WiFE: WiFi and Vision based Intelligent Facial-Gesture Emotion Recognition

Yu Gu[1] , **Xiang Zhang**[1] , **Zhi Liu**[2] and **Fuji Ren**[3]

[1]Hefei University Of Technology
[2]Shizuoka University
[3]Tokushima University

Abstract

Emotion is an essential part of Artificial Intelligence (AI) and human mental health. Current emotion recognition research mainly focuses on single modality (e.g., facial expression), while human emotion expressions are multi-modal in nature. In this paper, we propose a hybrid emotion recognition system leveraging two emotion-rich and tightly-coupled modalities, i.e., facial expression and body gesture. However, unbiased and fine-grained facial expression and gesture recognition remain a major problem. To this end, unlike our rivals relying on contact or even invasive sensors, we explore the commodity WiFi signal for device-free and contactless gesture recognition, while adopting a vision-based facial expression. However, there exist two design challenges, i.e., how to improve the sensitivity of WiFi signals and how to process the large-volume, heterogeneous, and non-synchronous data contributed by the two-modalities. For the former, we propose a signal sensitivity enhancement method based on the Rician K factor theory; for the latter, we combine CNN and RNN to mine the high-level features of bi-modal data, and perform a score-level fusion for fine-grained recognition. To evaluate the proposed method, we build a first-of-its-kind Vision-CSI Emotion Database (VCED) and conduct extensive experiments. Empirical results show the superiority of the bi-modality by achieving 79.05% recognition accuracy for seven emotions, as compared with 66.48% and 66.67% recognition accuracy by gesture-only based solution and facial-only based solution, respectively. The VCED database download link is https://drive.google.com/open?id=1OdNhCWDS28qT21V8YHdCNRjHLbe042eG or https://pan.baidu.com/s/1BJ7LglQ0qUZheD9VF5_JDA(extract code:h0vu).

1 Introduction

Three decades ago, Minsky in his book *The Society of Mind* asked: "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions" [Minsky, 1987]. Since then, emotion becomes an essential part of AI. It's also a key piece of human mental health [Gross and Muñoz, 1995], and emotion recognition rises as a fundamental issue [Cowie *et al.*, 2001]. It is proved to be effective in many areas, e.g., analyzing the EEG (Electroencephalography) for fatigue detection [Chai *et al.*, 2016] and emotional regulation [Yen *et al.*, 2018]. However, emotion recognition is challenging since emotion expression is person-dependent and multi-modal in nature [Anagnostopoulos *et al.*, 2015].

Previous research in this topic mainly focuses on single modality like facial expression [Li and Deng, 2018] or speech [Albanie *et al.*, 2018]. But because emotional expression is diversified with different people, a single modality may not be able to capture the real emotion fully. For example, some people tend to rely on body language to express emotion rather than facial expression. Thus, there is a recent trend of exploring multi-modality for more reliable and accurate emotion recognition, e.g., facial-audio [Noroozi *et al.*, 2019] or facial-EEG [Huang *et al.*, 2017]. However, the correlation between two modalities still remains a major problem.

In general, facial expression and body gesture are emotion-rich and tightly-coupled in emotion expression. Yan and Zhang [Yan and Zhang, 2009] find that traditional image-based facial recognition system is not accurate enough, but gestures can help people better-analyzing emotions behind people's facial expressions, where vision is used for capturing the spontaneous facial expression and body gesture [Gunes and Piccardi, 2006]. But due to the line-of-sight (LOS) constraint, a camera may not be able to cover the facial expressions and full-body gestures, leading to information loss. It is well known that WiFi signals can be affected by human presence or activities due to the multi-path effect.

To this end, we leverage commodity WiFi for a full-coverage, contactless, and fine-grained gesture recognition while adopting vision for facial expression recognition. There exist two design challenges, i.e., how to improve the sensitivity of WiFi signals and how to process the large-volume, heterogeneous, and nonsynchronous data contributed by the two-modalities. For the former, we propose a signal sensitivity enhancement method based on the Rician K factor theory; for the latter, we combine CNN and RNN to mine the high-level features of bi-modal data, and perform a score-

level fusion for fine-grained recognition. To evaluate the proposed method, we build a first-of-its-kind Vision-CSI Emotion Database (VCED) and conduct extensive experiments. Our contributions are summarized as follows:

- We build the first open Vision-CSI Emotion Database (VCED) for facial expression-gesture bimodal emotion recognition. VCED contains a total of 1750 video clips, and corresponding CSI sequences (data collected from ten volunteers, containing seven emotions), which will be publicly available to researchers free of charge. We further propose a WiFi CSI sensitivity enhancement method based on the Rician-K factor theory to collect better CSI data.

- To the best of our knowledge, we are the first to propose the facial expression-gesture bi-modal for emotion recognition with video and CSI data.

- We have implemented a prototype system which combines CNN and RNN to mine the high-level features of bi-modal data, and perform a score-level fusion for fine-grained recognition. The prototype is evaluated with extensive experiments, and the results demonstrate the effectiveness of the propose facial expression-gesture bi-modal emotion recognition method.

2 Vision-CSI Emotion Database

2.1 Database Design

Generally, emotions are associated with facial expression, speech, and body gestures. To make the database 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 to use *Acted Facial Expressions in the Wild* (AFEW) [Dhall *et al.*, 2012] dataset to help gain knowledge on how others behave with different emotions. AFEW is a temporal and multi-modal database 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 database to help to select emotion representation templates. To be more specific, we choose seven universal emotions to build our database and choose templates for each emotion. Hence emotion reactions will not be dispersed. We select the videos with body gestures in this AFEW and show them to the volunteers who participate in data collection. Then the volunteers will vote for the five most popular templates (we assume that these five templates are the most reasonable templates) as the final templates used in the database building.

As to the second requirement, the templates are used only as guidance instead of rules. Figure 1 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. Note that although templates are given to the subjects when we build VCED, VCED still has more freedom than other similar databases such as CK+

[Lucey *et al.*, 2010] and JAFFE [Lyons *et al.*, 1998]. Moreover, we create our database with different illumination conditions, head positions, and intensities of expressing emotions to capture the diversity in daily life better.

2.2 Database Collection and Brief Introduction

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 Tr antenna sends WiFi signals (can be replaced by a regular router), and the other three Rx antennas receive WiFi signals and extract CSI data using csitool [Halperin *et al.*, 2011]. As illustrated in Figure 1, 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 leap plate is placed between Tr-Rx3 to increase gesture sensitivity, which will be further explained in Section2.3.

In general, VCED has 7 emotions (i.e. *Angry, Disgust, Fear, Happy, Neutral, Sad* and *Surprise*), 35 templates (5 templates are selected for each emotion). Ten volunteers (7 male and 3 female, whose ages range from 23 to 25) repeat each template five times. Finally, 1750 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 VCED, we provide raw data (the suffix is .dat) and cropped data, which only contains emotion-related actions (the suffix is .mat). The total size of VCED is 43GB.

2.3 Leveraging Rician-K Factor Theory for WiFi CSI Data Collection

As shown in Figure.1, Tr-Rx3 antenna pair has low sensitivity to gestures due to their location. Based on the derivation of the Rican-K factor theory, we find that reducing the component of the direct signal can improve its performance, thus we add a lead plate between Tr-Rx3.

Rician-K factor

Rician-K factor is defined as the ratio of the power in the LOS path to the energy in the NLOS (Non-Line-Of-Sight) path. The baseband in-phase/quadrature-phase (I/Q) representation $x(t)$ of the received signal $x(t)$ (t is time index) can be represented as follows [Tepedelenlioglu *et al.*, 2003]:

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

where K is the Rician Factor, Ω denotes the total received power, θ_0 and ϕ_0 are the angle of arrival and phase of the LOS, respectively. f_D is the maximum Doppler frequency, and $h(t)$ is the diffuse component given by the sum of a large number of multipath components, constituting a complex Gaussian process.

The edge distribution probability density function $p(r)$ of the envelope of the received signal can be expressed by K
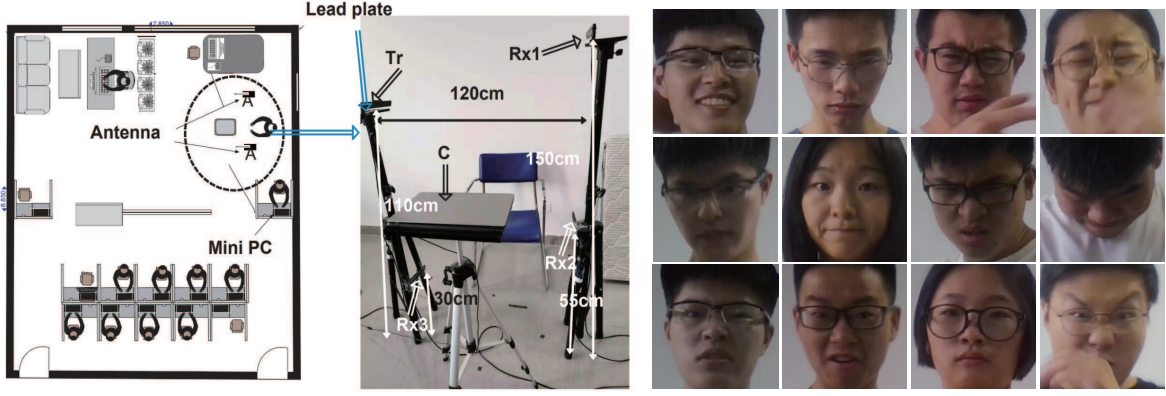
Figure 1: The left is a schematic diagram of our data acquisition system. Tr is the transmitting antenna, Rxs are receiving antennas. C is the computer, which is used for shooting video. We use a lead plate to block the Tr. The right part of the figure shows examples in our VCED database, the leftmost three (from top to bottom) are the happy, angry, and disgust expressions of the same volunteer, and the other nine pictures are the angry expressions of the other nine volunteers.

and Ω as:

$$p(r) = \frac{2(K+1)r}{\Omega} e^{(-K - \frac{(K+1)r^2}{\Omega})} \cdot I_0(2r\sqrt{\frac{K(K+1)}{\Omega}}) \tag{2}$$

where $I_0(.)$ is the 0th-order modified Bessel function of the first kind, r is the value of envelope. If there is no LOS component, i.e., when $K=0$, the above formula becomes the Rayleigh distribution.

How to use the Rician-K factor to assist WiFi data collection?

CSI at a particular moment can be expressed by CFR (channel frequency response), which could be further expressed as the superposition of dynamic path CFR and static path CFR as follows:

$$H(f,t) = H_s(f,t) + H_d(f,t) \tag{3}$$

The moments of the Rice distribution can expressed by [Stüber and Stèuber, 1996]:

$$\begin{aligned}\mu_n &:= E[R^n(t)] \\ &= (\delta^2)^{n/2}\Gamma(n/2+1)e^{(-K)}F_1(n/2+1;1;K),\end{aligned} \tag{4}$$

where $R(t)$ is the envelope of received signal, $\delta = \sqrt{\frac{\Omega}{K+1}}$, $F_1(.;.;.)$ is the confluent hypergeometric function, and $\Gamma(.)$ is the gamma function. Since K and δ are unknown, if we want to estimate K, we need at least two different moments of $R(t)$. Assume function depends only on K as follows:

$$f_{n,m}(K) := \frac{\mu_n^m}{\mu_m^n}, n \neq m \tag{5}$$

We can estimate K by inverting $f_{n,m}(K)$, and the estimation of K based on the mth and nth components is as follows:

$$\widehat{K}_{n,m} = f_{-1}^{n,m}(\frac{\widehat{u}_n^m}{\widehat{u}_m^n}) \tag{6}$$

$$\widehat{u}_k = \frac{1}{N}\sum_{l=0}^{N-1} R^k(lTs) \tag{7}$$

Here N is the number of available samples, and Ts is the sampling period. Generally, we use the second and fourth components to estimate K [Greenstein *et al.*, 1999]:

$$\widehat{K}_{2,4} = \frac{-2\widehat{\mu}_2^2 + \widehat{\mu}_4 - \widehat{\mu}_2\sqrt{2\widehat{\mu}_2^2 - \widehat{\mu}_4}}{\widehat{\mu}_2^2 - \widehat{\mu}_4} \tag{8}$$

As described in Eq. (3), the received signal can be divided into two parts: static path signal and dynamic path signal. The receiving signal has a time-varying amplitude in complex plane [Wang *et al.*, 2016]:

$$|H(f,\theta)|^2 = |H_s(f)|^2 + |H_d(f)|^2 + 2|H_s(f)||H_d(f)|cos\theta, \tag{9}$$

where θ is the phase difference between the static vector and the dynamic vector. The term $2|H_s(f)||H_d(f)|cos\theta$ will cause the amplitude fluctuation of the CSI. When the amplitude of motion and θ are both constant, $|H_s(f)|$ and $|H_d(f)|$ will be the factors affecting the fluctuation range.

Since antenna does not move when we collect data, i.e., $f_D = 0$, we could simplify Eq. (1) as:

$$x(t) = \sqrt{\frac{K\Omega}{K+1}}e^{\phi_0} + \sqrt{\frac{\Omega}{K+1}}h(t) \tag{10}$$

When 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 Eq. (10) and ignoring the transmitted power, we define H_s and H_d as follows:

$$H_s = \sqrt{\frac{K}{K+1}} + \sqrt{\frac{1}{K+1}} \cdot \rho \tag{11}$$

$$H_d = \sqrt{\frac{1}{K+1}} \cdot (1-\rho) \tag{12}$$

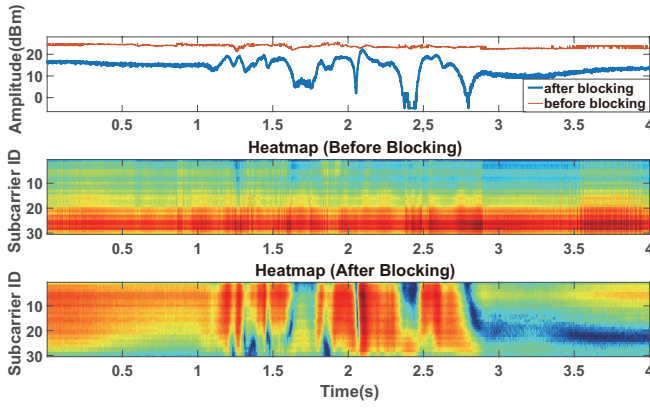where ρ is the proportion of static paths in the NLOS component. Combine with equation (9), we can obtain the following

Figure 2: The amplitude, energy of the user's same motion (turn over his body) before/after blocking the transmitting antenna.

equation:

$$|H|^2 = |H_s|^2 + |H_d|^2 + 2|H_s||H_d|cos\theta$$
$$= \frac{K + \rho^2 + 2\sqrt{K}\rho cos\alpha}{K+1} + \frac{(1-\rho)^2}{K+1} +$$
$$\frac{2(1-\rho)\sqrt{K + \rho^2 + 2\sqrt{K}\rho cos\alpha}}{K+1}cos\theta$$

(13)

where α is the phase difference of the LOS component to the NLOS component in the static vector. It is obvious that K and ρ are the factors affecting the range of waveform fluctuation. We assume that all NLOS components belong to the dynamic vector, i.e., $\rho = 0$ and assume that $\alpha = \pi/2$. Then the following equation can be obtained:

$$f(K) = |H_s||H_d| = \frac{\sqrt{K}}{K+1}$$

(14)

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

(15)

When $K > 1$, $f(K)$ decreases as K increases. When there are not many obstacles in between the transmitting antenna and its receiver, most of the NLOS paths belong to dynamic vector, and this can ensure that ρ is relatively small. In other words, adding an appropriate obstacle to lower the K (Note that don't block too many transmitting signals) can make CSI more sensitive to gestures. The effect of blocking the LOS signal is shown in Figure 2, where we can observe that blocking the LOS signal reduces the average amplitude of the CSI, however, enhances the sensitivity of the CSI to the gesture.

3 System Design

3.1 System Overview

After having the database, we introduce the Vision-CSI bi-modal emotion recognition system, as illustrated in Figure 3. For CSI data, after filtering the high-frequency noise, SVM is used for classifying the emotions by its associated body gestures. For video data, we use three kinds of Densenet

(Densenet121, Densenet169, and Densenet 201, where the integer number represents the number of the layers) to extract the static features of the video frames. We use the VGG-LSTM network structure to extract the temporal characteristics of the video. By exploring both the temporal and spatial features, the facial expression can be better captured. Finally SVM is used to fuse the final results.

3.2 Data Pre-processing

The captured CSI data contains a large amount of Gaussian white noise due to environmental and electromagnetic interference. To extract useful information while removing irrelevant information, we use a simple but effective Butterworth low-pass filter to filter the CSI data first.

For facial expression recognition, the first step required is to obtain human faces. We extract human faces from all video frames using the Multi-task Cascaded Convolutional Networks (MTCNN) [Zhang *et al.*, 2016] to save the computation load. MTCNN can adjust the head position and detect more faces than the dlib detector [King, 2009]. The faces are aligned at a fixed direction before training our networks, as shown in Figure 3. The size of each image is 256x256.

3.3 SVM for CSI Classification

We extract the *length, variance, maximum, minimum*, and *mean values* of all 90 subcarriers of each CSI data as classification features to train the SVM classifier. Due to the characteristics of different periods of a gesture that may vary, we divide each CSI data into eight segments for each subcarrier evenly. Then the variance is calculated for each part. Thus we can obtain 1080 features from each CSI data file. After features extracted, we use SVM for the CSI-based emotion recognition, details of which are omitted due to the page limitation.

3.4 Neural Networks for Vision-based Recognition

We implement vision-based emotion recognition by two depth-based approaches, i.e., Densenet-SVM and VGG-LSTM, in which Densent-SVM for static features and VGG-LSTM for temporal features. In particular, the Densent network extracts the characteristics of each video frame into SVM for classification, this process considers static features of all frames. VGG-LSTM use VGG to obtain the characteristics of each frame, and then sends the features of each frame to the LSTM network chronological for training, this method considers the temporal characteristics of video clips.

We use the FER2013 database [Goodfellow *et al.*, 2013] [1] to pre-train the four selected CNN (three kinds of Densenet and VGG16) models. After pretraining the four selected neural networks, we use our database to fine-tune the pre-trained network structures. Then we extract features by the fine-tuned models from the last layer. The feature dimension for

---

[1] Note that the FER2013 database is introduced during the ICML 2013 Challenges in Representation Learning. It is a large-scale and unconstrained database collected automatically by the Google image search API, FER2013 is well-known for its quality in gray-scale and its adequate number of facial expressions, which is also the reason we use it to pre-train the CNN models.
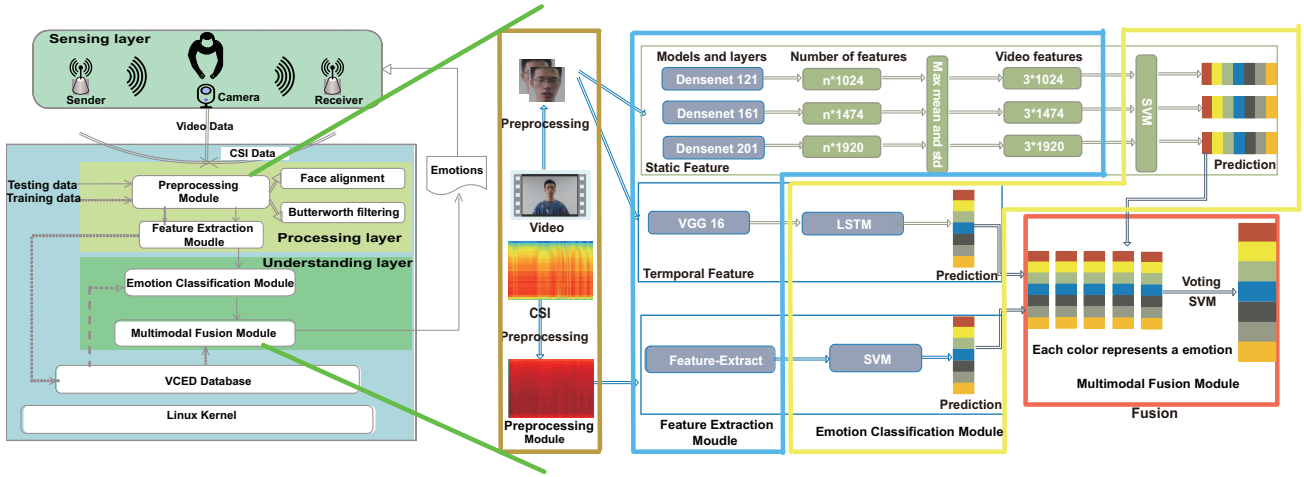
Figure 3: Overview of our method, The left is the overall flowchart of our emotion recognition system, and the right is the specific data processing method. There are three main parts in our hybrid system, Densent-SVM for static image features, VGG-LSTM for temporal videos, and SVM for CSI data. The results from these classifiers will be fused using SVM.

each video is relevant to the number of detected faces and the layer dimension.

For Densenet models, we first normalize the features by dividing them with the maximum value. Since different videos have different lengths, the extracted feature dimensions also various. Then we calculate the *mean, max* and *standard deviation* for features extracted from each video. In this way, the characteristics of each video are all three-dimensional. After completing the dimensionality reduction, we use SVM to classify the characteristics of different videos.

For the VGG model, we first normalize the features by dividing each value with the maximum value. Then, to balance the temporal characteristics of the videos, we complement the data with 0 to make them equal in length and send them to the LSTM chronological for processing.

3.5 Fusion of the Results

After obtaining the recognition results from all five classifiers, we combine them using SVM for the final result. Due to space limitations, we omit the details here.

4 Evaluations

4.1 Evaluation Setup

We build a prototype system to evaluate the performance, as shown in Figure 1. During the experimental evaluation process, we select three volunteers' data (two male and one female,named zx ysl and hmm in our dataset) as the test data, and the remaining seven volunteers' data as the training data. Thus our training and testing process are individually independent.

4.2 Evaluation Results

When with CSI data only, the overall accuracy of the CSI-SVM method is 66.48%, and the confusion matrix is shown in Figure 4. When with video data only, the overall accuracy achieved by Densent121, Densent169 and Densent201

are 64.57%, 64% and 61.14% respectively. The confusion matrices are shown in Figure 5, 7 and 8. For VGG-LSTM, it mainly considers the temporal characteristics of video clips, and its overall accuracy is 66.67%. Its confusion matrix is shown as Figure 9.

We initially consider the use of weighted voting to fuse the classification results of the five classification models. The overall accuracy reaches 79.05%, where the weights are 0.4,0.3,0.1,0.1,0.1 for CSI-SVM, VGG-LSTM, Densent121, Densent169 and Densent201 respectively. The highest classification accuracy is *disgust*, which is 89.33%. The correct classification rate of *happy, neutral, sad* and *surprise* are all over 84%. The lowest classification rate is *angry*, which only 54.67%, The probability of mis-classifying *angry* to *sad* is 30.67%, which is similar to video-based schemes. The misclassification rate of other emotions is lower than CSI-based scheme.

We initially consider the use of weighted voting to fuse the classification results of the five classification models. The overall accuracy reaches 79.05%, where the weights are 0.4,0.3,0.1,0.1,0.1 for CSI-SVM, VGG-LSTM, Densent121, Densent169 and Densent201 respectively. The highest classification accuracy is *disgust*, which is 89.33%. The correct classification rate of *happy, neutral, sad* and *surprise* are all over 84%. The lowest classification rate is *angry*, which only 54.67%, The probability of mis-classifying *angry* to *sad* is 30.67%, which is similar to video-based schemes. The misclassification rate of other emotions is lower than CSI-based scheme. When using SVM to fusion all predictor's results, the final accuracy reaches 83.24%, which is much higher than the CSI-only based scheme and vision-only based schemes. This indicates the effectiveness of our bimodal solution. The accuracy confusion matrix is shown in Figure 6. The highest classification accuracy is *surprise*, which is 92%, close to the CSI-SVM method, while the lowest classification accuracy is *sad*, which is 73.33%. Notice:However, we think 10-foldSVM is unfair, because there are only 3 volunteer's data
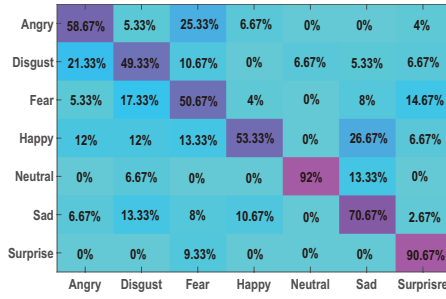
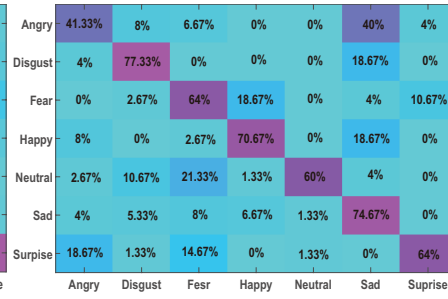| Figure 4 (CSI-SVM)	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprisre
Angry	58.67%	5.33%	25.33%	6.67%	0%	0%	4%
Disgust	21.33%	49.33%	10.67%	0%	6.67%	5.33%	6.67%
Fear	5.33%	17.33%	50.67%	4%	0%	8%	14.67%
Happy	12%	12%	13.33%	53.33%	0%	26.67%	6.67%
Neutral	0%	6.67%	0%	0%	92%	13.33%	0%
Sad	6.67%	13.33%	8%	10.67%	0%	70.67%	2.67%
Surprise	0%	0%	9.33%	0%	0%	0%	90.67%

Figure 4: Confusion matrix of CSI-SVM method.

Figure 5 (Densent121)	Angry	Disgust	Fesr	Happy	Neutral	Sad	Suprise
Angry	41.33%	8%	6.67%	0%	0%	40%	4%
Disgust	4%	77.33%	0%	0%	0%	18.67%	0%
Fear	0%	2.67%	64%	18.67%	0%	4%	10.67%
Happy	8%	0%	2.67%	70.67%	0%	18.67%	0%
Neutral	2.67%	10.67%	21.33%	1.33%	60%	4%	0%
Sad	4%	5.33%	8%	6.67%	1.33%	74.67%	0%
Surprise	18.67%	1.33%	14.67%	0%	1.33%	0%	64%

Figure 5: Confusion matrix of Densent121 method.

Figure 6 (SVM-based fusion)	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	81.33%	6.67%	2.67%	0%	0%	6.67%	2.67%
Disgust	4%	86.67%	0%	2.67%	2.67%	4%	0%
Fear	4%	0%	88%	5.33%	0%	2.67%	0%
Happy	6.67%	2.67%	2.67%	78.67%	6.67%	1.33%	1.33%
Neutral	0%	4%	6.67%	4%	82.67%	2.67%	0%
Sad	9.33%	10.67%	2.67%	2.67%	1.33%	73.33%	0%
Surprise	4%	1.33%	1.33%	0%	0%	1.33%	92%
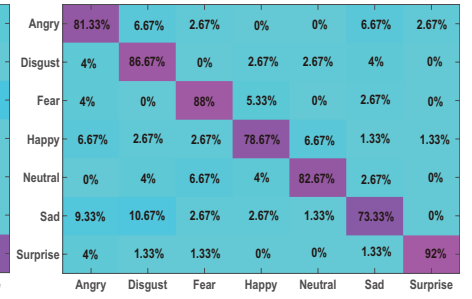
Figure 6: Confusion matrix of SVM-based fusion method.

Figure 7 (Densent169)	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	44%	17.33%	8%	0%	0%	30.67%	0%
Disgust	9.33%	70.67%	6.67%	0%	0%	13.33%	0%
Fear	5.33%	4%	68%	16%	0%	6.67%	0%
Happy	4%	0%	8%	70.67%	0%	17.33%	0%
Neutral	2.67%	10.67%	20%	0%	64%	2.67%	0%
Sad	6.67%	21.33%	8%	0%	4%	60%	0%
Surprise	12%	1.33%	8%	0%	0%	8%	70.67%
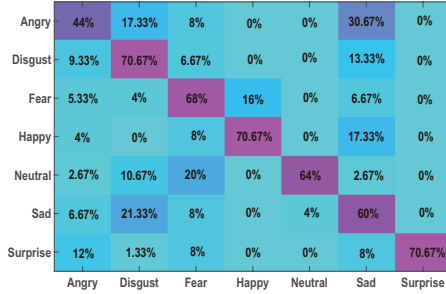
Figure 7: Confusion matrix of Densent169 method.

Figure 8 (Densent201)	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	34.67%	16%	6.67%	0%	0%	42.67%	0%
Disgust	1.33%	92%	2.67%	1.33%	0%	26.67%	0%
Fear	2.67%	8%	42.67%	21.33%	0%	9.33%	16%
Happy	5.33%	0%	0%	84%	0%	10.67%	0%
Neutral	1.33%	14.67%	22.67%	4%	57.33%	0%	0%
Sad	1.33%	26.67%	2.67%	13.33%	6.67%	49.33%	0%
Surprise	2.67%	21.33%	4%	0%	4%	0%	68%
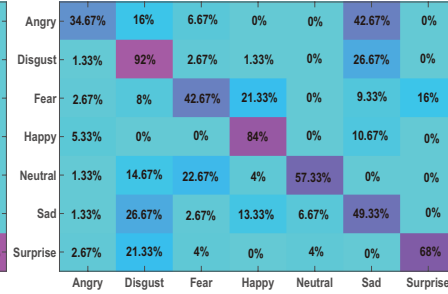
Figure 8: Confusion matrix of Densent201 method.

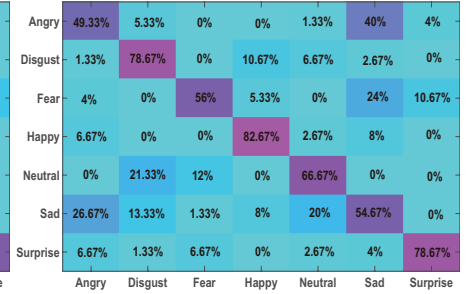| Figure 9 (VGG-LSTM)	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	49.33%	5.33%	0%	0%	1.33%	40%	4%
Disgust	1.33%	78.67%	0%	10.67%	6.67%	2.67%	0%
Fear	4%	0%	56%	5.33%	0%	24%	10.67%
Happy	6.67%	0%	0%	82.67%	2.67%	8%	0%
Neutral	0%	21.33%	12%	0%	66.67%	0%	0%
Sad	26.67%	13.33%	1.33%	8%	20%	54.67%	0%
Surprise	6.67%	1.33%	6.67%	0%	2.67%	4%	78.67%

Figure 9: Confusion matrix of VGG-LSTM method.

for test, thus we only use weighted voting as finally result.
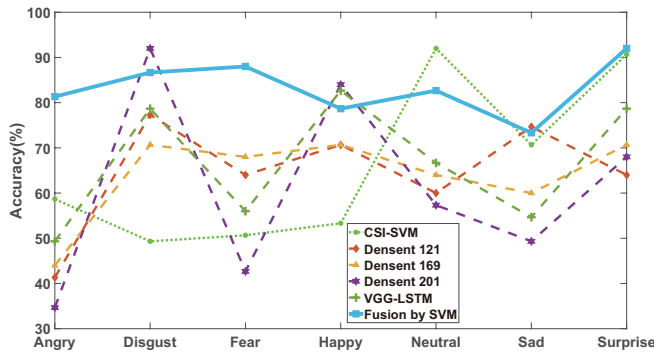

Figure 10: Classify accuracy of different methods.

4.3 Evaluation Analysis

Figure 4 shows the confusion matrix of the CSI-SVM method. Among them, the highest classification accuracies are for classifying *surprise* and *neutral*, both over 90%. This is because the duration of *surprise* is shorter than the other kinds of emotions, and the associated gestures' amplitude is larger than the others, and *Neutral* is usually calmer than other kinds of emotions. The accuracy of *disgust* is the lowest, only 49.33%. The highest misclassification rate is to misrecognize *happy* to *sad*, which reaches 26.67%.

Figure 5 shows the confusion matrix of the Densent121 method. The *disgust* has the highest classification accuracy

as 77.33%. The lowest classification accuracy is the probability of recognizing *angry* as 41.33%. The highest misclassification rate is misclassifying *angry* to *sad*, which reaches 40%. This is because there are significant differences between the intensities and manners of facial expressions of different people when expressing the same emotion, which leads to low classification accuracy. In some scenarios, the hands and arms cover the face (such as when wiping tears, covering mouth in surprise), this also poses challenges for vision-based emotion recognition.

As shown in Figure 5, 7 and 8, these three kinds of Densent networks have similar results in general. They differ in some subtle ways; for example, Densent201 has a classification accuracy rate of 92% for *disgust*. Other kinds of Densent networks do not have such a high classification rate for *disgust*. This because different Densent with various depth, thus the features they eventually extract will be different.

Comparing the results obtained from the CSI and video, we can notice that CSI performs better for *neutral* and *surprise* (all over 90%), but vision-based scheme performs better for *disgust* and *happy*, these two methods can complement with each other to recognizing these four emotions better. We can also find that they also make a big difference in terms of misclassification rates. For example, all neural networks for videos have the highest misclassification rate of misclassifying *angry* to *sad*. But for CSI, the rate of misclassifying *angry* to *sad* is 0%.

We show the fused emotion recognition result and all single-approach based recognition results obtained by each classifier in Figure 10, the experimental results prove that both gestures and facial expressions are beneficial for emo-

tion recognition, their combination can also complement each other and bring significant benefits. The vision-based method complements the shortcomings of the gesture-based approach in identifying *disgust*, while the gesture-based scheme makes up for the defects of the vision-based approach in identifying *neutral* emotions. Even for *angry* emotion, where the accuracy of these two methods all terrible, the performance of the fusion scheme is still excellent; this is because the two ways have different misclassifications of *angry*, they can still be complementary in performance.

5 Conclusion

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 theory and combined CNN and RNN to mine the high-level features of bi-modal data to process the large-volume, heterogeneous and non-synchronous data contributed by the two-modalities, and perform a score-level fusion for fine-grained recognition. We built a first-of-its-kind Vision-CSI Emotion Database (VCED) and a prototype system to evaluate the proposed method. The Empirical results show the superiority of the bi-modality by achieving 83.24% recognition accuracy for seven emotions, as compared with 66.48% and 66.67% recognition accuracy by gesture-only based solution and facial-only based solution, respectively.

References

[Albanie *et al.*, 2018] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 292–301. ACM, 2018.

[Anagnostopoulos *et al.*, 2015] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

[Chai *et al.*, 2016] Rifai Chai, Ganesh R Naik, Tuan Nghia Nguyen, Sai Ho Ling, Yvonne Tran, Ashley Craig, and Hung T Nguyen. Driver fatigue classification with independent component by entropy rate bound minimization analysis in an eeg-based system. *IEEE journal of biomedical and health informatics*, 21(3):715–724, 2016.

[Cowie *et al.*, 2001] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[Dhall *et al.*, 2012] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34–41, 2012.

[Goodfellow *et al.*, 2013] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[Greenstein *et al.*, 1999] Larry J Greenstein, David G Michelson, and Vinko Erceg. Moment-method estimation of the ricean k-factor. *IEEE Communications Letters*, 3(6):175–176, 1999.

[Gross and Muñoz, 1995] James J Gross and Ricardo F Muñoz. Emotion regulation and mental health. *Clinical psychology: Science and practice*, 2(2):151–164, 1995.

[Gunes and Piccardi, 2006] Hatice Gunes and Massimo Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1148–1153. IEEE, 2006.

[Halperin *et al.*, 2011] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Computer Communication Review*, 41(1):53–53, 2011.

[Huang *et al.*, 2017] Yongrui Huang, Jianhao Yang, Pengkai Liao, and Jiahui Pan. Fusion of facial expressions and eeg for multimodal emotion recognition. *Computational intelligence and neuroscience*, 2017, 2017.

[King, 2009] Davis E King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[Li and Deng, 2018] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*, 2018.

[Lucey *et al.*, 2010] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[Lyons *et al.*, 1998] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[Minsky, 1987] Marvin Minsky. The society of mind. In *The Personalist Forum*, volume 3, pages 19–32. JSTOR, 1987.

[Noroozi *et al.*, 2019] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, 10(1):60–75, 2019.

[Stüber and Stèuber, 1996] Gordon L Stüber and Gordon L Stèuber. *Principles of mobile communication*, volume 2. Springer, 1996.

[Tepedelenlioglu *et al.*, 2003] Cihan Tepedelenlioglu, Ali Abdi, and Georgios B Giannakis. The ricean k factor: estimation and performance analysis. *IEEE Transactions on Wireless Communications*, 2(4):799–810, 2003.

[Wang *et al.*, 2016] Hao Wang, Daqing Zhang, Junyi Ma, Yasha Wang, Yuxiang Wang, Dan Wu, Tao Gu, and Bing 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*, pages 25–36. ACM, 2016.

[Yan and Zhang, 2009] Yan Yan and Yu-Jin Zhang. State-of-the-art on video-based face recognition. In *Encyclopedia of Artificial Intelligence*, pages 1455–1461. IGI Global, 2009.

[Yen *et al.*, 2018] Ju-Yu Yen, Yi-Chun Yeh, Peng-Wei Wang, Tai-Ling Liu, Yun-Yu Chen, and Chih-Hung Ko. Emotional regulation in young adults with internet gaming disorder. *International journal of environmental research and public health*, 15(1):30, 2018.

[Zhang *et al.*, 2016] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.