WiFE: WiFi and Vision based Unobtrusive Emotion Recognition via Gesture and Facial Expression

Yu Gu, *Senior Member, IEEE,* Xiang Zhang, Huan Yan, Jingyang Huang, Zhi Liu, *Senior Member, IEEE,* Mianxiong Dong, *Senior Member, IEEE,* Fuji Ren, *Senior Member, IEEE*

Abstract—Emotion plays a critical role in making the computer more human-like. As the first and most essential step, emotion recognition emerges recently as a hot but relatively nascent topic, i.e., current research mainly focuses on single modality (e.g., facial expression) while human emotion expressions are multi-modal in nature. To this end, we propose a unobtrusive emotion recognition system leveraging two emotion-rich and tightly-coupled modalities, i.e., gesture and facial expression. The system design faces two major challenges, namely, how to capture the emotional expression in both modalities without disturbing the subject and how to leverage the relationship between modalities for recognizing the emotion. For the former, we explore WiFi and vision for unobtrusive and contactless gesture and facial expression sensing, respectively. For the latter, we propose a novel deep learning framework named Multi-Source Learning (MSL) to efficiently exploit both self-correlation in the modality and cross-correlation between modalities for fine-grained emotion recognition. To evaluate the proposed method, we prototype the system on low-cost commodity WiFi and vision devices, build a first-of-its-kind WiFi-Vision emotion dataset, and conduct extensive experiments. Empirical results not only verify the effectiveness of WiFE in emotion recognition, but also confirm the superiority of multi-modality over single-modality.

Index Terms—Gesture Recognition, Facial Expression, Emotion Recognition, Multimodal, Channel State Information

1 Introduction

THREE decades ago, Minsky mentioned in his book *The Society of Mind*: “***The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions***” [1]. Since then, emotion has been recognized as a critical role in making machines more human-like and attracts much research attention. As a result, emotion recognition, the first and most important step in affective computing, emerges as a hot but relatively nascent topic [2].

Charles Darwin claimed that the expression of emotion usually involves facial expression, behavioral response and physical response [3]. He proposed the universality hypothesis about emotion expression, i.e., the expression of emotion is universal across race or culture. It has been further tested by psychologists like Paul Ekman [4] and Carrol lzard [5]. Their empirical studies have revealed a number of interesting findings. For instance, facial expression is likely to be unique to each emotion and conveys the internal states of mind [6]. These psychological studies have laid the foundation of affective computing [7], and triggered the blossom of modern research on emotion recognition [8].

However, previous research on this topic mainly focuses on single modality like facial expression [9]. Emotion expression is person-dependent and multi-modal in nature [10]. Therefore, there is a recent trend of exploring multimodality for more reliable and accurate emotion recognition, e.g., facial-audio [11] or facial-EEG (electroencephalogram) [12]. Careful selection of modalities is key to the multi-modal emotion recognition. A noted Jungian analyst, Irene Claremont de Castillejo, once pointed out, “***Emotion always has its roots in the unconscious and manifests itself in the body***”. A gesture is a movement that a person makes with a part of his hands, head or face to express emotion or information. It is naturally correlated to facial expression, and constitutes an important non-verbal social affective cue [13]. Hence, we focus on exploring gesture and facial expression, two tightly-coupled and emotion-rich modalities [14], for a fine-grained emotion recognition, which comes with two major design challenges.

The first challenge is how to capture emotional expression in gesture and facial expression without disturbing the subject? Currently, facial expression is mainly handled by the vision-based methods while body gesture can be captured by sensors. But sensors that are in general contact or even invasive, which may interfere the subject and thus contaminate the emotional cues. Therefore, we tend to a newly-developed unobtrusive alternative for sensors, i.e., WiFi.

The ubiquitous WiFi has been shown to be capable of capturing subtle body movements like respiration and heart-beat due to the multi-path effect over the human body [15]. Its Channel State Information (CSI) has been explored as a substitution of wearable sensors for unobtrusive and contactless gesture recognition [16]. Recently, Gu *et al.* pro-

- *Yu Gu and Fuji Ren are with I*$^{+}$ *Lab, School of Computer Science and Engineering, University of Electronic Science and Technology of China. E-mail: yugu.bruce@ieee.org, renfuji@uestc.edu.cn*
- *Xiang Zhang, Huan Yan and Jiangyang Huang are with the School of Computer and Information, Hefei University of Technology, China.*
- *Zhi Liu is with Department of Computer and Network Engineering, The University of Electro-Communications, Japan.*
- *Mianxing Dong was with the Dept. of Sciences and Informatics, Muroran Institute of Technology, Japan.*
- *Corresponding author:Xiang Zhang.*

posed EmoSense, a CSI-based emotion recognition system leveraging the body language only [17]. Though EmoSense demonstrates the possibility of leveraging CSI only for emotion recognition, its performance is constrained to the limited emotional features of a sole modality. It motivates us to involve facial expression, a particularly salient stimuli for delivering emotional signal.

The sensitivity of WiFi CSI over human motion plays an important role in recovering the subtle gesture like an imperceptible nod. Unlike our rivals mostly relying on the empirical study for tuning the antenna layout, we propose a Rician-K factor based theoretical model to enhance the sensing granularity. The key idea is to highlight the gesture-induced information by suppressing the gesture-unrelated information on channel response. In particular, we amplify the ratio of the Non-Line-of-Sight (NLoS) component corresponding to gestures along signal propagation over channel response via weakening the Line-of-Sight (LoS) part. Both theoretic analysis and experimental results have verified that the proposed model can significantly improve the sensitivity of CSI over gestures with little implementation cost.

The second challenge is how to efficiently exploit the correlations among the large-volume and heterogeneous data contributed by the two modalities for emotion recognition? Most of the current multi-modal emotion recognition is based on early-fusion or late-fusion [18]. The former combines the features extracted by multiple encoders (encoders are used to extract features from the multi-modal data) and sends them to a single decoder (the decoder is used to process the features for classification) for recognition. The latter trains a decoder for each encoder, and applies voting or retraining to the output of each decoder to obtain the recognition result. However, early-fusion is vulnerable to data loss, and late-fusion requires training multiple decoders, increasing the training complexity significantly [19]. Moreover, late-fusion is unable to perform knowledge sharing between modalities at feature-level.

To fill in the gap, we propose a novel Multi-Source Learning (MSL) framework to exploit both self-correlation in the modality and cross-correlation between modalities describing the same physical event, i.e, emotional expression. For each modality, MSL employs one encoder to extract the emotion-related self-correlated features, which will be fed into a shared decoder allowing modality-related knowledge being interacted for getting cross-correlated features through the parameter sharing mechanism [20]. The blended features will then help each modality to decide its own output. MSL makes the final decision based on voting over modality-related outputs. Moreover, a gesture usually involves different parts of the body. To describe the correlation among them, we specifically design a fine-grained feature based on their velocity differences derived by performing the DWT (Discrete Wavelet Transform) on CSI.

To evaluate the performance of WiFE, we prototype it with low-cost off-the-shelf WiFi and vision devices and build a first-of-its-kind WiFi-Vision emotion dataset. WiFE dataset has 7 emotions (i.e. Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise) and 35 kinds of emotional expression (5 for each emotion) performed by twenty volunteers, resulting in 3500 video clips as well as the corresponding CSI sequences. Each video clip has a frame rate of 30Hz and a resolution of 720p. The sampling rate for collecting CSI data is set to 500 packets/second, and the CSI profile contains data of 90 subcarriers from 3 receiving antennas. The total size of WiFE dataset is about 82GB. We conduct extensive experiments on WiFE dataset and the empirical results confirm the superiority of the bi-modality by achieving 86.54% recognition accuracy for seven emotions on average, as compared with 70.2% and 74.51% recognition accuracy by gesture-only and facial-only solutions, respectively. We also verify that MSL outperforms the state-of-the-art early-fusion method (79.97%) with robustness against data loss, and late-fusion method (86.29%) with 33.9% less complexity, respectively.

In summary, our contributions are summarized as follows:

- To the best of our knowledge, we are among the first to leverage WiFi and vision for unobtrusive gesture and facial expression emotion recognition. We build the first bi-modal WiFi-Vision emotion dataset, which will be released to the public.
- We propose a Rician-K factor based model to enhance the sensing granularity of WiFi CSI to the human gesture. It can be seamlessly integrated into other similar WiFi sensing systems with little implementation cost. To the best of our knowledge, it is the first quantitative mode on leveraging Rician fading for enabling sub-wavelength gesture capture.
- We design a multi-source learning framework to efficiently exploit the correlations among modalities for fine-grained emotion recognition. It allows modality-related knowledge being interacted through the parameter sharing mechanism. Moreover, we design a handcrafted fine-grained feature decomposing one gesture into different parts of the body through velocity differences derived by performing the DWT on CSI.
- We prototype WiFE with low-cost commodity WiFi and vision devices, and evaluate it with extensive experiments. The results not only verify its effectiveness in emotion recognition, but also confirm the superiority of multi-modality over single-modality.

2 Related Work

WiFE involves two research areas, i.e., gesture and facial expression recognition. In this section, we will introduce the representative research on both topics, and outline the motivation of our research.

2.1 Body Gesture Recognition

Gesture recognition is an important part of human-computer interaction. Most current gesture recognition solutions are based on contact or even invasive sensors [21], which may interfere the subject and thus hinder the practicality. Therefore, there are other studies devoted to contactless gesture recognition, such as a vision-based system using a hybrid RGB and depth-sensing approach to extract body gestures [22], or a ultrasonic system extracting gestures from sound waves [23]. However, they still exist several shortages

TABLE 1: Acronyms used in the paper

Acronyms	Full Name
AFEW	Acted Facial Expressions in the Wild
CFR	Channel Frequency Response
CSI	Channel State Information
Densenet	Dense Convectional Network
DWT	Discrete Wavelet Transform
EEG	Electroencephalography
ECG	Electrocardiograph
EOG	Electrocardiography
FCC	Full-Connected Layers Classifier
LOS	Line-of-Sight
LSTM	Long Short-Term Memory
MTCNN	Multi-task Cascaded Convolutional Networks
MTL	Multi-Task Learning
MSL	Multi-Source Learning
NLOS	Non-Line-of-Sight
PCA	Principal Component Analysis
RSS	Received Signal Strength
SVM	Support Vector Machines
VGG	Visual Geometry Group

like the illumination requirement, the installation cost, and the instrumentation overhead.

Recently, device-free sensing technology based on WiFi signals has attracted widespread attention from researchers due to its advantages (easy deployment, wide availability, privacy protection and low cost) [24]. The basic principle is that when the WiFi signal encounters the human body during the propagation process, it will experience reflection, refraction, diffraction and scattering, which will disturb the normal propagation of the signal. By analyzing the received signal and detecting the characteristics of the signal disturbance, human events can be identified [25], [26].

In early days, researchers use WiFi RSS (Received Signal Strength) to recognize some coarse-grained events, such as daily activities. Since the channel state information (CSI) can be extracted from the physical layer of the wireless network card [27], more detailed information can be extracted from WiFi signals, and more fine-grained gesture sensing can be realized [28]. WiFi CSI can be used to track accurate breath even heartbeat [29], to recognize more fine-grained gestures [16]. Researchers even realized human body keypoint imaging [30] with WiFi CSI.

2.2 Facial Expression Recognition

Emotion recognition is one fundamental research issue of affective computing [31]. State-of-the-art emotion recognition schemes mainly focus on the inherent facial expressions, facial expressions are also the most commonly used methods for judging emotions by humans, it is an important channel for humans to convey information through emotion [32], [33].

For facial expression, current work mainly leverage vision based methods (images or videos). The features used in current facial expression recognition schemes can be divided into two categories from the perspective of the time domain, namely, static features based on the a single picture and dynamic features based on the facial change sequence of multiple pictures. In particular, [34] use the spatial information of a single image to recognize the emotions. [35] extend this and considers the temporal correlation between successive frames of the captured video.

Current vision based Facial Expression recognition (FER) methods mainly huddle in date-level and feature-level for emotion recognition. Data-level methods focus on cleaning the sample images directly. The prevailing approach is to downweight uncertain samples by building a multi-branch deep network to model the latent label distribution to pinpoint those uncertain ones [36], [37]. Weighting samples attempts to constrain the influence of uncertainties in data and prevent the network from over-fitting uncertain images. A more aggressive idea is to relabel those uncertain images via an attention mechanism [38].

Feature-level methods concentrate on generating discriminative data representation via handcraft features or deep features, respectively. The former usually explicitly extracts folds and geometry changes in local regions or global areas caused by facial expressions [39], [40]. However, handcrafted features are not robust and are computationally intensive in general. Recently, neural networks have been frequently explored for extracting deep features for FER. A popular idea is to design novel loss functions to enhance the discriminative power of the network to learn salient deep features [41], [42]. Besides, another tempting approach is to leverage the attention mechanism to focus on crucial local regions like areas around mouth and eyes for more distinctive representations of facial expressions [43], [44].

2.3 Gesture and Physiological Based Emotion Recognition

Human emotional expressions are multi-modal in nature. Besides facial expression, recent research also focuses on interpreting emotions through body gestures [17], [45], as well as vital signs such as electroencephalogram (EEG) [46] and electrocardiogram (ECG) signals [47].

Body gestures are one of the most important forms of non-verbal communication. They include movements of hands, head and other parts of the body that allow individuals to convey a variety of feelings, thoughts and emotions [48], [49], [50]. Shiffrar *et.al* [51] show that body movements and postures encode rich information about a person's status, including their awareness, intention, and emotional state. For example, nodding as a sign of affirmation or consent is probably innate, turning to the sides as a sign of refusal is a gesture we learn during early childhood [52]. Some researchers find that human participants of a study could not correctly identify facial expressions associated with winning or losing a point in a professional tennis game when facial images were presented alone, whereas they were able to correctly identify this distinction with images of just the body or images that included both the body and the face [50]. Recently, Luo *et al.* created a Body Language Dataset (BOLD) to recognize bodily expression of emotion in the wild [45]. Our previous work, i.e., EmoSense, leverages WiFi-based sensing technology to sense fine-grained body gestures for emotion recognition [17].

There are also many studies that utilize physiological signals (EEG, ECG) [53] for emotion recognition, because facial expressions or body gestures can be disguised while physiological signals are much harder to control. However, the acquisition of physiological signals requires the subject to wear contact or even invasive sensors, which could disturb generating emotions or expressing them.

Some previous studies have used wireless signals to achieve emotion recognition [17], [54], [55]. EQ-Radio uses FMCW radar to obtain fine-grained heartbeat signals, and then uses SVM to classify the inputs into Sadness, Anger, Pleasure and Joy after extracting features from the heartbeat signals. Khan et al [55] used a deep learning model to process breathing and heartbeat signals obtained from wireless signals to perform emotion recognition. The wireless signal they used is 5.8GHz radar, and their solution enables the classification of four types of emotions: Disgust, Joy, Relax and Scary. Both solutions are based on vital signs such as respiration and heart beat. However, in practical scenarios the changes in wireless sensing signals due to vital signs can be confused with those due to body movements, thus both solutions require the user to be at rest. In addition, these solutions use specialized sensors and are more expensive. Our previous work Emosense [17] uses the more generalized Wi-Fi as the sensing signal and is based on body movements rather than vital signs, but the Emosense dataset contains only the simple actions of sitting at a table and it can only distinguish between the four emotions Happy, Sad, Anger and Fear. Our solution enables passive 7-classes emotion recognition in more complex and variable scenarios.

2.4 Hybrid Emotion Recognition

Some work explored multi-modal based hybrid emotion recognition. Almost all hybrid emotion recognition methods contain vision modality, researchers adds other sources, e.g. speech [56], body gesture [57], and physiological signal [58], to construct multi-modal systems towards even better performance. Hybrid schemes based on multi-modal fusion can make up for the shortcomings of single modal. Through effective fusion (feature-level or decision-level fusion), the performance of these schemes are greatly improved compared with the methods based on single modality , and multi-modal based schemes can also allow for more robust predictions [59].

Multi-modal fusion is the concept of integrating information from multiple modalities to predict an outcome measure. The current multi-modal fusion schemes are mainly early-fusion (feature-level) and late-fusion (decision-level) [18], the difference between the two solutions is that when and what they combine different modalities.

Early-fusion fuses features immediately after they are extracted by encoders (often by simply concatenating their representations), and late-fusion performs integration after each of the modalities/decoders has made a decision (e.g., classification or regression) [19]. Each of these two methods has advantages and disadvantages, early-fusion [60] can exploit correlations and interactions between the features of different modality, and it only requires to train one decoder, make the training easier than late-fusion, however, it can't work when some modalities are missing. In contrast, late fusion [56] fuse all results output by all decoders use fusion mechanisms such as voting, or retraining. Early-fusion is not resilient to data loss, i.e., some modalities are missing. One the other hand, late-fusion can handle such modality-missing at the cost of training one decoder for each modality. Moreover, it ignores the feature level interaction between modalities. To fill in the gap, inspired by the multi-task learning [61], we propose the MSL method to realize the feature level interaction between modalities based on parameters sharing even with missing modalities.

3 System Overview

Fig. 1 presents an overview of the system architecture of WiFE, which mainly consists of two data streams (CSI for gesture and Vision for facial expression) and three modules, i.e., WiFi-based gesture capture, Vision-based facial expression capture and Multi-modal fusion. Tab. 1 includes all acronyms used in the paper.

In the first module, we first leverage the Ricean-*K* factor to enhance CSI for better capturing the gestures. Then, the collected CSI representing the variation of channel response induced by gestures is pre-processed to reduce the background noise and data dimension. Then, we design a fine-grained feature decomposing a gesture into different parts of the body via velocity differences by performing DWT on CSI to better preserve the self-correlation of a gesture. Lastly, we use three kinds of Densenet and LSTM to extract static and temporal features from DWT feature map and raw CSI data, respectively.

In the second module, we rely on both temporal and spatial features to better preserve the facial expression. Specifically, we first leverage the up-to-date Multi-task Cascaded Convolutional Networks (MTCNN) to extract face-related clips from video frames. Then we use the widely-used Densenet to extract the static features (spatial features) from these clips. Lastly, we also use the state-of-the-art VGG-LSTM network to extract the temporal characteristics between these clips. Exploring both the spatio-temporal features help better preserve the facial expression.

In the last module, we propose a MSL framework to perform bi-modal fusion for emotion recognition. In particular, MSL blends the CSI and Vision features to explore the correlations between gesture and facial expression conveying the same emotion. Moreover, compared to static images, a video provides rich dynamic features describing the temporal correlations of the emotion. Therefore, MSL leverages not also the spatial features from both modalities but also the temporal features for fine-grained emotion recognition. Next, we elaborate each module in detail.

3.1 WiFi-based Gesture Capture

In this part, we introduce how to capture gesture through WiFi sensing. Firstly, we explain how to collect the WiFi CSI. Secondly, we propose a Rician-*K* factor based model to enhance CSI for capturing gestures. Lastly, we present a DWT-based feature decomposing a gesture to get self-correlations in the temporal-spatial domains.

3.1.1 CSI Collection

CSI describes the signal's attenuation on its propagation paths, such as scattering, multi-path fading or shadowing fading, and power decay over distance. In frequency domain,it can be characterized as:

$$\vec{Y} = \vec{H} \cdot \vec{X} + \vec{N} \tag{1}$$

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3285777

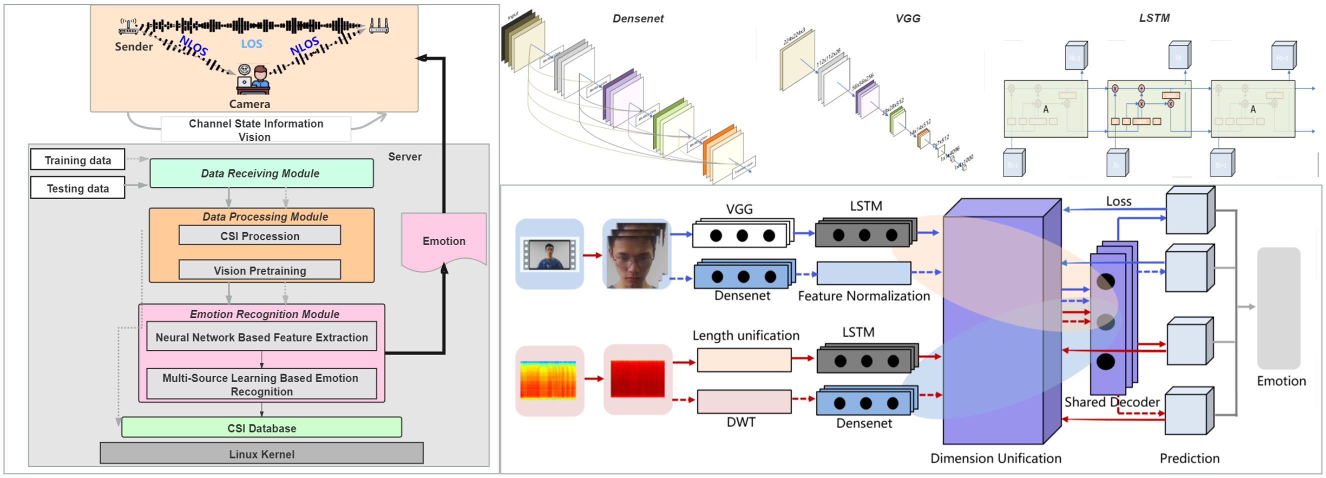


Fig. 1: An overview of the WiFE. The input of the WiFi-based gesture capture model is PCA-processed CSI data, while the input of the vision-based facial expression capture model is images containing only the face extracted by MTCNN.

Where $\vec{Y}$ and $\vec{X}$ are the received and transmitted signal vectors, respectively. $\vec{N}$ is the additive white Gaussian noise, and $\vec{H}$ is the channel matrix representing CSI information.

The WiFi spectrum is divided into multiple orthogonal subcarriers. For each subcarrier, the channel frequency response (CFR) can be expressed as:

$$H_i = L_i + j \cdot Q_i = |H_i| e^{j\angle H_i} \tag{2}$$

Where i is the subcarrier index. L_i and Q_i are the Channel Frequency Response. $|H_i|$ and $\angle H_i$ are the amplitude and phase of ith subcarrier, respectively. The dimension of H is $N_t \times N_r \times N_s \times T$, where N_t, N_r and N_s are the number of transmitting antennas, receiving antennas and subcarriers, respectively. T is the length of CSI data, and $N_t = 1$, $N_r = 3$ and $N_s = 30$ in our WiFE system. T is a variable depending on the duration of data acquisition:

$$\begin{bmatrix} CSI_{1,1} & CSI_{1,2} & \ldots & CSI_{1,30} \\ CSI_{2,1} & CSI_{2,2} & \ldots & CSI_{2,30} \\ CSI_{3,1} & CSI_{3,2} & \ldots & CSI_{3,30} \end{bmatrix} \tag{3}$$

Formula 3 shows one CSI package obtained at some time point. It is a $1 \times 3 \times 30$ matrix. Each $CSI_{a,b}$ represents a CSI value. a and b denotes the transmit-receive antenna pair and subcarrier number, respectively.

CFR (channel frequency response) can be expressed simply as the superposition of dynamic path CFR and static CFR, and it can be represented as:

$$H(f,t) = H_s(f,t) + H_d(f,t) \tag{4}$$

H_s and H_d represent the static and dynamic CFR, respectively. The dynamic CFR can be written as:

$$H_d(f,t) = \sum_{k \in D} h_k(f,t) e^{-j2\pi f \tau_k(t)} \tag{5}$$

where f and $\tau_k(t)$ represent the carrier frequency and the propagation delay on the k^{th} path, respectively. D is the set of dynamic paths, and h_k is the signal attenuation.

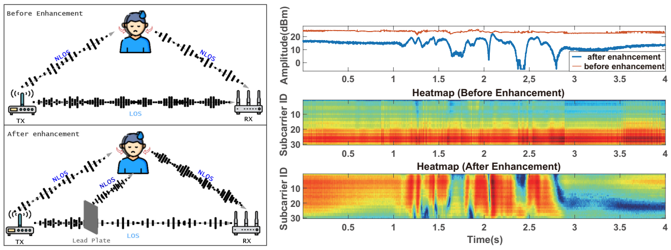


Fig. 2: The effect of CSI enhancement via Ricean fading: the sensitivity of CSI to gesture has been improved by over 4 times

3.1.2 CSI Enhancement

As shown in Fig. 2, a person acts like a mirror to WiFi signals, causing the NLoS signal propagation paths. The signal attenuations caused by a gesture only happen in the NLoS paths of WiFi CSI. Therefore, we could improve the system sensitivity to gestures by amplifying the ratio of the NLoS component corresponding to gestures via weakening the LoS component, e.g., deploying a flat steel plate in front of the transmitting antenna.

In wireless communications, when there is a line of sight between the transmitter and the receiver, the received signal can be written as the sum of a complex exponential and a narrowband Gaussian process, which is known as the ”LOS component”. The ratio of the powers of the LOS component to the whole received power is the Rician factor, which measures the relative strength of the LOS, and represents the link quality. The baseband $x(t)$ In/Quadrature phase (I/Q) representation of the received signal can be modeled as [62]:

$$x(t) = \sqrt{\frac{K\Omega}{K+1}} e^{j(2\Pi f_D cos(\theta_0)t)+\phi_0} + \sqrt{\frac{\Omega}{K+1}} h(t) \tag{6}$$

where K is the Ricean Factor, Ω denotes the total received power and θ_0 and ϕ_0 are the AoA(Angle of Arrival) and phase of the LOS, respectively. f_D is the maximum Doppler frequency, and $h(t)$ is the diffuse components.

In our WiFE system, considering that the antenna layout is fixed, i.e., $f_D = 0$, we can simplify Equation 6 to:

$$x(t) = \sqrt{\frac{K\Omega}{K+1}} e^{\phi_0} + \sqrt{\frac{\Omega}{K+1}} h(t) \tag{7}$$

As described in Equation.4, the received signal can be divided into two parts: static paths and dynamic paths signal, and the received signal has a time-varying amplitude in complex plane according to [63]:

$$|H(f,\theta)|^2 = |H_s(f)|^2 + |H_d(f)|^2 + 2|H_s(f)||H_d(f)|cos\theta \tag{8}$$

θ is the phase difference between the static vector and the dynamic vector, the part that causes the amplitude fluctuation of the CSI waveform is $2|H_s(f)||H_d(f)|cos\theta$. It can be seen that in the case where the range and position of the motion are constant, θ is constant, and the factor affecting the fluctuation is $|H_s(f)|$ and $|H_d(f)|$.

In the case where the torso does not block LOS, all LOS components and part of NLOS components belong to the static path; part of NLOS components belong to the dynamic path. Combined with Equation (7) and set the transmitted power as 1, we can define the $|H_s|$ and $|H_d|$ as follows:

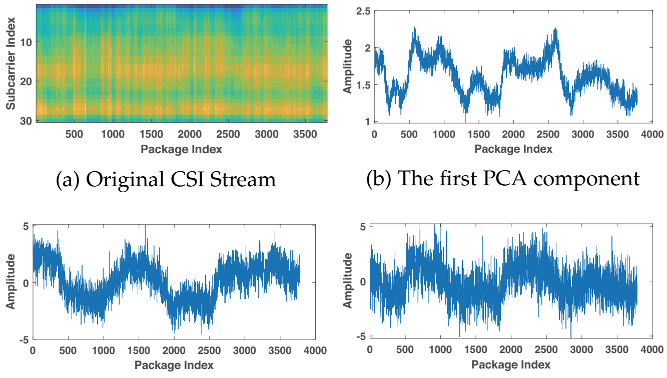
$$|H_s| = \frac{K}{K+1} + \frac{1}{K+1} \cdot \rho \tag{9}$$

$$|H_d| = \frac{1}{K+1} \cdot (1-\rho) \tag{10}$$

Where ρ is the proportion of static paths in the NLOS component.

Combined with Equation (9) and (10), we can get the following equation:

$$\begin{aligned} |H|^2 &= |H_s|^2 + |H_d|^2 + 2|H_s||H_d|cos\theta \\ &= \frac{(K+\rho)^2}{(K+1)^2} + \frac{(1-\rho)^2}{(K+1)^2} \\ &+ \frac{2(K+\rho)(1-\rho)}{(K+1)^2} cos\theta \end{aligned} \tag{11}$$



(a) Original CSI Stream

(b) The first PCA component

(c) The second PCA component

(d) The third PCA component

Fig. 3: An example demonstrating the effect of PCA on the original CSI for noise and dimensionality deduction

Signal amplitude variation caused by motion can be quantified as:

$$f(K,\rho) = 2|H_s||H_d|cos\theta = \frac{2(K+\rho)(1-\rho)}{(K+1)^2} cos\theta \tag{12}$$

The value of the above formula is related to three variables, namely θ, K and ρ. Consider that the change in phase difference caused by same motion is relatively stable, we omit θ without considering.

$$f'(K) = \frac{2(1-\rho)(-K^2 - 2\rho K + 1 - 2\rho)}{(K+1)^4} \tag{13}$$

When $K > 1 - 2\rho$, $f(K,\rho)$ decreases as K increases, under normal circumstances, only a small part of the signal of the omnidirectional antenna can be reflected by the human body, which means that ρ is generally bigger than 0.5. In practice, we could simply deploy an obstacle like a flat steel plate in front of the transmitting antenna to bound off the LoS signal to increase $f(K)$.

Taking the derivative of Equation 12 of ρ, we can get:

$$f'(\rho) = \frac{2(-2\rho + K - 1)}{(K+1)^2} \tag{14}$$

When K does not change, WiFi sensing ability increases in the interval [0,$\frac{1-K}{2}$] as ρ increases, and decreases in the interval [$\frac{1-K}{2}$,1] as ρ increases.

When $K > 1$, $\frac{1-K}{2} < 0$, according to Equation 14, sensing ability monotonously decreases over the interval [0,1] as ρ increase. **When** $0 <= K <= 1$, it means $0.5 >= \frac{1-K}{2} >= 0$, according to equation 14, the sensing ability increases in the interval [0,$\frac{1-K}{2}$] as ρ increase, and decreases in the interval [$\frac{1-K}{2}$,1] as ρ increase. In a real-world environment, K is usually not less than 1 because the quality of WiFi connection would be poor.

Will blocking the LOS make perception capability worse? Blocking the LOS path can reduce the Rice-K value, but will blocking must improve the detection capability? Equation 12 have two main variables, K and ρ. Whether there such a situation that block the LOS path can reduce K, but ρ increased, results in poorer motion perception ability? We believe that such a situation is hard to happen. Even for the worst case, which is blocking the LOS path to increase the NLOS power by L scale, L is not allocated to the dynamic vector at all. At this time, the values of $|H_s|$ and $|H_d|$ are unchanged, their product will not change, and the motion detection capability will not be worse, just equal to the original situation. In the actual indoor environment, it is difficult to make the blocked LOS signal do not propagate towards the human body at all due to reflections from indoor objects. Nevertheless, no matter ρ is larger, smaller, or unchanged, the motion detection capability will not be deteriorated by block the LOS (Unless the occlusion affects the signal reception at receiver).

Fig. 2 demonstrates such an example (the system setting will be introduced later in Section 4.1.2). A centimeter-level gesture like a gentle nod is hard to capture for the original CSI. But after enhancement, it can be clearly recorded. The sensitivity of CSI to the gesture in terms of the amplitude fluctuation has been improved by 4.35 times applying our Rician-K based CSI enhancement method.

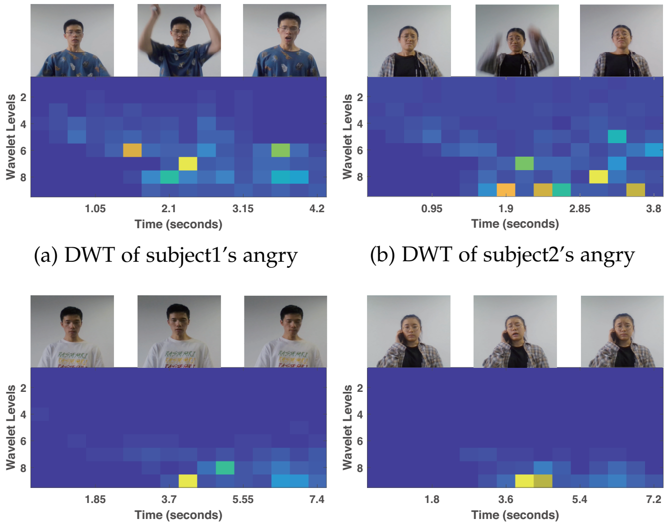
In Section 4, extensive real-world experiments suggest that our CSI enhancement method enables subwavelength level gesture capture with little implementation overhead.

3.1.3 *CSI Feature Extraction*

After enhancement, the CSI data should be preprocessed to remove the impulse background noise and reduce the data dimension. Conventional filters like the low-pass filter are unable to handle the impulse noise due to its high energy and bandwidth [64]. Consequently, we apply Principal Component Analysis (PCA) to fulfill both noise canceling and dimension reduction at the same time [65].

PCA projects each CSI data onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. In our case, a gesture leaves different but correlated influences on different subcarriers due to their continuous wavelengths. Therefore, PCA is able to extract such correlations from the original data while excluding unrelated information, especially the background noise.

Fig. 3 shows an example, where a sad gesture of subject 2 lasting for 3.2 seconds is captured by the enhanced CSI (its corresponding vision snapshots are shown in Fig. 4c). The original CSI can capture the gesture in a vague shape formed by the yellow pixels in Fig. 3b, while this shape has been significantly sharpened in the first three PCA components. Since the first component presents the best performance, we leverage it only instead of all subcarriers in our system, reducing the complexity to $1/30$.



(a) DWT of subject1's angry

(b) DWT of subject2's angry

(c) DWT of subject1's sad

(d) DWT of subject2's sad

Fig. 4: Examples of the correlation feature decomposing a gesture based on DWT

A gesture may consist of a set of movements performed by different parts of the body at different time with different speeds. Therefore, decomposing a gesture would provide a more fine-grained view on the emotional expression. Wang *et al.* showed that the energy profile of different signal frequencies can quantify the movement speeds of different human body parts [64]. Therefore, we apply the DWT, a Time-Frequency analysis tool, on the first PCA component to calculate the energy in different levels at any given time, where each level corresponds to a frequency range. The higher the energy is, the more likely it is caused by a faster moving body part.

Figure 4 shows an example, where both modality data describing two different subjects' emotional cues for the angry and sad emotions have been displayed in a synchronized manner. It is quite interesting to see visually that for both emotions the two subjects share similar ways of expression. On the other hand, angry is an emotion that usually has a much more intensive way of expression than sad involving quick hand and arm movements. Such expressional differences have been successfully captured by the DWT. For instance, Fig. 4a and Fig. 4c recorded the emotional cues of angry and sad for subject 1. DWT has clearly shown their significant energy differences.

After obtaining the DWT feature map, we use three kinds of Densenet(Densenet121,169 and 201) [66] to extract features of different depths as CSI static features, these Densenet are all pre-trained with ImageNet [67], and we use a 512-dimensional LSTM network to extract temporal features from raw CSI data.

3.2 Vision-based Facial Expression Capture

In this part, we introduce how to capture facial expression based on computer vision. Firstly, we explain how to accurately crop faces in a video. Then, we pre-train our deep learning network consisting of the Densenet and VGG [68] using an existing facial expression dataset (FER2013) for better handling the extracted face-related clips. Lastly, we leverage the pre-trained network to explore the spatial and temporal features of these clips.

3.2.1 *Face detection and Alignment*

Accurate and stable face tracking is the key first step for understanding facial expressions. Recently, Zhang *et al.* show that the MTCNN exploits the inherent correlation between face detection and facial expression recognition and constitutes a better solution than the widely-used Dlib detector in handling issues like the changing head position [69]. Therefore, we use the MTCNN approach to crop faces in the video clips and align them at a fixed direction, as shown in Figure 4. The size of each image is 256×256.

3.2.2 *Vision Pre-training*

Pre-training is commonly used in CV due to the existence of many well-constructed and well-labeled image datasets. On one hand, pre-training acts like a initial parameter setter and plays an important role in subsequent supervised training. On the other hand, it relieves the overfitting problem on a small dataset. For example, pre-training on ImageNet, the world's largest labeled image dataset containing 22,000 categories and 15 million images, is a dominant paradigm to initialize the backbones of object detection and segmentation models [70]. However, ImageNet is not a good choice for our problem since it does not include the facial expression. Therefore, we choose FER2013 [71] to tune our network, which consists of Densenets and VGG

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2023.3285777

shown in Fig. 1. FER2013 is a large-scale dataset specifically collected for facial expression by the Google image search API. It is well-known for its quality in gray-scale and its adequate number of facial expressions, which is also the reason we choose it over other datasets.

3.2.3 Vision Feature Extraction

The pre-trained Densenet and VGG-LSTM are used to explore the spatial and temporal characteristics of the face-related clips, respectively. In particular, the Densenet network is responsible for extracting the spatial facial expression features of each video frame, similiar to CSI, we also use three kinds of Densenet for static features extraction. While the VGG is employed to obtain the static features of each frame, and then send them to the LSTM network in chronological order for exploiting the temporal correlations of a facial expression.

For the Densenet, since the cropped face-related clips have different lengths, the dimension of the extracted features for each clip also varies. Therefore, we first normalize the features of a video clip. Then, we calculate the *mean*, *max* and *standard deviation* of the normalized features. In this way, the dimension of features from different face-related clips will be unified. The length of the CSI data is also different, but we unified each CSI data into a same size image when generating DWT images. Therefore a segment of CSI data can only generate one DWT image, and thus gesture feature extraction don't need feature normalization.

3.3 Multi-Source Learning

After obtaining features of gesture and facial expression , we need to effectively fuse them to recover the corresponding emotion. In this part, we introduce our MSL framework.The key idea is that different modalities actually describe the same physical expression of an emotion, but only from different perspectives.

Fig. 5 compares MSL with three prevailing learning frameworks, i.e., early-fusion and late-fusion and MTL. As introduced in Section 2.4, early-fusion performs feature-level aggregation to train one decoder for multiple encoders, while late-fusion trains a decoder for each encoder and fuse the outputs based on voting or retraining, respectively. Early-fusion exploits correlations and interactions between the features of different modality, and it only requires to train one decoder, making the training much easier than late-fusion. However, it needs all modalities to make a decision. In contrast, late-fusion fuses the outputs of all decoders via voting or retraining. Therefore, it is resilient to the modality missing issue. But late-fusion ignores the feature-level interaction between modalities. To fill in the gap, we first focus on a recent framework named Multi-Task Learning (MTL).

As shown in Fig. 5c, MTL is able to use only one model to solve multiple tasks [72] due to its ability of information exchange between tasks. The key idea is that different tasks can share an encoder since the underlying features used in different tasks are similar. Training of each decoder is only affected by the corresponding task, while training of the encoder is affected by all tasks. The weighted losses of different tasks form the overall loss for the encoder.

Inspired by MTL, we can also use only one decoder for all encoders since all encoders actually describe the same event (physical expression of an emotion) from different perspectives. Therefore, we propose MSL to better fuse modalities. Unlike MTL, multi-modal fusion belongs to the single-task learning. It combines multi-modal information for one task. As shown in Figure 5d, MSL uses one encoder for each modality to fully explore the emotional features within each modality, and leverages only one decoder for knowledge sharing like MTL. In this paper, the decoder is composed of three fully-connected layers, and the dimensions of these fully-connected layers are 1080, 1080 and 7, respectively.

During the training process of MSL, one encoder is only affected by its corresponding input. But the decoder is affected by all encoders. The loss of different encoders will be weighted and summed to get the total loss to train the decoder. In this way, knowledge about the emotion can be exchanged between different modalities through the parameter sharing mechanism. The decoder blending features from different encoders (modalities) then helps each modality make its own decision. Lastly, MSL makes the final decision based on voting or re-training on these decisions, like late-fusion.

Compared to early-fusion, MSL is resilient to the modality missing issue since the decoder is shared by all encoders and there is not here is unnecessary to combine the features from different encoders like early-fusion. Compared to late-fusion, the independent training of each decoder is saved since only one shared decoder should be handled, significantly reducing the training efforts.

In this paper, we use softmax loss as the loss function. Specifically, softmax loss is computed as follows:

$$L_s = -\sum_{i=1}^{m} log \frac{e^{w_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{w_j^T x_i + b_j}} \tag{15}$$

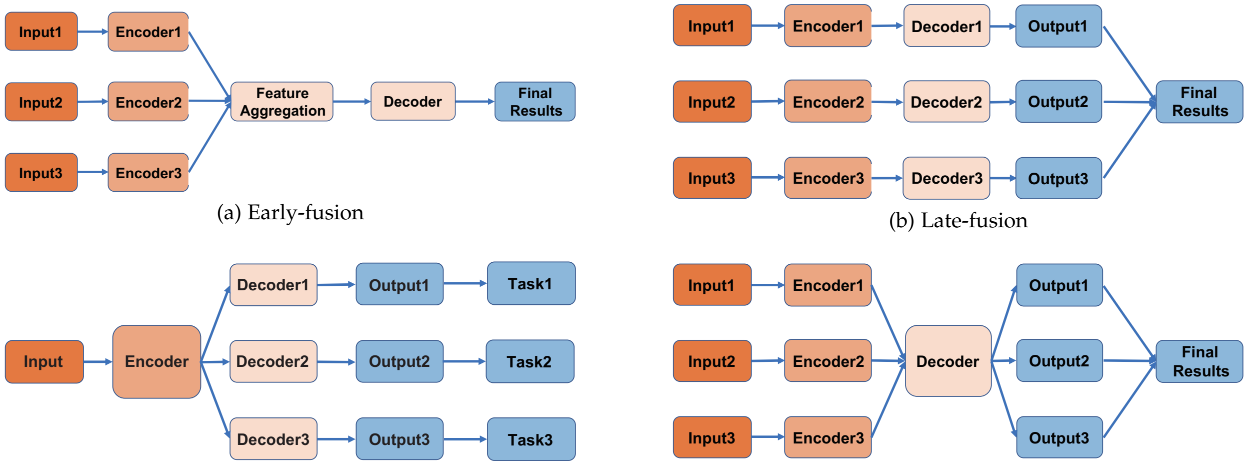
where w_j and b_j are the weight and bias, respectively. $n = |C|$ is the number of classes, and m is the mini-batch size.

The final loss is obtained as follows:

$$L = \sum_{b=1}^{q} L_{s_b} \tag{16}$$

Where q denotes the number of branches, and here we set q to 8 (8 encoders). λ_b is the weighted ratio of the loss of the bth encoders. L_{s_b} is the softmax loss of the bth branch.

In particular, as shown in Fig 1, WiFE leverages MSL based method to handle vision and CSI data, where eight encoders are built to extract temporal and spatial features from both modalities. As explained in the previous parts, these encoders are based on the Densenet, VGG, LSTM and DWT. Due to the final result is obtained at the result layer, our MSL framework can be flexibly split. Since different encoders generate different dimensions of features, they should go through a 1080-dimensional fully-connected layer to unify the feature dimensions before sending to the shared decoder. After obtaining the recognition results of each encoder, we obtain the final result by weighted voting. According to our empirical experiences, the weighted ratios of loss of the eight encoders are set to $5:5:5:1:1:1:2:2$, and the weighted ratios of decision of the eight encoders are



(a) Early-fusion

(b) Late-fusion

(c) Multi-Task Learing (MTL)

(d) Multi-Source Learning (MSL)

Fig. 5: (a) Early-fusion aggregates all features to train one decoder for multiple encoders; (b) Late-fusion trains a decoder for each encoder, and fuse the output based on voting or retraining; (c) MTL uses multiple decoders to exchange knowledge ; (d) MSL processes multiple inputs for one task using only one decoder It exchanges knowledge through sharing decoder parameters.

set to $3 : 4 : 3 : 2 : 6 : 2 : 4 : 2$, respectively. The order of these weights corresponds to Densenet 121, Densenet169, Densenet201 and VGG-LSTM for vision, Densenet 121, Densenet169, Densenet201 and DTW-LSTM for gesture.

4 Dataset Construction and Performance Evaluations

In this section, we first introduce our WiFE dataset, and then use this dataset to evaluate our system.

4.1 WiFE dataset Construction

To the maximum of our knowledge, there exists none WiFi-Vision bi-modal dataset. Therefore, we have to build the first WiFi-Vision emotion dataset to evaluate our system. In this part, we will introduce its design and realization in detail.

4.1.1 Dataset Design

Generally, emotions are associated with facial expression, speech, and body gestures. To make the dataset more meaningful, the following two requirements should be satisfied: the associated gestures should be coherent with the emotions, and different subjects' emotions should not be identical (i.e., subjects express their emotions independently).

As to the first requirement, we choose *Acted Facial Expressions in the Wild* (AFEW) [73] dataset to help gain knowledge on how others behave with different emotions. AFEW is a temporal and multi-modal dataset that provides vastly different environmental conditions in both audio and video, and it contains clips with spontaneous expressions collected from various movies/TV series. We learn from the AFEW dataset to help select emotion representation templates. To be more specific, we choose seven universal emotions to build our dataset and choose templates for each emotion. Hence emotion reactions will not be dispersed. We select the videos with body gestures in the AFEW and show them to the volunteers who participate in data collection. Then the volunteers vote for the five most popular templates (we assume that these five templates are most reasonable) as the final templates used in the dataset building. As to the second requirement, the templates are used only as guidance instead of rules. Figure 6 shows the anger emotion expressed by different subjects according to the same template, where we can observe that each person retains his/her own independence.

4.1.2 Dataset Collection and Brief Introduction of Data

During the data collection process, we use a laptop computer to gather the video and two Mini PC with four antennas to obtain the CSI. The MiniPC is with Ubuntu 12.04, one transmitting (T_r) antenna sends WiFi signals (can be replaced by a regular WiFi router), and the other three receiving (R_x) antennas receive WiFi signals and extract CSI data using CSITool [27]. As illustrated in Figure 6, the laptop is placed in the center to collect vision information, and we put the WiFi antennas on both sides of the shelf for collecting the CSI data containing gesture information. A lead plate is placed between T_r and R_{x3} to increase gesture sensitivity, which will be further explained in Section4.4.1.

In general, WiFE has 7 emotions (i.e. *Angry*, *Disgust*, *Fear*, *Happy*, *Neutral*, *Sad* and *Surprise*), 35 templates (5 templates are selected for each emotion). Twenty volunteers (12 male and 8 female, whose ages range from 22 to 26) repeat each template five times. Finally, 3500 video clips and the corresponding CSI sequences are collected. Each video clip has a frame rate of 30Hz and a resolution of 720p. For video data, we provide the original video and the cropped video containing only the blocks of the facial expressions. These videos are saved in mp4/avi format. The packet rate used for collecting CSI data is 500 packets/second, and the CSI file contains data of 90 subcarriers of 3 receiving antennas. For CSI in WiFE, we provide raw data (the suffix is .dat) and cropped data, which only contains emotion-related actions (the suffix is .mat). The total size is 82GB.

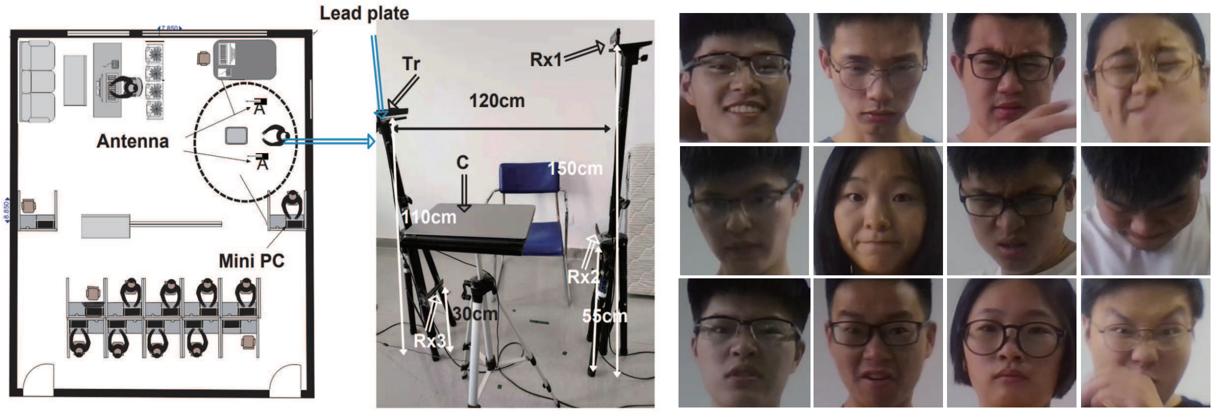
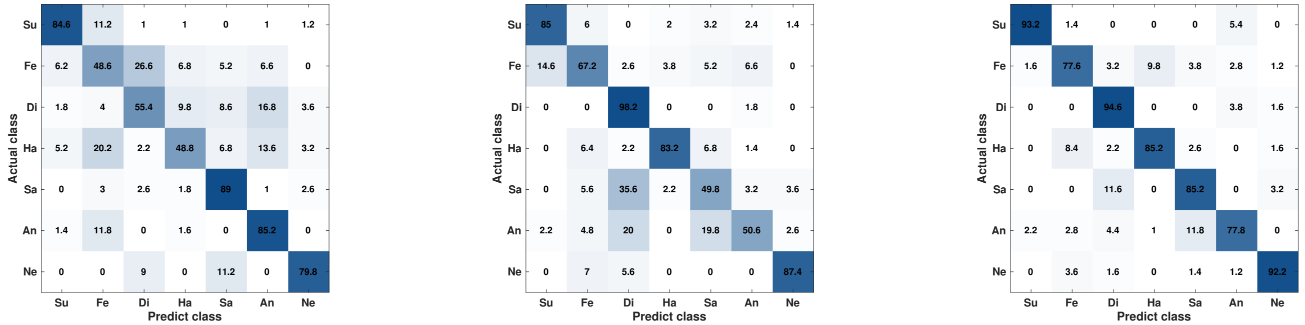


Fig. 6: A snapshot of our WiFE system and its dataset



(a) Gesture-only

(b) Vision-only

(c) Bi-modality

Fig. 7: (a) Gesture-only; (b) Facial-only; (c) Bi-modality.

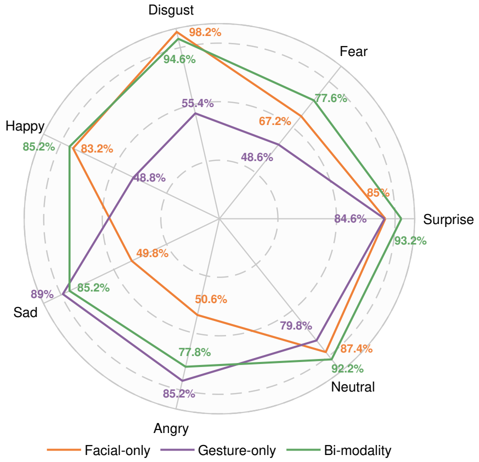


Fig. 8: Comparison between gesture-only (CSI), facial-only (Densenet+VGG-LSTM) and bi-modality.

4.2 Overall Performance

We systematically evaluate WiFE on our dataset via ten-fold subject-independent cross-validation. Fig. 7 presents three confuse matrices corresponding to gesture-only, vision-only and bi-modal settings, respectively.

Firstly, we confirm that bi-modality indeed improves the overall recognition accuracy over single-modality by achieving 86.54% over 70.2% for gesture-only and 74.51% for vision-only. Moreover, it indicates that WiFE can effectively leverage correlations of emotional clues in two modalities for better recognizing emotions.

Secondly, physical clues of the same emotion in different modalities could be complementary to each other. Fig. 8 shows a direct comparison among three different settings, where the orange line, purple line and green line denote facial-only, gesture-only and bi-modality, respectively. We take happy as an example, which ranks second-last (48.8%) for gesture-only, ranks second (83.2%) in facial. On the other hand, angry, poorly recognized via facial expression (50.6%) while has the highest recognition accuracy (85.2%) for gesture-only. This observation leads to another important question: ***What is the best combination of modalities for recognizing some emotion?*** It calls for future research with more sophisticated psychological experiments driven by state-of-the-art computational methods.

Lastly, facial expression constitutes a better descriptor of emotions than gesture. On average, it achieves 74.51% recognition accuracy compared to 70.2% for gesture. For some specific emotions with unique expressions like disgust, facial expression is almost 2 times better that gesture (98.2% vs 55.4%). However, for some emotions which usually accompany large-differentiated gestures, the gesture modality can achieves better results, such as angry (85.2% vs 50.6%) and sad (89% vs 49.8%). The recognition performance of different emotions varies significantly, and we believe this is because different emotions are expressed in different modalities with different degrees of variation. For example, angry, sad and disgust are easily confused in facial expressions (they are very close to each other in the continuous Valence-Arousal emotional space). But in gesture modality, angry is usually accompanied by intensive body movements and can be easily distinguished from sad and disgust. On the other hand, for fear, its body movements are easily confused with disgust (moving away from a feared or disgusted object). But, in the facial expression modality, the difference between fear and disgust is more significant.

4.3 Comparative Study with State-of-the-art Rivals

In this part, we compare our method with some current solutions based on facial, gesture and multimodal emotion recognition.

As shown in Table 2, we have compared our method with some other emotion recognition schemes. For facial expression recognition, MSL achieves a higher accuracy (69.2%) than the region attention based method RAN (67.09%) [74] and the multi-attention based scheme DAN (67.69%) [44]. This is because MSL incorporates the results of three kinds of Densenet, which enables layered features extraction, while the other two schemes focus on extracting more effective features on one kind of backbone network.

For gesture recognition, our previous work (EmoSense [17]) leverages shallow learning approaches (KNN,SVM), and only gets 57.97% recognition accuracy. WiGRUNT relies on a dual attention network and achieves the best result (68.29%) while MSL delivers an recognition accuracy of 67.94%. This is because gesture contains less emotional cues than facial expression, and there exists no dedicated datasets large enough for pre-training the CSI-related network. As a result, the attention mechanism seems more effective under this situation.

TABLE 2: Comparison of Methods

Method	Modality	Feature	Accuracy
RAN [74]	Facial Expression	Static	67.09%
DAN [44]	Facial Expression	Static	67.69%
MSL	Facial Expression	Static	69.2%
Emosense [17]	Gesture	Static	57.97%
WiGRUNT [75]	Gesture	Static	68.29%
MSL	Gesture	Static	67.94%
Tzirakis [76]	Bimodality	Static+Temporal	79.97%
Ortega [60]	Bimodality	Static+Temporal	86.29%
MSL	Bimodality	Static+Temporal	86.54%

For multimodal emotion recognition, we compare our method with a early-fusion method [76] as well as a late-fusion scheme [60]. This time MSL achieves the best recognition results (86.54%). Moreover, the accuracy of multimodal-based emotion recognition is much better than unimodal solutions, confirming the superiority of our solution and multimodal emotion recognition.

4.4 Impact of Different Settings on Performance

In this part, we will go through the design flow of WiFE and study the impact of different settings on the performance.

TABLE 3: Impact of CSI enhancement

CSI Setting	Accuracy
Original CSI	68.67%
Enhanced CSI	**83.67%**

4.4.1 Impact of CSI Enhancement Method

Firstly, we consider the effectiveness of our CSI enhancement method. In Section 3.1.2, we used an example to show that a centimeter-level gesture (a gentle nod) is clearly recorded after CSI enhancement. More specifically, the sensitivity of CSI to gesture in terms of the amplitude fluctuation has been improved by 4.35 times after enhancement. However, constructing the dataset again with the original CSI is both time-consuming and labor-intensive. Moreover, it is extremely difficult to make sure the volunteers have the exact facial expression and gesture for both settings. Therefore, we construct a miniature dataset consisting of two trained volunteers performing three emotions (angry, surprise and happy) while keeping other settings. For each CSI setting, we have 2(volunteers)× 3(emotions)× 5(templates)×5(times)= 150 entries. Table 3 concludes the result, where the original CSI and enhanced CSI achieve 68.67% and 83.67% recognition accuracy on average, respectively. Therefore, our CSI enhancement method has improved the performance by 21% on this mini dataset. We also notice that the average accurate obtained on this mini dataset is quite close to our WiFE dataset (83.67% vs 86.54%).

4.4.2 Impact of different Encoder Settings

TABLE 4: Impact of different Encoder Settings

Combination	Accuracy
Vision Static	69.2%
CSI Static	67.94%
Vision Static + Vision Temporal	74.51%
Vision Static + CSI Static	81.97%
Vision Static + CSI Temporal	75.03%
CSI Static + CSI Temporal	70.2%
CSI Static + Vision Temporal	77.97%
All	**86.54%**

MSL employs multiple encoders to extract temporal-spatial features from different modalities. Therefore, it is important to study the combinations of encoders on the system performance. To this end, Table 4 shows that the impact of four different combinations of encoders in recognition accuracy. If only Densenet is used for vision and CSI spatial features, 69.2% and 67.94% accuracy can be achieved, respectively. If we add VGG-LSTM for the vision temporal features, we could improve the accuracy to 74.51%, which is the performance upper-bound for vision-only. But if we add CSI static (gesture), the recognition accuracy reaches 81.97%, and CSI static + vision temporal also performance better than CSI static + CSI temporal. In other words, a new modality may compensate for existing modalities and thus advance the overall performance, just like we explained in Fig. 8. If we fully leverage the spatial-temporal features of both modalities, the performance can be further improved to 86.54%.

TABLE 5: Impact of different fusion schemes

Scheme	Accuracy
Late-fusion	86.29%
Early-fusion	79.97%
Multi-Source Learning (MSL)	**86.54%**

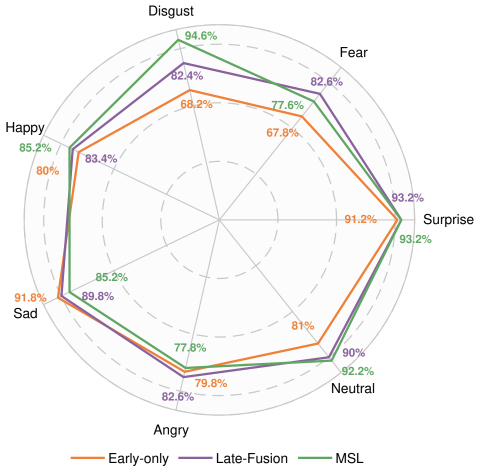


Fig. 9: Recognition accuracy of different fusion schemes over 7 emotions.

4.4.3 Impact of MSL

In this part, we compare MSL with current mainstream multi-modal fusion schemes (early-fusion and late-fusion). For early-fusion, we combine all modality-related features to make the final decision through the same neural network structure as in [60], [76]. For late-fusion and MSL, we use the weighted voting to get the final decision after obtaining the recognition result of every single modality. For all three methods, the decoder structure is the same (two 1080-dimensional fully connected layers and one 7-dimensional fully connected layer.), and the evaluation results are shown in Table 5. MSL yields the best performance compared to other fusion schemes. It indicates that the knowledge sharing mechanism at feature-level is beneficial for emotion recognition.

Figure 9 demonstrates the recognition accuracy of different emotions of the three fusion schemes. For early-fusion, the recognition results are not even over 7 emotions. Some emotions like disgust (68.2%) have low recognition accuracy, while some other emotions like sad (91.8%) can be easily identified. This is because early-fusion combines all the modal features to train the classifier, during which it pursues the highest overall accuracy, leading to the unbalance performance over different emotions. Late-fusion and MSL do not present such phenomenon but due to different reasons. For late-fusion, fusion happens in the decision-level and does not affect each modality. For MSL, knowledge sharing in the decoder ensures that the blended features help optimize each encoder.

TABLE 6: Impact of modality-missing

	Early-fusion	Late-Fusion	MSL
CSI-only	/	68.06%	**70.2%**
Vision-only	/	**75%**	74.51%
Bi-modality	79.97%	86.29%	**86.54%**

TABLE 7: Comparison of network complexity between MSL and Late-Fusion

Schemes	Parameters	Running time (per sample)
Late-Fusion	35,344,129	9.1ms
MSL	**23,363,647**	**2ms**

Table 6 describes the performance variation for the modality-missing issue, where we consider three different modality settings, i.e, CSI-only, vision-only and bi-modality. For early-fusion, it requires features of every encoders (modality) to make the fusion decision. Therefore, it fails to work if missing any encoder. Late-fusion achieves 68.06%, 75% and 86.29% for the three setting, respectively. As for MSL, late-fusion also witnesses the performance improvement from single-modality to bi-modality. Moreover, it backs up our previous observation that vision outperforms CSI with MSL, implying that facial expression constitutes a stable and reliable modality with rich emotional cues.

4.4.4 Efficiency of MSL

Efficiency is another key factor affecting the practicality of the proposed method. To this end, we compare the computational complexity between later-fusion and MSL in Table 7. It is clear that late-fusion embodies $35,344,129$ parameters in its deep network while MSL reduces the number by 33.9%, i.e., $11,980,482$ network parameters. Therefore, a less complicated network leads to faster running time. Late-fusion takes 5.3 ms to process every test sample on average while MSL only needs 2 ms. In other words, MSL is 4.55 times faster than late-fusion.

4.4.5 Evaluations on RAF-DB

Another possible concern is that the vision-only results reported on our self-collected dataset sometimes do not align with existing vision-only literature. To this end, we have conducted experiments on RAF-DB [41], and achieved an overall recognition accuracy of 88.69%. Specifically, the recognition accuracy for different emotion classes are 88.15%, 55.41%, 65.53%, 5.19%, 87.87%, 80.86% and 89.12% (Surprise, Fear, Disgust, Happy, Sad, Anger and Neutral).

We think the difference in results are due to the presence of actions and occlusions in our dataset. For sadness, our dataset has actions such as wiping tears, bowing the head, and sobbing that obscure the face. In contrast, the images captured by facial expression datasets such as RAF-DB are selected, so the data distribution is not the same as our dataset.

5 LIMITATIONS

In this section, we focus on several shortcomings of the proposed framework.

Raised costs. Combing different modalities in emotion sensing naturally leads to raised costs in data collection and processing. On one hand, extra devices are needed compared to single-modality solutions, raising costs in hardware, design as well as deployment. On the other hand, the collected multimodal data are heterogeneous in nature, and thus the data learning network becomes more complex and sophisticated, significantly raising computational costs in training. In our case, multiple complex feature extractors are designed to ensure recognition accuracy, making the training process both computation-intensive and time-consuming. On possible solution is to use more efficient feature extraction methods (like RAN [74] and DAN [44]) to build multimodal learning models using only 1-2 feature extractors for each modality to significantly reduce the training complexity.

Privacy concerns. Privacy exposure is a major concern in real-world affective applications and vision constitutes the most worried sensing channel. To this end, we could substitute the RGB camera with a depth camera, which only captures the depth information of the monitored targets to reduce the risk of privacy breaches. Naturally, the information loss, especially in facial expressions, could lead to performance degeneration in recognition accuracy.

Wifi sensing is only capable of recording target movements due to its inherent sensing paradigm, and thus it poses very little privacy exposure risks compared to the visual channel.

Performance drop facing data missing. Another shortcoming is that data missing in a modality in a multimodal network could significantly degenerate the overall performance. Though our proposed MSL method can ease this issue, we still observe performance downgrade in experiments. We believe this is because we do not consider inter-modal interactions in depth. The semantics of the multimodal data we acquire is consistent (label consistent) and we are considering using neural-symbolic learning to improve the MSL architecture. Specifically, we could build an end-to-end multi-modal learning architecture and use semantic consistency to supervise the training of feature extractors for different modalities and the decoder.

Inapplicable situations. The proposed multi-modal methods could fail in the following situations: 1) facial covering. Facial expression is key to emotion recognition. As with other state-of-the-art research using vision for emotion recognition, missing facial expressions could significantly damage system performance. 2) motion artifacts. Motion artifacts could interfere with the wifi channel data in multipath propagation and thus jeopardize the network in recognizing emotional information in body movements. Specifically, different from the wearable sensors or vision-based methods, the former is directly worn on the user, so it will not be disturbed by motion artifacts, and the latter has a very fine-grained spatial discrimination and can easily distinguish different motions. The principle of wireless sensing is that electromagnetic waves propagating in space will be reflected when they touch the human body. The movement of the human body will cause the reflection path to change. Therefore, at the receiver, changes in signal indicators (RSS or CSI, etc.) caused by changes in the reflection path are used to perform tasks such as activity recognition. And the signal path changes caused by the activities of multiple motions will be confused and cannot be distinguished.

6 Conclusion and Future Work

This paper proposed a hybrid emotion recognition system leveraging two emotion-rich and tightly-coupled modalities, i.e, facial expression and body gesture. Unlike our rivals relying on contact or even invasive sensors, we explored the commodity WiFi signal for device-free and contactless gesture recognition, while adopting a vision-based facial expression. We proposed a signal sensitivity enhancement method based on the Rician-K factor and a Multi-Source Learning (MSL) method to mine the temporal-spatial features of bi-modal data to process the large-volume and heterogeneous data contributed by the two-modalities. We build a first-of-its-kind WiFi-Vision emotion dataset (WiFE dataset) and a prototype system on low-cost commodity WiFi and vision devices to evaluate the proposed method. The empirical results show the superiority of the bi-modality by achieving 85.33% recognition accuracy for seven emotions, as compared with 68.95% and 72% recognition accuracy by gesture-only based solution and facial-only based solution, respectively. Moreover, we also confirm the efficiency of our CSI enhancement method and MSL framework by comparing them with state-of-the-art rivals.

For the future work, we intend to significantly extend our current dataset with more volunteers and more emotional templates to approach the AFEW dataset. A natural and large-scale emotional bi-modal dataset is both critical to fine-tune our system and to the psychological understandings on emotions. Moreover, the parameter sharing mechanism of MSL provides the possibility of knowledge exchange for different modalities. But this exchange may bring positive knowledge and negative knowledge at the same time. The former is helping and the latter could significantly deteriorate the system performance. Therefore, we believe that the attention mechanism to the parameter sharing is critical to overcome this defect.

References

[1] M. Minsky, "The society of mind," in *The Personalist Forum*, vol. 3, no. 1. JSTOR, 1987, pp. 19–32.

[2] A. Majumder, L. Behera, and V. K. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 103–114, 2018.

[3] C. Darwin, *The expression of the emotions in man and animals*. Oxford University Press, 1998.

[4] P. Ekman, *Basic Emotions*. John Wiley & Sons Ltd, 2008.

[5] C. Izard, J. Kagan, and R. Zajonc, *Emotions, Cognition, and Behavior*. Cambridge University Press, 1984.

[6] D. Matsumoto, *Facial Expressions of Emotions*. New York: Guilford Press, 2018.

[7] R. W. Picard, *Affective Computing*. MIT Press, 2000.

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[9] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022.

[10] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[11] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2019.

[12] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and eeg for multimodal emotion recognition," *Computational intelligence and neuroscience*, vol. 2017, 2017.

[13] G. Castellano, G. Caridakis, A. Camurri, K. Karpouzis, G. Volpe, and S. Kollias, "Body gesture and facial expression analysis for automatic affect recognition," *Blueprint for affective computing: A sourcebook*, pp. 245–255, 2010.

[14] Y. Yan and Y.-J. Zhang, "State-of-the-art on video-based face recognition," in *Encyclopedia of Artificial Intelligence*. IGI Global, 2009, pp. 1455–1461.

[15] Y. Gu, X. Zhang, H. Yan, Z. Liu, and Y. Ji, "Real-time vital signs monitoring based on cots wifi devices," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1320–1324.

[16] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 313–325.

[17] Y. Gu, Y. Wang, T. Liu, Y. Ji, Z. Liu, P. Li, X. Wang, X. An, and F. Ren, "Emosense: Computational intelligence driven emotion sensing via wireless channel data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 3, pp. 216–226, 2020.

[18] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[19] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[20] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[21] D. Kim, O. Hilliges, S. Izadi, A. D. Butler, J. Chen, I. Oikonomidis, and P. Olivier, "Digits: freehand 3d interactions anywhere using a wrist-worn gloveless sensor," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*, 2012, pp. 167–176.

[22] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*. Ieee, 2011, pp. 1297–1304.

[23] S. Gupta, D. Morris, S. Patel, and D. Tan, "Soundwave: using the doppler effect to sense gestures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1911–1914.

[24] Y. Ma, G. Zhou, and S. Wang, "Wifi sensing with channel state information: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.

[25] X. Zhang, Y. Gu, H. Yan, Y. Wang, M. Dong, K. Ota, F. Ren, and Y. Ji, "Wital: A cots wifi devices based vital signs monitoring system using nlos sensing model," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 3, pp. 629–641, 2023.

[26] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "Wifi-based real-time breathing and heart rate monitoring during sleep," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[27] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11 n traces with channel state information," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 1, pp. 53–53, 2011.

[28] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "Besense: Leveraging wifi channel data and computational intelligence for behavior analysis," *IEEE Computational Intelligence Magazine*, vol. 14, no. 4, pp. 31–41, 2019.

[29] J. Liu, Y. Chen, Y. Wang, X. Chen, J. Cheng, and J. Yang, "Monitoring vital signs and postures during sleep using wifi signals," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2071–2084, 2018.

[30] F. Wang, S. Zhou, S. Panev, J. Han, and D. Huang, "Person-in-wifi: Fine-grained person perception using wifi," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5452–5461.

[31] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[32] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 38–50, 2016.

[33] Y. Gu, H. Yan, X. Zhang, Z. Liu, and F. Ren, "3-d facial expression recognition via attention-based multichannel data fusion network," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[34] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[35] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European conference on computer vision*. Springer, 2016, pp. 425–442.

[36] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6248–6257.

[37] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[38] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6897–6906.

[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[40] S. Khan, L. Chen, and H. Yan, "Co-clustering to reveal salient facial features for expression recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 348–360, 2017.

[41] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[42] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2402–2411.

[43] H. Zhang, W. Su, J. Yu, and Z. Wang, "Weakly supervised local-global relation network for facial expression recognition," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 1040–1046.

[44] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *arXiv preprint arXiv:2109.07270*, 2021.

[45] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "Arbee: Towards automated recognition of bodily expression of emotion in the wild," *International Journal of Computer Vision*, vol. 128, no. 1, pp. 1–25, 2020.

[46] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[47] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection," *IEEE Transactions on affective computing*, vol. 3, no. 1, pp. 102–115, 2011.

[48] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, 2018.

[49] H. K. Meeren, C. C. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proceedings of the National Academy of Sciences*, vol. 102, no. 45, pp. 16 518–16 523, 2005.

[50] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[51] M. Shiffrar, M. D. Kaiser, and A. Chouchourelou, "Seeing human movement as inherently social," *The science of social vision*, pp. 248–264, 2011.

[52] B. Pease and A. Pease, *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam, 2008.

[53] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2017.

[54] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proceedings of the 22nd annual international conference on mobile computing and networking*, 2016, pp. 95–108.

[55] A. N. Khan, A. A. Ihalage, Y. Ma, B. Liu, Y. Liu, and Y. Hao, "Deep learning framework for subject-independent emotion detection using wireless signals," *Plos one*, vol. 16, no. 2, p. e0242946, 2021.

[56] C. Liu, T. Tang, K. Lv, and M. Wang, "Multi-feature based emotion recognition for video clips," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 630–634.

[57] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4011–4018, 2019.

[58] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[59] Y. Jiang, W. Li, M. S. Hossain, M. Chen, A. Alelaiwi, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Information Fusion*, vol. 53, pp. 209–221, 2020.

[60] J. D. Ortega, P. Cardinal, and A. L. Koerich, "Emotion recognition using fusion of audio and video features," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 3847–3852.

[61] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2051–2060.

[62] C. Tepedelenlioglu, A. Abdi, and G. B. Giannakis, "The ricean k factor: estimation and performance analysis," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 799–810, 2003.

[63] H. Wang, D. Zhang, J. Ma, Y. Wang, Y. Wang, D. Wu, T. Gu, and B. Xie, "Human respiration detection with commodity wifi devices: do user location and body orientation matter?" in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 25–36.

[64] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st annual international conference on mobile computing and networking*, 2015, pp. 65–76.

[65] R. Zhou, M. Hao, X. Lu, M. Tang, and Y. Fu, "Device-free localization based on csi fingerprints and deep neural networks," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2018, pp. 1–9.

[66] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[68] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[69] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[70] T. Kaneko, K. Hiramatsu, and K. Kashino, "Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016, pp. 327–331.

[71] I. J. G. *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International conference on neural information processing*, 2013, pp. 117–124.

[72] S. Chennupati, G. Sistu, S. Yogamani, and S. A Rawashdeh, "Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[73] A. Dhall, R. Goecke, S. Lucey, T. Gedeon *et al.*, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[74] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[75] Y. Gu, X. Zhang, Y. Wang, M. Wang, H. Yan, Y. Ji, Z. Liu, J. Li, and M. Dong, "Wigrunt: Wifi-enabled gesture recognition using dual-attention network," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 736–746, 2022.

[76] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

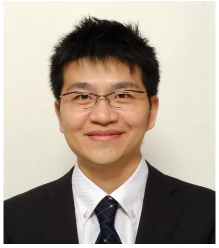
Yu Gu (M'10-SM'12) received his B.E. degree from the Special Classes for the Gifted Young, University of Science and Technology of China, Hefei, China, in 2004, and his D.E. degree from the same university in 2010. In 2006, he was an Intern with Microsoft Research Asia, Beijing, China, for seven months. From 2007 to 2008, he was a Visiting Scholar at the University of Tsukuba, Tsukuba, Japan. From 2010 to 2012, he was a JSPS Research Fellow with the National Institute of Informatics, Tokyo, Japan. Since 2012, he has been a Professor and Dean Assistant at the School of Computer and Information, Hefei University of Technology. Now he is a Professor of the University of Electronic Science and Technology of China since 2023. His current research interests include pervasive computing and affective computing. He was the recipient of the IEEE Scalcom 2009 Excellent Paper Award, NLP-KE2017 Best Paper Award, and IEEE CCIS 2018 Best Student Paper Award. He is a member of ACM and a senior member of IEEE.

Xiang Zhang received the B.E. degree from Hefei University of Technology, China, in 2017, and his Ph.D. degree from the same university in 2023. In 2022, he was a Visiting Scholar at the National Institute of Informatics, Tokyo, Japan. Now he is a Postdoc of the University of Science and Technology of China since 2023. His research interests include intelligent information processing and wireless sensing and affective computing. He is a TPC Member of IEEE ICME and Globecom. He has served as a reviewer for IEEE TNNLS, Information Fusion, Pervasive and Mobile Computing, Pattern Recognition, and Ad Hoc Network.

Huan Yan received the B.E. degree from Hefei University of Technology, China, in 2017, and his D.E. degree from the same university in 2023. His research interests include intelligent information processing and wireless sensing and affective computing.

Jingyang Huang received the B.Eng. degree from Anhui University, China in 2017, and received the Ph.D. degree at School of Cyberspace Security from University of Science and Technology of China in 2022. Now, he is a Lecturer at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), and a member of the HFUT-S2AC Group. His research interests lie Human-computer interaction, Wireless sensing, Wireless communication, and Machine learning.

Zhi Liu (SM11-M14) received the Ph.D. degree in informatics in National Institute of Informatics. He is an associate professor at The University of Electro-Communications, Japan. His research interest includes video network transmission, vehicular networks and mobile edge computing. He is a member of IEICE and a senior member of IEEE.



Mianxiong Dong received B.S., M.S. and Ph.D. in Computer Science and Engineering from The University of Aizu, Japan. He is the youngest ever Vice President and Professor of Muroran Institute of Technology, Japan. He was a JSPS Research Fellow with School of Computer Science and Engineering, The University of Aizu, Japan and was a visiting scholar with BBCR group at the University of Waterloo, Canada supported by JSPS Excellent Young Researcher Overseas Visit Program from April 2010 to August 2011. Dr. Dong was selected as a Foreigner Research Fellow (a total of 3 recipients all over Japan) by NEC C&C Foundation in 2011. He is the recipient of IEEE TCSC Early Career Award 2016, IEEE SCSTC Outstanding Young Researcher Award 2017, The 12th IEEE ComSoc Asia-Pacific Young Researcher Award 2017, Funai Research Award 2018 and NISTEP Researcher 2018 (one of only 11 people in Japan) in recognition of significant contributions in science and technology. He is Clarivate Analytics 2019 Highly Cited Researcher (Web of Science).

Fuji Ren received his Ph. D. degree in 1991 from the Faculty of Engineering, Hokkaido University, Japan. From 1991 to 1994, he worked at CSK as a chief researcher. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor of the Faculty of Engineering, Tokushima University. He is a Chair Professor of University of Electronic Science and Technology of China from 2022. His current research interests include Natural Language Processing, Artificial Intelligence, Affective Computing, Emotional Robot. He is the Academician of The Engineering Academy of Japan and EU Academy of Sciences. He is a senior member of IEEE, Editor-in-Chief of the International Journal of Advanced Intelligence, a vice president of CAAI, a Fellow of The Japan Federation of Engineering Societies, a Fellow of IEICE, and a Fellow of CAAI. He is the President of the International Advanced Information Institute, Japan.