WiFi and Vision enabled Multimodal Emotion Recognition

Yuanwei Hou
School of Software and Microelectronics
Peking University
Beijing, China
hyw@pku.edu.cn

Xiang Zhang
School of Computer and Information
Hefei University of Technology
Hefei, China
zhangxiang@mail.hfut.edu.cn

Yu Gu
School of Computer and Information
Hefei University of Technology
Hefei, China
yugu.bruce@ieee.org

Weiping Li
School of Software and Microelectronics
Peking University
Beijing, China
wpli@ss.pku.edu.cn

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9838315

***Abstract*—Emotion recognition plays a vital role in current research on human-computer interaction, and human emotion expressions are multi-modal. In this paper, we propose a passive multi-modal emotion recognition system based on facial expression and gesture. To achieve the system design, two major challenges must be addressed, namely, how to capture facial expression and gesture without disturbing the subject, and how to use the correlation between the two modalities to better recognize emotions. For the former, we use WiFi and vision for the passive gesture and facial expression capture, respectively. For the latter, we design a Multi-Source Learning method inspired by Multi-Task Learning to efficiently exploit the correlation between modalities for better emotion recognition. Finally, to evaluate the effectiveness of our system, we use low-cost vision and WiFi devices to prototype the system and build a WiFi-Vision emotion dataset for related research, and we verify the effectiveness of our system in emotion recognition and the superiority of multi-modality over single-modality through conduct extensive experiments.**

***Index Terms*—emotion recognition, channel state information, vision, dataset, multi-source learning**

I. Introduction

Emotion recognition plays a vital role in current research on human-computer interaction, and human emotion expressions are multi-modal in nature, [1] choosing a more expressive modality can make it easier to realize emotion recognition. Facial expressions are undoubtedly the most intuitive modalities that can reflect human emotions, and facial expressions constitute important nonverbal emotional cues [2]. Besides, body gesture is also a modality that can easily express emotions [3], people can use part of their body to express emotions, especially the relatively strong emotions. In this paper, we focus on exploring these two modalities to achieve fine-grained emotion recognition. And in the process of system implementation, there are two challenges to be solved.

The first challenge is how to capture emotional expression without disturbing the subject? In current researches, facial expressions are usually captured by cameras, while body gestures are mainly captured by wearable sensors. However, wearable sensors that are generally in contact or even intrusive state may interfere with objects, thereby contaminating emotional cues. Therefore, we tend to use a newly developed deviceless alternative sensor for gesture capture, namely WiFi [4], [5]. Due to the multipath effect on the human body [6], COTS WiFi has been shown to be able to capture human movement. However, the sensitivity of WiFi channel status information (CSI) to human motion plays a vital role in gesture recognition. In this paper, we propose a CSI enhancement model that leveraging Rician fading to enhance its sensing granularity.

The second challenge is how to effectively leverage the correlation between a large amount of data from the facial and gesture modalities for better emotion recognition? Current researches on multimodal fusion mainly rely on early-fusion or late-fusion based methods [7]. However, early-fusion is easily affected by data loss, and late-fusion needs to train multiple decoders [8]. To fill in this gap, we proposed a Multi-Source Learning (MSL) framework inspired by Multi-Task Learning to take advantage of cross-correlation between modalities. MSL uses an encoder to extract useful features for each modality, and then feeds them to a shared decoder. Through parameter sharing, different modalities can exchange knowledge to extract cross-correlated features. Finally, the cross-correlated features help each modality to output its recognition result, and the final decision is made via voting over all outputs.

To evaluate our idea, we prototyped it using low-cost off-the-shelf vision and WiFi equipment, and built the first WiFi-Vision emotion dataset. Extensive experiments have confirmed the effectiveness of the bi-modality-based method by reaching 83.81% recognition accuracy, as compared with 68.38% and 70.29% recognition accuracy by gesture-only and facial-only solutions, respectively. We also verify that MSL based method surpasses the early-fusion method (80.76%) with robustness against data loss, and the late-fusion method (83.43%) with less computing consumption, respectively.

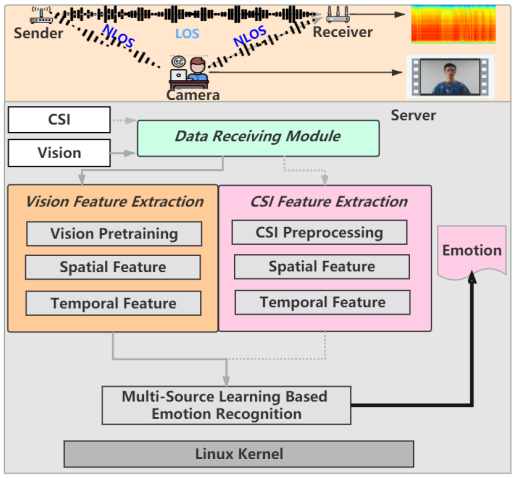

Fig. 1: System Overview

In summary, our contributions are summarized as follows:

- We are the first one to leverage WiFi and vision for contactless emotion recognition. And we build the first open WiFi-Vision emotion dataset for public research.
- We propose a CSI enhancement model based on Rician fading theory for enabling sub-wavelength gesture capture.
- We design a Multi-Source Learning framework to exploit the cross-correlation between modalities for better emotion recognition. Its efficiency has been verified via our dataset.

II. System Design

A. Overview

The overview of our emotion recognition system is shown in Fig. 1, the system consists of three data processing parts, i.e., Vision-based facial expression feature extraction, WiFi-based gesture feature extraction, and MSL-based emotion recognition.

In the first part, we leverage the Multi-task Cascaded Convolutional Networks (MTCNN) [9] to detect face frames from videos and crop the frame to a suitable size. Then we use two kinds of Densenet to extract the spatial features from these cropped frames. Finally, we use the VGG-LSTM network to extract the temporal features.

In the second part, we first leverage the Ricean fading based model to enhance CSI for better capturing the fine-grained body gestures. Then, we visualize CSI data into CSI maps for processing using convolutional neural networks. Finally, we use Densenet121, Densenet169, and VGG-LSTM to extract static and temporal features from CSI maps, respectively.

In the last part, we propose a MSL framework inspired by multi-task learning to perform bi-modal emotion recognition. In particular, MSL combines visual and CSI features to explore the correlation between gestures and facial expressions that convey the same emotion. In the next part, we describe our system in detail.

B. Vision-based Facial Expression Feature Extraction

In this part, we introduce how to extract emotional-related features from captured videos. We first explain how to crop the face correctly from video frames. Then, drawing on previous research, in order to help the network converge faster, we use the FER2013 face dataset to pre-train the networks we used in this paper. Finally, we leverage the pre-trained networks to extract the emotional-related features of these videos for bi-modal emotion recognition.

1) Face Detection and Alignment: Stable face tracking is a vital step for vision-based emotion recognition. In this work, we crop the faces using the Multi-task Cascaded Convolutional Networks (MTCNN) [9], which not only can detect more faces than the dlib detector (a popular image tool that can be used for face detection), but also can adjust the head position. The faces are cropped and aligned at a fixed direction before training our nets, and the size of each image is 256×256.

2) Vision Pre-training: Since there are many well-structured and well-labeled image data sets, pre-training is often used in computer vision tasks. On the one hand, pre-training is like the initial parameter setting procedure, which is critical in the subsequent training process. On the other hand, it alleviates the overfitting problem of small data sets. In this paper, We use the FER2013 dataset [10] (The FER2013 dataset is introduced during the ICML 2013 Challenges in Representation Learning, it is a large-scale facial expression dataset.) to pre-train the three selected CNN (two kinds of Densenet and VGG16) models. After pretraining the three selected neural networks, we use our dataset to fine-tune the pre-trained network structures.

3) Vision Feature Extraction: According to previous researches [11], [12], we implement vision-based feature extraction by two depth-based approaches, i.e., Densenet and VGG-LSTM, in which Densenet for static features and VGG-LSTM for temporal features. In particular, the Densenet network extracts the characteristics of each video frame, this process considers static features of all frames. VGG-LSTM uses 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.

For Densenet models, after extracting features for each frame in the video, we first normalize the features by dividing them by the maximum value. Since different videos have different lengths, the extracted feature dimensions also various, thus we calculate the *mean, max* and *standard deviation* for features extracted from each video. In this way, the features of each video are all three-dimensional, the features extracted by Densent are input into the Multi-Source-Learning based fusion scheme to obtain the final result.

C. WiFi-based Gesture Extraction

1) CSI Collection: Channel State Information (CSI) is a kind of fine-grained physical layer (PHY) information, it

describes the signal's attenuation factors on each transmission path, such as scattering, multipath fading or shadowing fading, power decay of distance, and other information. In the frequency domain, the channel model can be expressed as:

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

Where Y is the received WiFi signal, X is the transmitted WiFi signal, and N is the noise caused by the environment and devices. H is the channel state information matrix that can reflect the attenuation of WiFi signal propagation. Human gestures have a special effect on the propagation of WiFi signals, and this effect can be obtained from the H matrix for wireless sensing tasks.

The frequency bands of current commercial WiFi devices are divided into multiple subcarriers, and for each subcarrier, the channel frequency response (CFR) can be expressed as:

$$H_i = |H_i| e^{j \angle H_i} \tag{2}$$

Where i is the subcarrier index. $|H_i|$ and $\angle H_i$ are the amplitude and phase of ith subcarrier, respectively. Thus the dimension of CFR matrix H is $A_t \times A_r \times A_s \times Ti$, A_t,A_r and A_s are the number of transmitting antennas, receiving antennas and subcarriers, respectively, where $A_t = 1$, $A_r = 3$ and $A_s = 30$ in our system.

2) CSI Pre-processing: The sensitivity of CSI to gesture plays a critical role in capturing the emotional expression, especially the subtle ones like an imperceptible nod. In this section, we present a CSI enhancement model based on Rician fading to highlight the gesture-induced information by suppressing the gesture-unrelated information on the channel response

In wireless communications, the Rician K factor is the ratio of the powers of the LOS component to the total received power, and the baseband $x(t)$ can be modeled as [13]:

$$x(t) = \sqrt{\frac{K\Omega}{K+1}} e^{j(2\Pi f_D cos(\theta_0)t) + \phi_0} + \sqrt{\frac{\Omega}{K+1}} h(t) \tag{3}$$

where K are Ω are the Ricean K factor and wholly received power, respectively. θ_0 is the angle of arrival, and ϕ_0 is the phase of the LOS. f_D and $h(t)$ are the maximum Doppler frequency and diffuse components, respectively.

In our system, due to the transmitting and receiving antennas are fixed, i.e., $f_D = 0$, we can simplify Equation 3 to:

$$x(t) = \sqrt{\frac{K\Omega}{K+1}} e^{\phi_0} + \sqrt{\frac{\Omega}{K+1}} h(t) \tag{4}$$

Since we only measure the effect of the factor K on the received signal, it can be written as follows,

$$x(t) = \sqrt{\frac{K}{K+1}} + \sqrt{\frac{1}{K+1}} \tag{5}$$

The LOS components and the part of NLOS components that are not affected by gestures are static. Other components are dynamic. Therefore, we can define H_s and H_d as follows if omitting the transmit power:

$$H_s = \sqrt{\frac{K}{K+1}} + \sqrt{\frac{1}{K+1}} \cdot \rho \tag{6}$$

$$H_d = \sqrt{\frac{1}{K+1}} \cdot (1-\rho) \tag{7}$$

Where ρ is the proportion of static paths in the NLOS component. The received signal consists of both static and dynamic paths. Thus the receiving signal has a time-varying amplitude [14]:

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

Combining Equation (6), we can get the following equation:

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

Where α is the phase difference of the LOS component to the NLOS component in the static paths. It can be seen that the factors affecting the range of waveform fluctuation caused by the motions are K and ρ.

We define $f(K) = |H_s| \cdot |H_d|$ to indicate the system sensitivity to the gesture. Assuming that all NLoS components belong to the dynamic paths, i.e., ρ = 0 and $\alpha = \frac{pi}{2}$, we obtain the following equations:

$$f(K) = |H_s||H_d| = \frac{\sqrt{K}}{K+1} \tag{9}$$

$$f'(K) = \frac{1-K}{2\sqrt{K}(1+K)^2} \tag{10}$$

When $K > 1$, $f(K)$ increases as K decreases. In summary, we only need to block a part of the LOS signal to achieve the purpose of CSI enhancement. In our system, we put a small piece of lead plate on the transmitting antenna to block the LOS signal

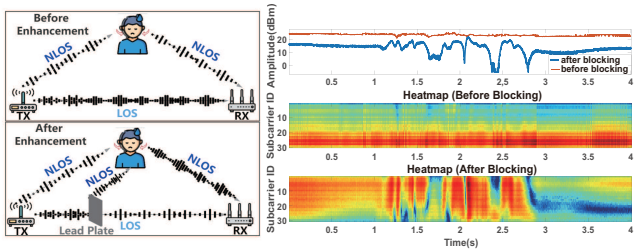

Fig. 2: The performance of our CSI enhancement method fading

As shown in Fig. 2. A person acts a gentle nod causing the NLoS signal propagation paths, and it is hard to be observed in the original CSI. But after enhancement by applying our Rician-K based CSI enhancement method, we can clearly observe such movements from recorded CSI.

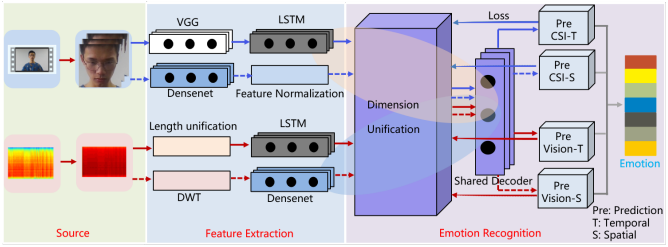

Fig. 3: MSL Based Emotion Recognition

3) CSI Feature Extraction: After CSI enhancement, the CSI data should be processed to remove the background noise. The captured CSI data contains a large amount of Gaussian white noise due to interference from environmental noises and electromagnetic noises. In order to extract useful information while removing irrelevant information, we filter the CSI data before the recognition. In this system, we use a simple but effective Butterworth low-pass filter with a passband of 10 and a stopband of 30.

In order to use the existing neural networks (DenseNet, VGG), in this paper, we first convert the high-dimensional CSI tensor into a two-dimensional heat map. After obtaining the CSI map, we use two kinds of Densenet (Densenet121 and 169, both pre-trained with FER2013) to extract features of different depths as CSI spatial features [15], while a 512-dimensional LSTM network is employed to extract temporal features. The extracted CSI features will be sent to the multi-modal emotion recognition part together with the visual features for emotion recognition.

D. MSL-based Emotion Recognition

After obtaining the spatial and temporal features of gestures and facial expressions, we need to effectively use the correlation between these features to achieve bimodal emotion recognition. In this part, we introduce our MSL-based emotion recognition method inspired by Multi-Task Learning (MTL).

The purpose of multi-task learning is to use only one model to solve multiple tasks, and can improve performance through the information exchange between tasks. Each task of MTL has a corresponding decoder, but shares an encoder, the information interaction of different tasks is realized through the shared encoder. Inspired by MTL [16], we propose MSL which uses only one decoder for all encoders, and each input feature corresponds to an encoder. Therefore, MSL realizes the information interaction of multiple input features through the shared decoder. Different from MTL, it shares a low-level feature encoder that is effective for multiple tasks, while MTL shares a task-specific high-level decoder.

In the training process of MSL, the decoder is shared by all inputs, and it is jointly trained by all inputs. Different inputs will be assigned weights according to their importance. The greater the weight, the greater the impact on the training process of the decoder. This training method encourages the decoder to be trained under the supervision of real labels and multiple inputs, so as to converge towards the goal of benefiting multiple inputs. Lastly, MSL makes the final decision based on voting.

Current multi-modal recognition researches mainly leverage early-fusion or late-fusion based methods. The former first aggregates the output from multiple encoders, then use the aggregate feature to obtain the final result through a decoder. While late-fusion only fuses the output results of different modalities in the result layer based on voting or retraining. Early-fusion needs to aggregate the features of different inputs, thus it cannot work in the absence of a kind of input, while MSL can still work in the absence of several inputs. Compared to late-fusion, since different inputs share a decoder, the network parameters and efficiency have been improved.

Our MSL-based emotion recognition system is shown in Fig 3, for each modality, there are three encoders (Densenet 121 and Densenet 169 for spatial features, VGG-LSTM for temporal features.) for extracting temporal and spatial features. The output features from all six encoders will jointly train the shared decoder. After obtaining the emotion recognition results of each encoder, we obtain the final result by weighted voting. And in this paper, the weight of the six encoders are set to $2:4:3:3:4:4$, and the weight of the final weighted voting are set to $5:5:2:1:1:2$, respectively.

III. Dataset Construction and Performance Evaluations

In this section, we first introduce the WiFi-Vision dataset, and then use this dataset to evaluate our system.

A. WiFi-Vision Dataset Construction

To the best of our knowledge, there exists no WiFi-Vision bi-modal dataset. To this end, we build the first WiFi-Vision emotion dataset to evaluate our system. We choose *Acted Facial Expressions in the Wild* (AFEW) [17] dataset to guide how people behave with different emotions. The body gesture in our WiFi-Vision dataset is set with reference to the AFEW dataset. Figure 4 shows a snapshot of our bi-modal system as well as some examples of our dataset, which will be released to the public soon.

During the dataset collection, we use a laptop to record the video data, and two use Mini PC (Intel 5300 NIC) with four antennas to obtain the CSI as shown in Figure 4. Our dataset contains 7 kinds of emotions (i.e. *Angry, Disgust, Fear, Happy, Neutral, Sad* and *Surprise*), 35 kinds of gesture and facial expression templates (5 templates are selected for each emotion). 10 volunteers (7 male and 3 female, whose ages range from 23 to 25) repeat each template 5 times. Thus finally, 1750 video and the corresponding CSI sequences are collected. Our dataset will be released to the public.

B. Overall Performance

We systematically evaluate our system on our dataset through ten-fold cross-validation. Fig.5 respectively indicate the accuracy corresponding to the gesture-only, vision-only, and gesture-vision bi-modal settings. First of all, the bi-modal

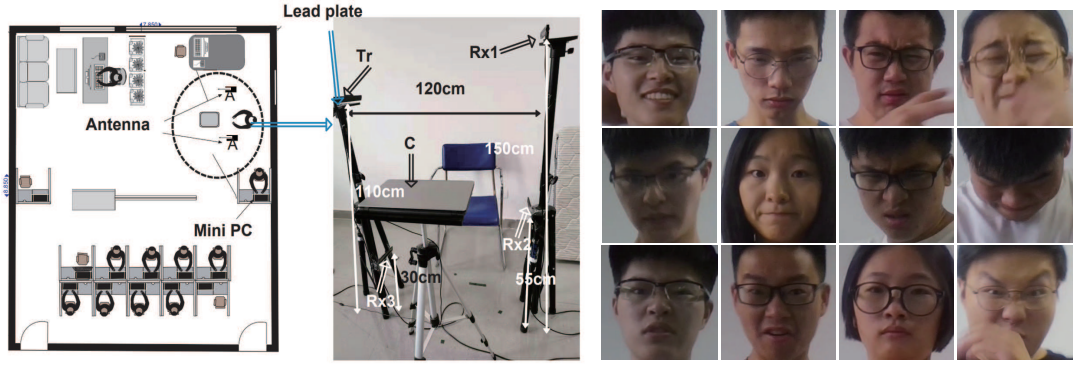

Fig. 4: A snapshot of our our system and its dataset

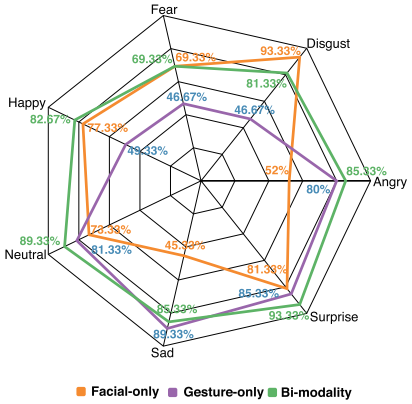

Fig. 5: Recognition accuracy of facial-only, gesture-only and bi-modal recognition

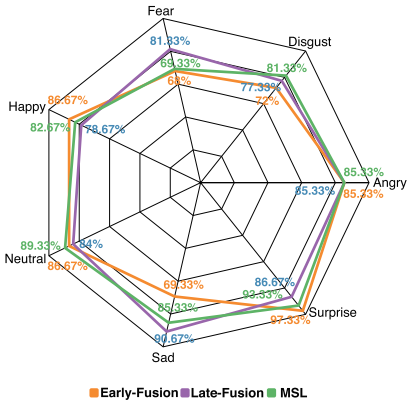

Fig. 6: Comparison between early-fusion, late-fusion and MSL

scheme achieves the best results, and its emotion recognition accuracy is 83.81%, while the accuracy of gesture-only and vision-only schemes are 68.38% and 70.29%, respectively. Moreover, it indicates that our system can leverage the information from two modalities to achieve better emotion recognition than single-modal recognition.

The reason why multi-modality is better than single-modality is that different modalities can complement each other's deficiencies. A direct comparison among three different settings is shown in Fig. 5. When recognizing happy, which ranks second-last (49.33%) for gesture-only method while has the higher recognition accuracy (77.33%) for the facial-only scheme. On the other hand, angry, poorly recognized via the vision-based method (52%), but performance good (80%) in the gesture-based scheme. Lastly, the facial-based scheme performs better than the gesture-based. On average, it achieves 70.29% emotion recognition accuracy compared to 68.38% for the gesture-based method.

TABLE I: Impact of different fusion schemes

Scheme	Accuracy
Late-fusion	83.43%
Early-fusion	80.76%
Multi-Source Learning (MSL)	**83.81%**

C. Compare MSL With Early-Fusion and Late-Fusion

In this section, we compare our proposed MSL-based method with the current mainly used early-fusion and late-fusion schemes. According to related researches, for early-fusion, we connect all features from each encoder to make the decision through a decoder [18]. And for late-fusion and MSL, we use the weighted voting to get the final results after obtaining the recognition result of each encoder, and the decoder structure used in the three schemes is the same. The final recognition results are shown in Table I. Firstly, we can confirm that MSL achieves the best performance compared to early-fusion and late-fusion schemes. It indicates that the information exchange mechanism at the shared decoder is beneficial for emotion recognition. Secondly, both early-fusion and late-fusion methods achieve a good result indicating that both methods are comparable for our dataset.

Figure 6 shows the recognition accuracy of early-fusion, late-fusion, and MSL based schemes. For early-fusion, the recognition results are not even over 7 emotions. Some emotions like sadness can be hardly identified, while some other emotions like surprise have high recognition accuracy. This is because the early-fusion based method combines all features

to train the decoder, during which it pursues the highest overall accuracy, leading to the unbalance performance over different emotions. Late-fusion and MSL are better than early-fusion but due to different reasons. For late-fusion, fusion happens at the decision level and does not affect each modality. For MSL, knowledge exchange in the decoder ensures that the blended features help optimize each encoder.

TABLE II: Impact of modality-missing

	Early-fusion	Late-Fusion	MSL
CSI-only	/	68.19%	**68.38%**
Vision-only	/	**71.62%**	70.29%
Bi-modality	80.76%	83.43%	**83.81%**

Table II 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 to make the final decision. Thus, it cannot work if missing any encoder's feature. Late-fusion achieves 68.19%, 71.62% and 83.43% for the three setting, respectively. Compare to MSL, late-fusion also witnesses the performance improvement from single-modality to bi-modality. Moreover, it supports our previous observation that vision is superior to CSI with MSL, which means that facial expressions constitute a stable and reliable way with rich emotional cues.

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

Schemes	Parameters	Running time (per sample)
Late-Fusion	22,362,515	6.61ms
MSL	**15,067,087**	**1.98ms**

Table II compares the computational complexity between late-fusion and MSL. It is clear that late-fusion embodies $22,362,515$ parameters in its deep network while MSL reduces the number by 32.62%, i.e., $15,067,087$ network parameters. Therefore, a less complicated network leads to faster running time. Late-fusion takes 6.61 ms to process every test sample on average while MSL only needs 1.98 ms. In other words, MSL is 70% faster than late-fusion.

IV. Conclusion and Future Work

This paper proposed a hybrid vision and CSI-assisted emotion recognition system leveraging facial expression and body gesture modalities. We adopted a vision-based facial expression recognition, while exploring the WiFi signal for contactless gesture recognition. We proposed a CSI enhancement method based on Rician fading theory and a Multi-Source Learning (MSL) framework to mine correlations between bi-modal data for better emotion recognition. We built the first WiFi-Vision emotion dataset use only low-cost commodity vision and WiFi devices to evaluate the proposed method. The empirical results showed the superiority of our system.

For future work, we will further study the modalities involved in emotion recognition and the potential relationships between these modalities, and use these relationships to achieve better emotion recognition. We will further explore the potential of MSL, study dynamic weight update methods, and use the attention mechanism to allow the framework to further improve the performance of MSL.

References

[1] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2019.

[2] Yu Luo, Jianbo Ye, Reginald B Adams, et al., "Arbee: Towards automated recognition of bodily expression of emotion in the wild," *IJCV*, vol. 128, no. 1, pp. 1–25, 2020.

[3] Ginevra Castellano, George Caridakis, Antonio Camurri, et al., "Body gesture and facial expression analysis for automatic affect recognition," *Blueprint for affective computing: A sourcebook*, pp. 245–255, 2010.

[4] Yu Gu, Xiang Zhang, Zhi Liu, and Fuji Ren, "Besense: Leveraging wifi channel data and computational intelligence for behavior analysis," *IEEE CIM*, vol. 14, no. 4, pp. 31–41, 2019.

[5] Yu Gu, Xiang Zhang, Zhi Liu, and Fuji Ren, "Wifi-based real-time breathing and heart rate monitoring during sleep," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[6] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng 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.

[7] Sidney K D'mello and Jacqueline Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM CSUR*, vol. 47, no. 3, pp. 1–36, 2015.

[8] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423–443, 2018.

[9] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE SPL*, vol. 23, no. 10, pp. 1499–1503, 2016.

[10] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, et al., "Challenges in representation learning: A report on three machine learning contests," in *ICONIP*. Springer, 2013, pp. 117–124.

[11] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao 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.

[12] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos 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.

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

[14] Hao Wang, Daqing Zhang, Junyi Ma, et al., "Human respiration detection with commodity wifi devices: do user location and body orientation matter?," in *ACM UbiComp*. ACM, 2016, pp. 25–36.

[15] Jia Deng, Wei Dong, Richard Socher, et al., "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*. IEEE, 2009, pp. 248–255.

[16] Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko, "Adashare: Learning what to share for efficient deep multi-task learning," *NeurIPS*, vol. 33, 2020.

[17] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al., "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[18] Juan DS Ortega, Patrick Cardinal, and Alessandro L Koerich, "Emotion recognition using fusion of audio and video features," in *IEEE SMC*. IEEE, 2019, pp. 3847–3852.