asks human's orientation (broadcast)
One of 5 speech commands is received
Yes
End
No
SRCN-DFER
Are you (recognized emotion)? (broadcast)
Speech command (Yes or NO)
Yes
Human-Robot Collaboration (Plays Music)

FIGURE 3. Overall flowchart of human-robot collaboration.

The details of four datasets are given as follows: (i) The numbers of humans for the NTUST-IRL, KDEF, JAFFE, and CK+ datasets are respectively 20, 97, 10, and 70, i.e., the total number of humans is 197. (ii) The NTUST-IRL and KDEF are RGB images; in contrast, CK+ and JAFFE are grayscale images. (iii) Gender: man/women. (iv) The resolutions of

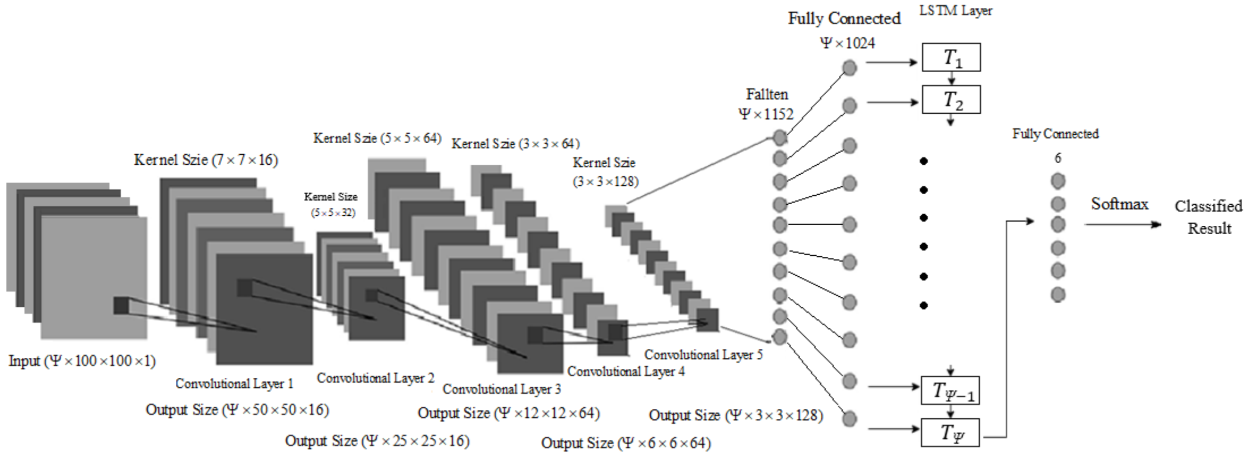


FIGURE 4. The architecture of STCN-DFER.

TABLE 1. Architecture of the proposed SRCN-DFER.

Property Layer	Kernel Size	Stride	Zero Padding	Output	Weight Number
Input				Ψ ×100×100×1	
Convolution1	7×7	1	3	Ψ ×100×100×16	800
Maxpooling1	2×2	2		Ψ ×50×50×16	0
Convolution2	5×5	1	2	Ψ ×50×50×32	12832
Maxpooling2	2×2	2		Ψ ×25×25×32	0
Convolution3	5×5	1	2	Ψ ×25×25×64	51,264
Maxpooling3	2×2	2		Ψ ×12×12×64	0
Convolution4	3×3	1	1	Ψ ×12×12×64	36,928
Maxpooling4	2×2	2		Ψ ×6×6×64	0
Convolution5	3×3	1	1	Ψ ×6×6×128	73856
Maxpooling5	2×2	2		Ψ ×3×3×128	0
Flatten				Ψ ×1152	0
Fully Conn.				Ψ ×1024	1,180,672
LSTM				256	1,311,744
Softmax				6	1,542
Total Weight					2,669,638

NTUST-IRL, KDEF, JAFFE, and CK+ are 100 × 100, 562 × 762, 256 × 256, and 640 × 480, respectively. (v) The training/testing numbers of six face emotions in NTUST-IRL, KDEF, JAFFE, and CK+ are 640 × 6.160×6, 80 × 6.40×6, 20 × 6.10×6, and 80 × 6.40×6, respectively.

The training procedure of SRCN-DFER is described as follows: (i) The CNN with fully connection and softmax layers is first trained by static images of 3 datasets: NTUST-IRL, KDEF, and JAFFE. (ii) After that, a pre-trained weight of CNN but without the fully connection (FC) and softmax (SM) layers is a part of initial weights. Together with the other small random initial weights are employed to train SRCN-DFER. (iii) Subsequently, the overall weight of SRCN-DFER is trained by 280 batches of the sequence images in CK+ dataset.

The loss function of categorical-cross entropy is used for the learning of SRCN-DFER [42]:

$$L(P) = -\sum_{m=1}^{M} t_m \ell og(p_m) \tag{1}$$

where t_m is the target signal of the $m-th$ facial emotion, M is the total number of facial emotions, $P = (p_1, p_2, \cdots, p_M)$ is the probability vector of classified output. Based on (1), the stochastic gradient descent (SGD) of "Adam" (2) is applied to learn the corresponding weights in the CNN except a stack of LSTMs or SRCN:

$$\hat{w}_i(k) = \hat{w}_i(k-1) - \eta_i(k)\hat{m}_i(k) \Big/ \sqrt{\hat{v}_i^{\max}(k) + \varepsilon} \tag{2}$$

where $\eta_i(k)$ is the initial rate 0.01 with the decay rate 10^{-5}, $g(k)$ is the gradient vector, $\varepsilon = 10^{-5}$ avoids a zero division. Moreover, we have

$$\begin{aligned} \hat{m}_i(k) &= m_i(k) \Big/ (1-\alpha_1^k), \\ m_i(k) &= \alpha_1 m_i(k-1) + (1-\alpha_1) g_i(k) \\ \hat{v}_i(k) &= v_i(k) \Big/ (1-\alpha_2^k), \\ v_i(k) &= \alpha_2 v_i(k-1) + (1-\alpha_2) g_i^2(k) \\ \hat{v}_i^{\max}(k) &= \max\left\{\hat{v}_i^{\max}(k-1), \hat{v}_i(k)\right\} \end{aligned} \tag{3}$$

where $g_i(k) = \partial L \big/ \partial \hat{w}_i$, $\alpha_1 = 0.9$, $\alpha_2 = 0.99$. Based on (2) and (3), the pre-training curve of CNN for 3 datasets with the number of static images 4354 and 1088 for training and testing is given in Fig. 5, which possesses the final training loss 0.0057 and testing loss 0.0083 after 1063 steps. The result is satisfactory due to the consistence between training loss and testing loss. Further validations will be given in section V. Since the sequence images for six face emotions are only 280 batches, the overall training curve for SRCN-DFER with $\Psi = 10$ is shown in Fig. 6, which has the training loss of 3.2 × 10^{-5} after 1043 steps. Since the number of face emotion sequences is much smaller than that of learning weight, we neglect the response of testing loss. Finally, the

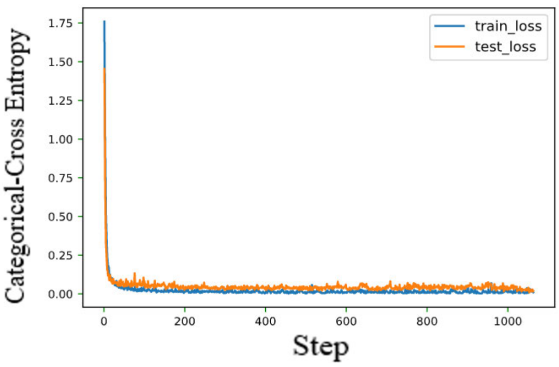


FIGURE 5. Pre-training response of CNN.

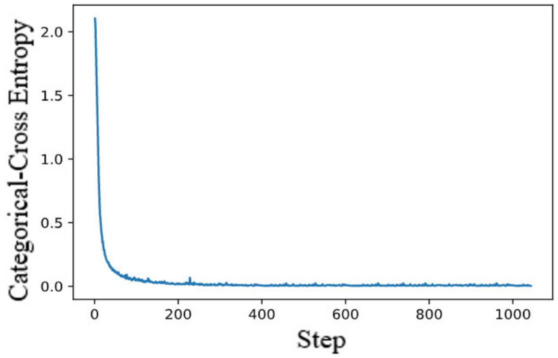


FIGURE 6. Training response of SRCN-DFER.

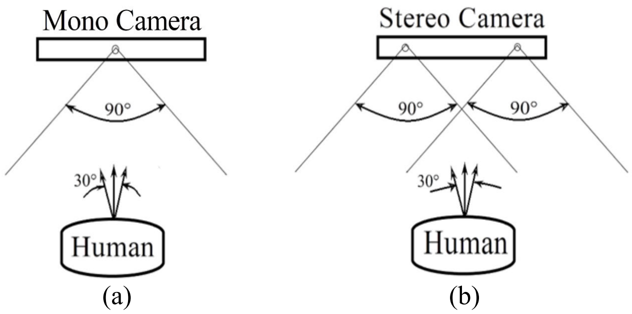


FIGURE 7. The compared FOV and performance between mono and stereo cameras.

proposed SRCN-DFER is described in Algorithm 1, possessing the similar concept of few-shot object detection [43].

The advantages of stereo camera are elaborated in the following. When the human is in the front of stereo camera, their total FOVs is much larger than 90° of mono camera (cf. Fig. 7). Because the facial emotion recognition is sensitive to the view angle of camera, the better recognition rate from them is the representative one. It is definitely better than that of mono camera at the expense of more processing time. Nevertheless, the increasing processing time is acceptable. The experimental video for the comparison between single and stereo cameras can refer to the URL: https://youtu.be/qgR7vyokSPo. It indicates that if the orientation of human face is larger than 15° with respect to the optical axis, the recognition for mono camera always fails. In contrast, stereo camera still successes. Since the face emotion recognition using stereo vision system is complex and seems unnecessary, the simultaneous comparison between left and right cameras with sharing the same learned weights and increasing viewing angle can improve the recognition rate. This advantage was rarely addressed in previous research.

TABLE 2. Architecture of the proposed SRCN-WSCR.

Property Layer	Kernel Size	Stride	Zero Padding	Output	Weight Number
Input				Ψ ×128×1	
Convolution1	5	1	4	Ψ ×128×16	96
Maxpooling1	2	2		Ψ ×64×16	0
Convolution2	7	1	6	Ψ ×64×32	3616
Maxpooling2	2	2		Ψ ×32×32	0
Flatten				Ψ ×1024	0
Fully Conn				Ψ ×128	131,200
LSTM				128	131,584
Fully Conn				64	8,256
Softmax				8	520
Total Weight					275,272

Algorithm 1 SRCN-DFER algorithm

Input: Subimage 100 × 100 from preprocessing; **Output**: Classification of 6 dynamic face emotions $DFE_i, = 1, 2, \cdots, 6$.

```
1: Using a set of static images trains and tests the weight of
   CNN in Table 1 by "Adam" SGD optimizer (2) and (3),
   and small random initial weight.
2: Using the pre-trained weight of CNN and small random
   initial weight for LSTM model, FC and SM layers trains
   the SRCN-DFER in Table 1 with suitable $\Psi$ by "Adam"
   SGD optimizer (2) and (3), and a sequence of dynamic face
   emotion images.
3: If the classified result is not satisfied, then it
   is back to step 2.
4: Output one of $DFE_i, = 1, 2, \cdots, 6$.
```

B. SRCN-WSCR

The eight speech commands, i.e., "Forward", "Backward", "Left", "Right", "Stop", "Follow", "Yes", and "No", are designed for the tasks of HRC. The commands of "Forward", "Backward", "Left", "Right", and "Stop" are employed to give the command of ODSR with respect to human. For example, "Left" and "Right" respectively command the ODSR at the left- and right-hand side of human with 2.5m between them. Likewise, "Forward" and "Backward" respectively command the ODSR in the front and rear side of human with 2.5m between them. Certainly, "Stop" command immediately stops the ODSR. The preprocessing of speech command is described in Fig. 8 or Algorithm 2. The architecture of the proposed SRCN-WSCR is described in Table 2, which is simpler than the SCRN-DFER in Table 1.

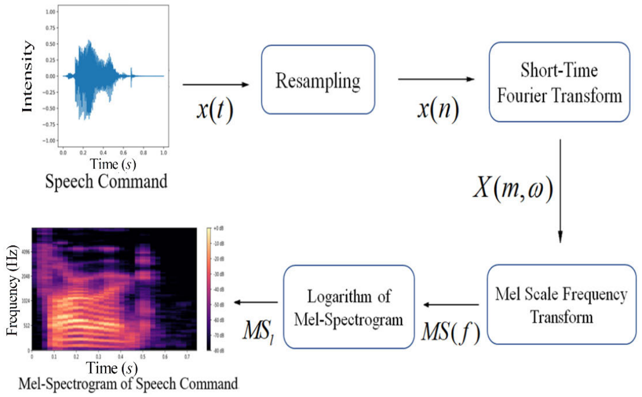

Intensity
Time (s)
Speech Command
x(t)
Resampling
x(n)
Short-Time Fourier Transform
X(m,ω)
Mel Scale Frequency Transform
MS(f)
Logarithm of Mel-Spectrogram
MS_l
Frequency (Hz)
Time (s)
Mel-Spectrogram of Speech Command

FIGURE 8. Processing of speech command.

Before the online application of speech command recognition, the training data from Google Speech Commands Dataset v0.02 is employed to train and test SRCN-WSCR algorithm. It includes (i) over 100,000 speech files with the time length of 1s for 35 classes, (ii) speech files with 6 different background noises. Based on "Adam" SGD optimizer, the training and testing losses of the SRCN-WSCR with $\Psi = 101$ are respectively 0.013 and 0.071 after 752 iterative steps by the number of training and testing voice files of 10372 and 2324. Since the dynamic feature of speech command is dominant, the number of LSTMs is increased in comparison to that in the SRCN-DFER. Since these voice files are sufficiently large and dynamic, no pre-trained response is given. Finally, Algorithm 2 is online applied to the wireless speech command recognition [19]. In its Step 3, is a formulation of common use to convert linear frequency to the Mel-scale frequency such that human speech is more easily distinguished.

```
Algorithm 2 SRCN-WSCR algorithm.
Input: Wireless speech command $x(t)$; Output: Classification of 8 speech commands $WSC_j, j = 1, 2, \cdots, 8$.
1: $x(t)x(n)$.
2: STFT: $X(m,\omega) = \sum_{n=0}^{N-1} x(n)H(n-m)e^{-j\omega n}$, where $N = 512$,
   $m = 160$, $H(n-m) = \left[1 - \cos(2(n-m)\pi \big/ (N-1))\right] \big/ 2$.
3: Mel-Spectrogram: $MS(f) = M_{mf}(f)S(m,\omega)$, where $S(m,\omega) = |X(m,\omega)|^2$, $M_{mf}(f) = 2595 \log_{10}\left(1 + f \big/ 700\right)$.
4: Logarithm Mel-Spectrogram: $MS_l = 10 \log_{10}\left(MS \big/ MS_M\right)$, where $MS_M = \max_{0 \le f \le f_s} \{MS(f)\}$,
   $f_s$ is a specific frequency.
5: Input $MS_l$ to Table 2 with suitable $\Psi$ and "Adam" SGD optimizer (2) and (3).
6: Output one of $WSC_j, j = 1, 2, \cdots, 8$.
```

TABLE 3. Comparison of SRCN-DFER between mono and stereo cameras at 3m with three view angles.

Accuracy (%) / Emotion	Mono Camera			Stereo Camera		
	-15°	0°	15°	-15°	0°	15°
Angry	98	98	92	98	100	100
Disgusted	86	98	86	100	100	100
Fearful	82	96	88	96	96	96
Happy	82	98	92	96	98	98
Sad	98	98	98	98	98	98
Surprised	98	100	98	100	100	100
Angle Avg. (%)	90	98	92	98	98.6	98.6
Average (%)	**93.3**			**98.4**		

TABLE 4. Recognition rate for SRCN-DFER at different distances.

Emotion (%) / Distance	Ang	Dis	Fea	Hap	Sad	Sur	Avg
Fixed 3m	98	100	96	98	100	100	98.6
Fixed 3.5m	98	96	96	98	100	100	98
Wandering 1-3m	100	94	96	98	98	100	97.6

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. VIDEO-BASED FACE EMOTION RECOGNITION

Most of video-based recognition rates of previous research (e.g., [11], [12], [13], [20], [28]) are only acceptable. Before the implementation of HRC, the video-based recognition rate between mono and stereo cameras with three view angles $-15°$, $0°$, and $15°$ at $3m$ are compared in Table 3.

The important observations of Table 3 are addressed as follows: (i) The average recognition rate of stereo camera is 98.4% (URL: https://youtu.be/qgR7vyokSPo), which is 4.9% better than that of mono camera. Moreover, it is better than that of previous research, e.g., [11], [12], [13], [20], [21], [22], [23], [24], [25], [26], [27], and [28]. (ii) As view angle is zero, i.e., the face in the right ahead of camera, the recognition rate of stereo camera is 0.6% slightly better than that of mono camera. Nevertheless, the recognition rate at view angle $-15°$ for stereo camera is 8.6% better than that of mono camera. (iii) The face emotions in Table 3 have not only right-left view angle changes but also up-down pose variations. (iv) In summary, the stereo camera detects face emotions separately, and the higher confidence is the final output result. Using the advantages of stereo camera for the recognition of the face emotions with different view angles, including left-right and up-down pose changes, yields a better result for the distance between 0.8 and 3.5m. (v) Although the previous study [28] has a satisfactory average recognition rate of 84.32%, its "fearful" emotion is the lowest (59.09%). On the contrary, the "fearful" in Table 3 at least has 82%.

The recognition rates for the SRCN-DFER at different distances are shown in Table 4, which is still excellent.

Furthermore, the comparisons among DCNN [16], the proposed SRCN-DFER, and ResNet50 [25] are presented in Table 5. The architecture of ResNet50 includes 49 conv-pooling layers and the last fully connected layer for the classification. The important observations of Table 5 are

TABLE 5. Comparsion among DCNN, the proposed SRCN-DFER, and ResNet 50 with stereo camera.

Method Comparison	DCNN [16]	SRCN-DFER	ReseNet50 [25]
Processing time (s)	0.26-0.32	0.34~0.38	0.35~0.41
Parameter number	66,764,806	2,669,638	23,593,734
Recognition Rate (%) of 6 Dynamic Face Emotions at 3m			
Angry	94	98	98
Disgusted	84	100	88
Fearful	90	96	82
Happy	96	98	96
Sad	98	100	94
Surprised	98	100	100
Average	**93.3**	**98.6**	**93**

TABLE 6. Comparison among 3D-CNN, LRCN, and SRCN-DFER.

Performance Method	Training Accuracy	Testing Accuracy	Training Loss	Testing Loss
LRCN [29]	92.5%	70.2%	0.283	0.702
3D-CNN [30]	92.4%	72.9%	0.192	0.865
SRCN-DFER	**98.7%**	**93.2%**	**0.064**	**0.315**

discussed as follows. (i) The proposed SRCN-DFER is 5.7% and 6% better than DCNN and ResNet50, respectively. (ii) The experimental video for ResNet50 is at the URL: https://youtu.be/tWigt50F_7M, which is acceptable. (iii) The computation time of SRCN-DFER is 0.015s averagely larger than that in DCNN but 0.02s smaller than that in ResNet50. The main reason is that 10 LSTMs with the same weight are required for the proposed approach. (iv) The recognized results of "Disgusted" and "Fearful" are improved by 19% [16] and 19.5% [25], respectively. (v) Two different persons, which have pose variations and are not in the training dataset, and slightly different backgrounds, are employed to further confirm the effectiveness of the proposed SRCN-DFER. The average recognition rate 97.6% is still excellent, cf. URL: https://youtu.be/Kz3fC0tjLLE. (vi) The learning weight in SRCN-DFER is only 4% and 11.3% in comparison to DCNN and ResNet50 such that overfitting problem can be reduced.

In Table 6, LRCN [29] and 3D-CNN [30] directly use many static sequential image for training; in contrast, SRCN-DFER has been pre-trained by static images of 3 datasets: NTUST-IRL, KDEF, and JAFFE. Then, the pre-trained weights of CNN and the small random weights for a stack of LSTMs, fully connection and softmax layers are set as the initial weight to train the SRCN-DFER by a sequence of dynamic images from CK+ dataset. From Table 6, it reveals that the SRCN-DFER through "Transfer Learning" can obtain a better performance due to extracting useful information from data in a related domain and transferring them used in target tasks [44].

TABLE 7. Confusion matrix of the SRCN-WSCR without noisy training data by the test data from google speech commands dataset v0.02.

True Est	Back-ward	Fol-low	For-ward	Left	No	Right	Stop	Yes
Backward	246	0	0	1	1	0	0	0
Follow	0	243	1	0	1	0	0	0
Forward	1	6	249	0	0	0	0	0
Left	0	0	0	248	0	4	0	1
No	3	1	0	0	247	0	0	0
Right	0	0	0	0	0	245	0	0
Stop	0	0	0	0	1	0	250	0
Yes	0	0	0	1	0	1	0	249
Acc (%)	98.4	97.2	99.6	99.2	98.8	98	100	99.6
Avg (%)	**98.8**							

TABLE 8. Confusion matrix of the SRCN-WSCR without noisy training data by the test data from google speech commands dataset v0.02 combined with mental chopsticks crashing noise.

True Est	Back-ward	Fol-low	For-ward	Left	No	Right	Stop	Yes
Backward	220	0	1	0	4	3	7	0
Follow	2	226	25	0	7	2	5	0
Forward	4	6	219	0	1	0	0	0
Left	8	1	1	228	4	4	0	6
No	1	1	0	0	209	0	3	4
Right	1	0	1	1	3	239	1	1
Stop	13	15	2	4	9	1	233	4
Yes	1	1	1	17	13	1	1	235
Acc (%)	88	90.4	87.6	91.2	83.6	95.6	93.2	94
Avg (%)	**90.4**							

B. WIRELESS-BASED SPEECH COMMAND RECOGNITION

At the outset, the confusion matrix of SRCN-WSCR without noisy training data by the test data from Google Speech Commands Dataset v0.02 is shown in Table 7, which is excellent. To verify its robustness, its confusion matrices using the original test data combined with metal and chopstick crashing noise and hand clapping noise are respectively presented in Table 8 and Table 9, which are satisfactory. To boot robust recognition of speech command, six noisy data from Google Speech Command Dataset are added to train SRCN-WSRC. Then, the confusion matrices of the SRCN-WSCR with noisy training data [25] for Table 8 and Table 9 cases are respectively presented in Table 10 and Table 11, which are much improved. It confirms the superiority of the proposed SRCN-WSCR.

C. HUMAN-ROBOT COLLABORATION

The resolution and sampling rate for this study are (3840 × 1080) and 30 FPS, respectively. In addition, the human-robot collaborations in Table 12 are exemplified in the following 6 scenarios. (i) In the beginning, ODSR and human

TABLE 9. Confusion matrix of the SRCN-WSCR without noisy training data by the test data from google speech commands dataset v0.02 combined with hand clapping noise.

True / Est	Back-ward	Fol-low	For-ward	Left	No	Right	Stop	Yes
Backward	228	1	2	6	7	9	4	0
Follow	3	224	13	5	5	0	1	0
Forward	4	18	230	0	0	0	0	0
Left	4	0	0	210	3	3	2	8
No	0	2	1	2	221	0	8	3
Right	2	2	2	11	10	234	1	1
Stop	6	3	1	2	0	2	231	1
Yes	3	0	1	14	4	2	3	236
Acc (%)	91.2	89.6	92	84	88.4	93.6	92.4	94.4
Avg (%)	**90.7**							

TABLE 10. Confusion matrix of the SRCN-WSCR with noisy training data by the test data from google speech commands dataset v0.02 combined with mental chopsticks crash noise.

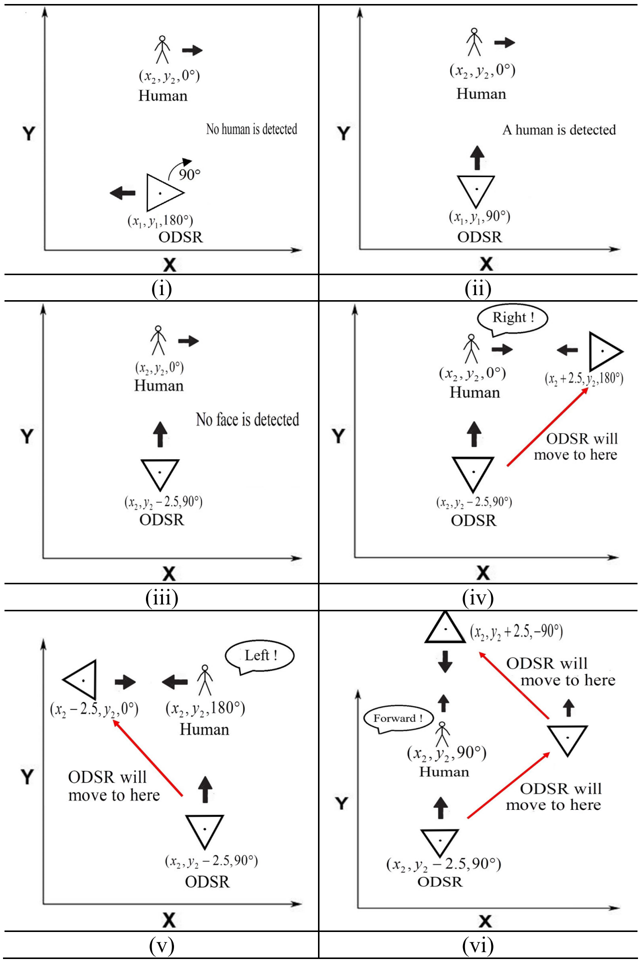
True / Est	Back-ward	Fol-low	For-ward	Left	No	Right	Stop	Yes
Backward	243	0	0	1	2	1	0	0
Follow	1	235	8	0	1	0	0	0
Forward	1	14	241	0	0	0	0	0
Left	1	0	0	239	2	2	0	1
No	2	0	0	0	242	0	1	0
Right	2	0	0	3	0	247	2	1
Stop	0	1	1	1	1	0	247	0
Yes	0	0	0	6	2	0	0	248
Acc (%)	97.2	94	96.4	95.6	96.8	98.8	98.8	99.2
Avg (%)	**97.1**							

are at $(x_1, y_1, 180°)$ and $(x_2, y_2, 0°)$, respectively, where the position is meter. Since the ODSR in Region 1 does not detect a human over 10s, based on the searching strategy in Fig. 2 ODSR rotates 90° in the clockwise (CW) orientation by the ASFTSC in [19]. (ii) In Region 2, a human is detected, and then ODSR is controlled to $(x_2, y_2 - 2.5, 90°)$. (iii) No face over 10s is detected by Faster R-CNN on ODSR. Then ODSR will broadcast "Where is your face orientation?" (iv) The human answers "Right", which indicates the orientation of human face in the right hand side of ODSR. After it is recognized by SRCN-WSCR, ODSR is controlled to $(x_2 + 2.5, y_2, 180°)$ in the alignment with the human face. (v) Likewise, the speech command "Left" is recognized by SRCN-WSCR, and ODSR is then controlled to $(x_2 - 2.5, y_2, 0°)$ in the alignment with the human face. (vi) If the speech command "Forward" is recognized by SRCN-WSCR, ODSR is passing through the waypoint $(x_2 + 2.5, y_2, 90°)$, and then is controlled to $(x_2, y_2 - 2.5, -90°)$ in the alignment with the human face.

TABLE 11. Confusion matrix of the SRCN-WSCR with noisy training data by the test data from google speech commands dataset v0.02 combined with hand clapping noise.

True / Est	Back-ward	Fol-low	For-ward	Left	No	Right	Stop	Yes
Backward	246	0	2	2	7	3	1	1
Follow	0	225	6	0	2	0	0	0
Forward	1	23	242	0	0	1	0	0
Left	1	0	0	236	1	2	1	1
No	0	0	0	1	239	0	2	0
Right	2	0	0	4	0	244	0	0
Stop	0	2	0	2	1	0	246	1
Yes	0	0	0	5	0	0	0	247
Acc (%)	98.4	90	96.8	94.4	95.6	97.6	98.4	98.8
Avg (%)	**96.25**							

TABLE 12. Scenarios of human-robot collaboration.



The operations in Table 12 do not discuss the "Backward" command since ODSR can detect a face at this status and the distance between is about 2.5*m*. The assigned distance of 2.5*m* is due to the environment constraint. Furthermore,

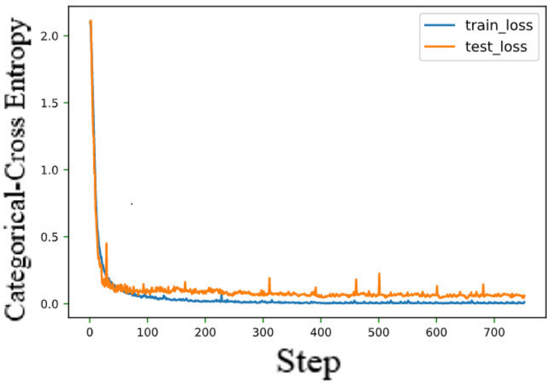


FIGURE 9. Training response of SRCN-WSCR.

the speech command “Stop” will stop the motion of ODSR in any circumstance. After facial emotion is recognized by SRCN-DFER, ODSR will broadcast “Are you (recognized emotion)?” Finally, the “Yes” or “No” speech command from human will be answered via wireless transmission. If “Yes”, the corresponding music will be playing. Otherwise, the continuous recognition by SRCN-DFER is implemented. The proposed approach is different from the relative pose estimation between two robots using an optimal Kalman filter [45] since it must have an operation with limited FOV.

The operations in Table 12 do not discuss the “Backward” command since ODSR can detect a face at this status since ODSR is just in the front of human with the distance of $2.5m$. The assigned distance of $2.5m$ is due to the environment constraint. Furthermore, the speech command “Stop” will stop the motion of ODSR in any circumstance. After facial emotion is recognized by SRCN-DFER, ODSR will broadcast “Are you (recognized emotion)?” Finally, the “Yes” or “No” from human will be answered via wireless transmission. If “Yes”, the corresponding music will be playing. Otherwise, the continuous recognition by SRCN-DFER is implemented.

The experimental video for human-robot collaboration is at the https://www.youtube.com/watch?v=J3DF30TdlzE. One representative human-robot collaboration with the “Happy” face emotion and its motion control response are presented in Table 13 and Fig. 10, respectively. They are explained in the following 13 portions. (i) To begin with, the wireless speech command “Follow” is received by ODSR. (ii) Based on SRCN-WSCR, the planned human-robot collaboration executes. Since the FOV of ODSR is in Region 1, no human is detected. (iii) Based on the searching strategy in Fig. 2, ODSR turns 90° in the clockwise (CW) orientation to Region 2 for the continuous face detection (see the 3rd subplot of Fig. 10(a)). (iv) A human in Region 2 is detected by Faster R-CNN on ODSR. (v) ODSR is controlled to $2.5m$ between them (see the 1st and 2nd subplots of Fig. 10(a) at $t = 57$s or Fig. 10(b)). (vi) Simultaneously, face detection (FD) using

TABLE 13. Important snapshots for human-robot collaboration with the “Happy” face emotion.

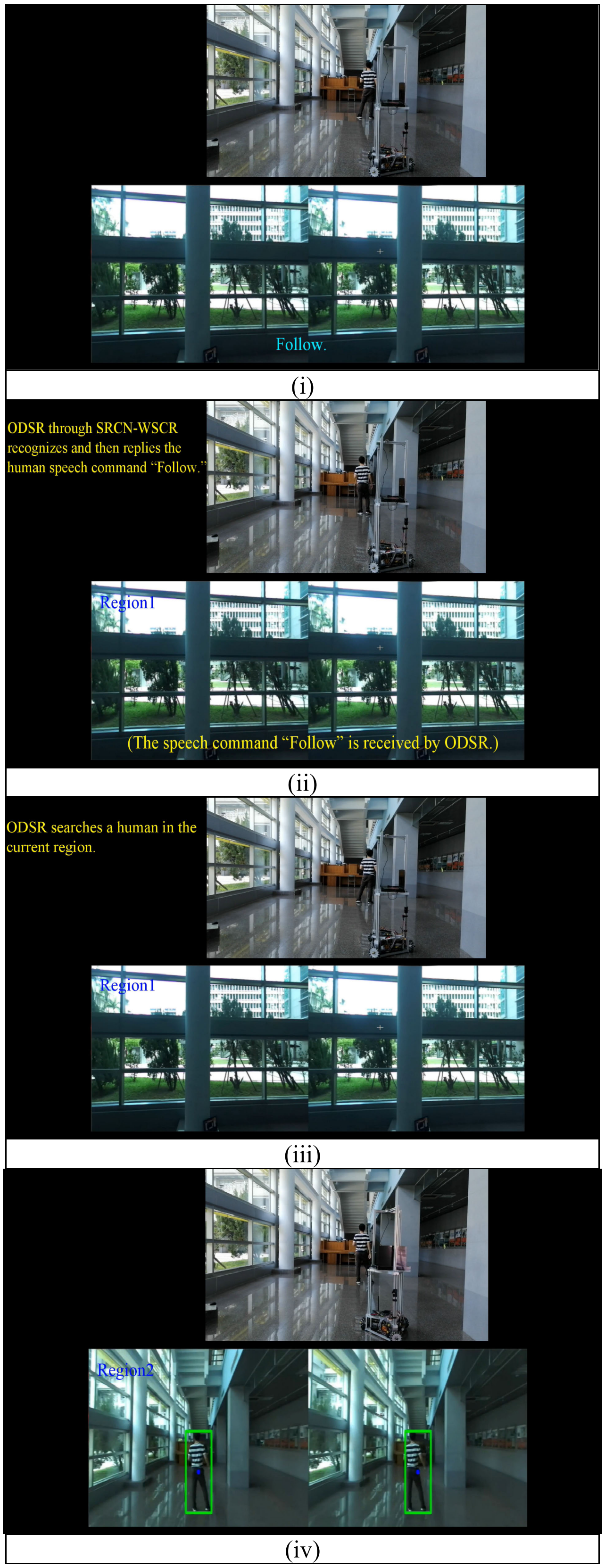
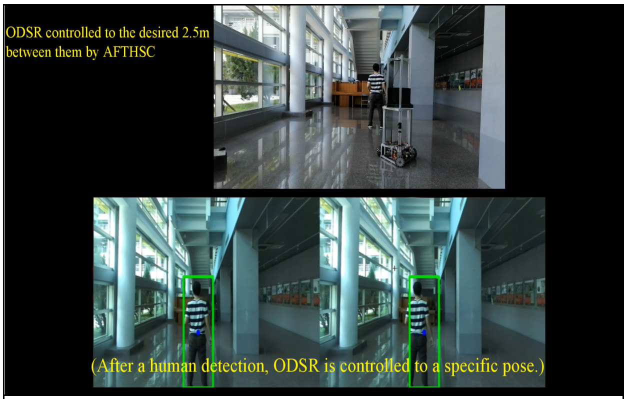

Follow.
(i)
ODSR through SRCN-WSCR recognizes and then replies the human speech command “Follow.”
Region1
(The speech command “Follow” is received by ODSR.)
(ii)
ODSR searches a human in the current region.
Region1
(iii)
Region2
(iv)

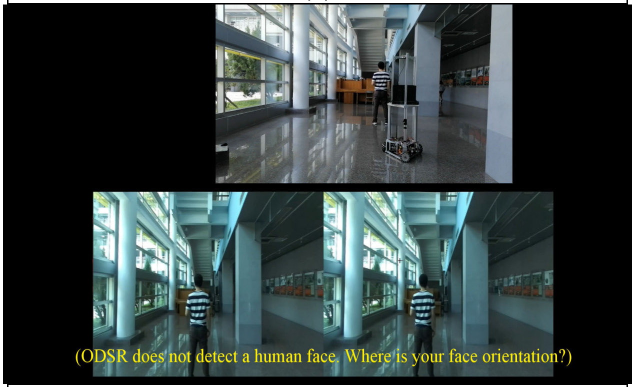
TABLE 13. ***(Continued.)*** **Important snapshots for human-robot collaboration with the "Happy" face emotion.**



ODSR controlled to the desired 2.5m between them by AFTHSC

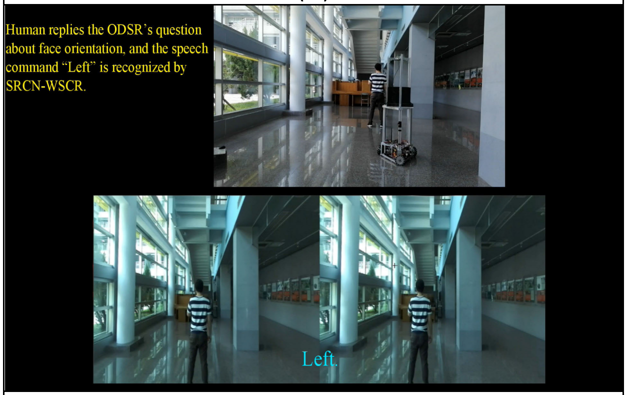
(After a human detection, ODSR is controlled to a specific pose.)

(v)



(ODSR does not detect a human face. Where is your face orientation?)

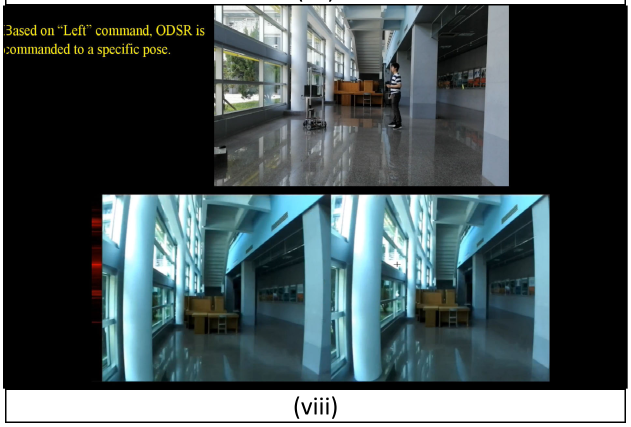
(vi)



Human replies the ODSR's question about face orientation, and the speech command "Left" is recognized by SRCN-WSCR.

Left.

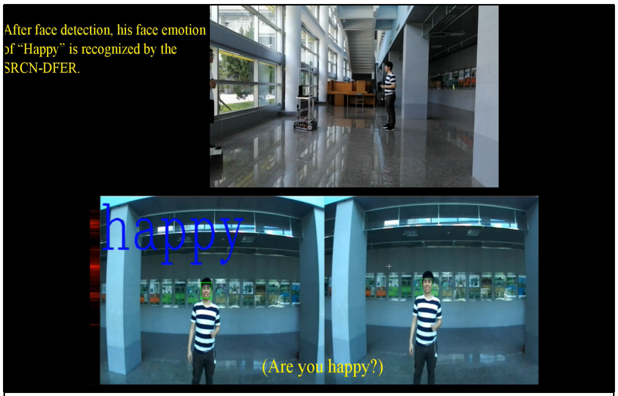
(vii)



Based on "Left" command, ODSR is commanded to a specific pose.

(viii)

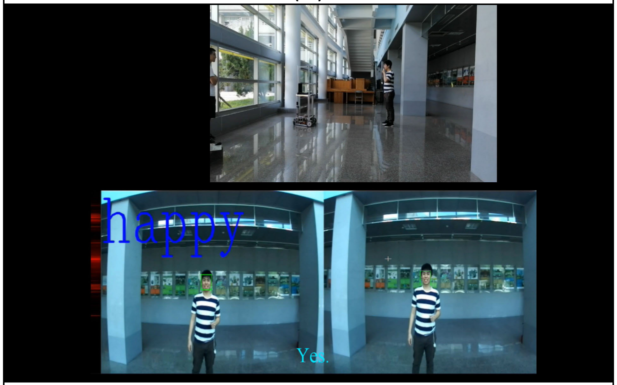
TABLE 13. ***(Continued.)*** **Important snapshots for human-robot collaboration with the "Happy" face emotion.**



After face detection, his face emotion of "Happy" is recognized by the SRCN-DFER.

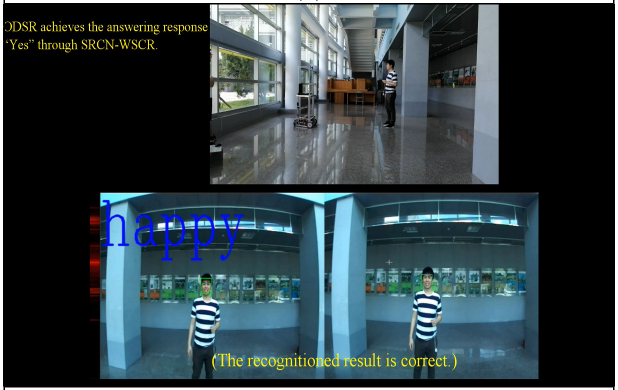
(Are you happy?)

(ix)



Yes.

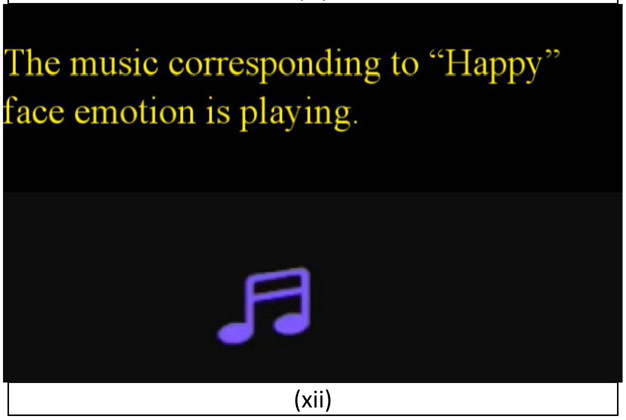
(x)



ODSR achieves the answering response "Yes" through SRCN-WSCR.

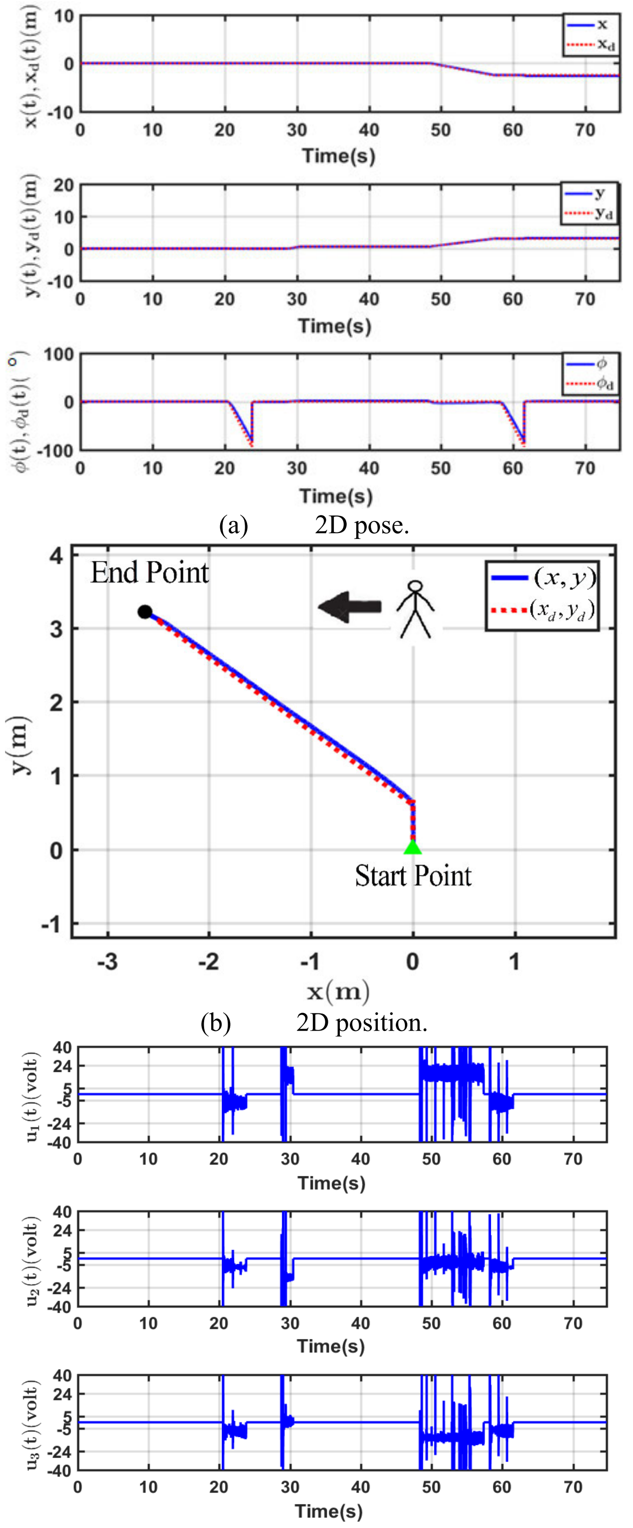
(The recognitioned result is correct.)

(xi)



The music corresponding to "Happy" face emotion is playing.

(xii)



(a) 2D pose.

(b) 2D position.

(c) Control input.

FIGURE 10. Response of HRC in Table 13 by the ASFTSC in [19].

Haar Cascade descriptor is implemented. Since no face is detected, ODSR will ask "Where is your face orientation?" (vii) Human answers "Left" to ODSR. (viii) After the use of SRCN-WSCR, ODSR moves to left side of human and turns 90° in the CW orientation to detect a face (see the 3rd subplot of Fig. 10(a)). (ix) After a face is detected by Haar Cascade feature descriptor, ODSR applies SRCN-DFER to recognize the human's face emotion. (x) ODSR will broadcasts "Are you happy?" (xi) Human answers "Yes" to ODSR. (xii) The corresponding music reflecting "Happy" emotion is playing. (xiii) In this experiment, the camera axis is the same as the motion axis of ODSR, i.e., Y-axis. The control response achieved by the adaptive stratified finite-time saturation control [19] is shown in Fig. 10(c). The simultaneous translation and rotation of ODSR is better than that of differential mobile robot [39], or car-like mobile robot [46].

VI. CONCLUSION

A creative design of SRCN for dynamic mapping of many machine learning problems, e.g., dynamic face emotion recognition, wireless speech command recognition, is established. From the outset, the CNN with fully connection and softmax layers is trained by static images to achieve the corresponding feature vector of facial emotion. Subsequently, SRCN-DFER with a stack of 10 LSTMs using the shared weight is trained by 280 batches of dynamic face emotion images. It is similar to the few-shot concept and achieves an average 98% recognition rate for different persons with pose variation and slightly different backgrounds. The performance is superior to many previous studies for dynamic face emotion recognition [11], [12], [13], [20], [21], [22], [23], [24], [25], [26], [27], [28]. Furthermore, the comparisons among DCNN [16], ResNet50 [25], and LRCN [29], and 3D-CNN [30] confirm the state-of-the-art performance. Since the files of speech command are sufficiently large and dynamic, a pre-trained CNN for SRCN-WSCR is not required. In contrast, its 101 LSTMs are larger than 10 LSTMs in the SRCN-DFER due to the strong dynamics of speech command. The proposed approaches not only effectively tackles the recognitions of dynamic mapping of facial emotion and speech command, but also prevents the overfitting problem in the noisy environment. Finally, the implementation of HRC, e.g., Table 13 and Fig. 10, is accomplished by the integration of trajectory tracking control of ODSR, searching and detection of human and face, preprocessing of speech command, dynamic face emotion and wireless speech command recognitions, and music playing. In the future, multiple ODSRs and humans, distributive UWB network for wireless navigation will be addressed.

REFERENCES

[1] A. Lim and H. G. Okuno, "The MEI robot: Towards using motherese to develop multimodal emotional intelligence," *IEEE Trans. Auto. Mental Develop.*, vol. 6, no. 2, pp. 126–138, Jun. 2014.

[2] C.-L. Hwang, B.-L. Chen, H.-T. Syu, C.-K. Wang, and M. Karkoub, "Humanoid robot's visual imitation of 3-D motion of a human subject using neural-network-based inverse kinematics," *IEEE Syst. J.*, vol. 10, no. 2, pp. 685–696, Jun. 2016.

[3] C.-L. Hwang and G.-H. Liao, "Real-time pose imitation by mid-size humanoid robot with servo-cradle-head RGB-D vision system," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 1, pp. 181–191, Jan. 2019.

[4] L. McCallum and P. W. McOwan, "Extending human–robot relationships based in music with virtual presence," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 4, pp. 955–960, Dec. 2018.

[5] J.-Y. Lin, M. Kawai, Y. Nishio, S. Cosentino, and A. Takanishi, "Development of performance system with musical dynamics expression on humanoid saxophonist robot," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1684–1690, Apr. 2019.

[6] C.-L. Hwang, D.-S. Wang, F.-C. Weng, and S.-L. Lai, "Interactions between specific human and omnidirectional mobile robot using deep learning approach: SSD-FN-KCF," *IEEE Access*, vol. 8, pp. 41186–41200, 2020.

[7] H. Wolfe, M. Peljhan, and Y. Visell, "Singing robots: How embodiment affects emotional responses to non-linguistic utterances," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 284–295, Apr. 2020.

[8] Z. Zhang, K. Qian, B. W. Schuller, and D. Wollherr, "An online robot collision detection and identification scheme by supervised learning and Bayesian decision theory," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1144–1156, Jul. 2021.

[9] S. Boucenna, P. Gaussier, and L. Hafemeister, "Development of first social referencing skills: Emotional interaction as a way to regulate robot behavior," *IEEE Trans. Auto. Mental Develop.*, vol. 6, no. 1, pp. 42–55, Mar. 2014.

[10] A. Zaraki, M. Pieroni, D. De Rossi, D. Mazzei, R. Garofalo, L. Cominelli, and M. B. Dehkordi, "Design and evaluation of a unique social perception system for human–robot interaction," *IEEE Trans. Cognit. Develop. Syst.*, vol. 9, no. 4, pp. 341–352, Dec. 2017.

[11] L. Chen, M. Zhou, M. Wu, J. She, Z. Liu, F. Dong, and K. Hirota, "Three-layer weighted fuzzy support vector regression for emotional intention understanding in human–robot interaction," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2524–2538, Oct. 2018.

[12] E. Benli, Y. Motai, and J. Rogers, "Visual perception for multiple human–robot interaction from motion behavior," *IEEE Syst. J.*, vol. 14, no. 2, pp. 2937–2948, Jun. 2020.

[13] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, and K. Hirota, "Weight-adapted convolution neural network for facial expression recognition in human–robot interaction," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 3, pp. 1473–1484, Mar. 2021.

[14] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the Human Face*. New York, NY, USA: Oxford Univ. Press, 1972.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[16] C.-K. Lee, "Deep CNN stereo camera based dynamic face emotion recognition to fulfill human–robot interaction tasks," M.S. thesis, Elect. Eng., Nat. Taiwan Univ. Sci. Technol., Taipei, Taiwan, Jun. 2020.

[17] Q. Liu, J. Chen, H. Yang, and Z. Yin, "Accurate stereo-vision-based flying droplet volume measurement method," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5000116.

[18] C.-L. Hwang, F.-C. Weng, W.-H. Hung, F. Wu, and C. Jan, "Simultaneous translation and rotation tracking design for sharp corner, obstacle avoidance, and time-varying terrain by hierarchical adaptive fixed-time saturated control," *Mech. Syst. Signal Process.*, vol. 161, Dec. 2021, Art. no. 107969.

[19] C.-L. Hwang, F.-C. Weng, D.-S. Wang, and F. Wu, "Experimental validation of speech improvement-based stratified adaptive finite-time saturation control of omnidirectional service robot," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 2, pp. 1317–1330, Feb. 2022.

[20] J. Lee, S. Kim, S. Kim, and K. Sohn, "Multi-modal recurrent attention networks for facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 6977–6991, 2020.

[21] Y. Hu, M. Lu, C. Xie, and X. Lu, "Driver drowsiness recognition via 3D conditional GAN and two-level attention bi-LSTM," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4755–4768, Dec. 2020.

[22] D. A. Chanti and A. Caplier, "Deep learning for spatio-temporal modeling of dynamic spontaneous emotions," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 363–3760, Apr.-Jun. 2021.

[23] S. Xie, H. Hu, and Y. Chen, "Facial expression recognition with two-branch disentangled generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2359–2371, Jun. 2021.

[24] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-GCN: Correlation based graph convolutional network for audio-video emotion recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 3793–3803, 2021.

[25] F. Zhang, M. Xu, and C. Xu, "Weakly-supervised facial expression recognition in the wild with noisy data," *IEEE Trans. Multimedia*, vol. 24, pp. 1800–1814, 2022.

[26] A. Shirian, S. Tripathi, and T. Guha, "Dynamic emotion modeling with learnable graphs and graph inception network," *IEEE Trans. Multimedia*, vol. 24, pp. 780–790, 2022.

[27] F. Becattini, F. Palai, and A. D. Bimbo, "Understanding human reactions looking at facial microexpressions with an event camera," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9112–9121, Dec. 2022, doi: 10.1109/TII.2022.3195063.

[28] G. Du, Z. Wang, B. Gao, S. Mumtaz, K. M. Abualnaja, and C. Du, "A convolution bidirectional long short-term memory neural network for driver emotion recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4570–4578, Jul. 2021.

[29] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[30] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[31] Z. Yue, S. Ding, S. Yang, L. Wang, and Y. Li, "Multimodal information fusion approach for noncontact heart rate estimation using facial videos and graph convolutional network," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2502813.

[32] C. Dhiman and D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics," *IEEE Trans. Image Process.*, vol. 29, pp. 3835–3844, 2020.

[33] H. Wang, P. Tang, Q. Li, and M. Cheng, "Emotion expression with fact transfer for video description," *IEEE Trans. Multimedia*, vol. 24, pp. 715–727, 2022.

[34] H. Liu, H. Zhao, J. Wang, S. Yuan, and W. Feng, "LSTM-GAN-AE: A promising approach for fault diagnosis in machine health monitoring," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[35] S. M. J. Jalali, S. Ahmadian, A. Kavousi-Fard, A. Khosravi, and S. Nahavandi, "Automated deep CNN-LSTM architecture design for solar irradiance forecasting," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 1, pp. 54–65, Jan. 2022.

[36] L. Yang, S. Song, J. Fan, B. Huo, E. Li, and Y. Liu, "An automatic deep segmentation network for pixel-level welding defect detection," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5003510.

[37] M. Vukovic, M. Stolar, and M. Lech, "Cognitive load estimation from speech commands to simulated aircraft," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 1111–1122, 2021.

[38] Y. Dong and Y.-D. Yao, "Secure mmWave-radar-based speaker verification for IoT smart home," *IEEE Internet Thing*, vol. 8, no. 5, pp. 3500–3510, May 2021.

[39] C.-L. Hwang, C.-C. Yang, and J. Y. Hung, "Path tracking of an autonomous ground vehicle with different payloads by hierarchical improved fuzzy dynamic sliding-mode control," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 899–914, Apr. 2018.

[40] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4131–4144, Aug. 2018.

[41] J. Chen, J. Zhang, X. Xu, C. Fu, D. Zhang, Q. Zhang, and Q. Xuan, "E-LSTM-D: A deep learning framework for dynamic network link prediction," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 6, pp. 3699–3712, Jun. 2021.

[42] S. R. Dubey, S. Chakraborty, S. K. Roy, S. Mukherjee, S. K. Singh, and B. B. Chaudhuri, "DiffGrad: An optimization method for convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4500–4511, Nov. 2020.

[43] X. Li, J. Deng, and Y. Fang, "Few-shot object detection on remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[44] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4500–4511, Nov. 2015.

[45] Y. Wang, X. Liu, Y. Kang, and S. S. Ge, "Anomaly resilient relative pose estimation for multiple nonholonomic mobile robot systems," *IEEE Syst. J.*, vol. 16, no. 1, pp. 659–670, Mar. 2022.

[46] C.-L. Hwang, "Comparison of path tracking control of a car-like mobile robot with and without motor dynamics," *IEEE/ASME Trans. Mechatronics*, vol. 21, no. 4, pp. 1801–1811, Aug. 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 at 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 a Professor of mechanical engineering, from 1996 to 2006. From 1998 to 1999, he was a Research Scholar at the George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA. From 2006 to 2011, he was a Professor at 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 Electrical and Computer Engineering of Auburn University, Auburn, AL, USA. He is the author or coauthor of many high-impact journal and conference papers in the related field. His current research interests include robotics, fuzzy neural modeling, classification and control, finite-time control, (distributed) visual or wireless localization or navigation systems, nonlinear multi-agent systems, remote control of UAV, face and speech emotion recognition, and human–robot collaborations. He was a recipient of the Excellent and Outstanding Research Awards from the National Taiwan University of Science and Technology, in 2018 and 2019. He is also selected as top 2% most influential scientists in ‘‘Career 1960 to 2021’’ and ‘‘Single 2021’’ by Stanford University.

YU-CHEN DENG received the B.E. degree in electrical engineering from the National Formosa University of Science and Technology, Chang-Hua, Taiwan, in 2019, and the M.E. degree from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2021. His research interests include dynamic facial emotion and voice recognitions via machine learning and robotics.

SHIH-EN PU received the B.E. degree in electrical engineering from the National Ocean University of Taiwan, Keelung, Taiwan, in 2020, and the M.E. degree from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2022. His research interests include wireless navigation and following control of multiple robots and computational intelligence.