International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 01 | January - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

EMOTION MINING USING WIFI AND VISION SETUP

Harshitha M[1], Charitra A N[2], Shivling B[3], Sunil Kumar D[4], Madhan L[5]

[1]*Harshitha M, ISE, Vidya Vikas Institute of Engineering & Technology Mysore*

[2] *Charitra A N, ISE, Vidya Vikas Institute of Engineering & Technology Mysore*

[3]*Shivling B, ISE, Vidya Vikas Institute of Engineering & Technology Mysore*

[4] *Sunil Kumar D, ISE, Vidya Vikas Institute of Engineering & Technology Mysore*

[5]*Madhan L, ISE, Vidya Vikas Institute of Engineering & Technology Mysore*

Abstract- This research investigates emotion recognition using a unique method that integrates computer vision and Wi-Fi signals. By monitoring the Wi-Fi patterns and analyzing the differences in it, and employing vision technology to capture facial expressions, we want to differentiate and understand human emotions in various contexts. Given the multimodal nature of human emotion expressions, we suggest this method that combines vision and Wi-Fi configurations. Additionally, we use the link between modalities for improved emotion recognition by employing the Multi Source Learning (MSL) approach, which was motivated by Multi-Task Learning.

Key Words—emotion recognition, channel state information, vision, dataset, multi-source learning

I. INTRODUCTION

Emotion recognition plays a important role in marketing, healthcare, human-computer interaction, and some medical therapies. Here, we can recognize emotions by the expressions that come out of a human's face and specific actions related to specific emotions. There are many challenges to be faced during recognition; we require a high-quality camera to capture gestures and actions of humans and should build a Wi-Fi and vision emotion dataset to verify or recognize the emotions expressed by humans. We need to conduct experiments using these modalities to verify efficacy of our setup and its superiority over single modality by conducting an experiment.

The challenge lies in capturing emotional expressions without disturbing the subject, as current research primarily uses cameras for facial expressions and wearable sensors for body gestures, which can interfere with objects and potentially contaminate emotional cues. The paper proposes a CSI enhancement model for

WiFi, a deviceless alternative sensor for gesture capture, to improve its sensitivity to human motion. The multipath effect on the human body allows COTS WiFi to capture human movement, but the CSI's sensing granularity is crucial for gesture recognition.

The challenge lies in leveraging the correlation between facial and gesture data for better emotion recognition. Current research relies on early-fusion or late-fusion methods, which are susceptible to data loss and require multiple decoders. To address this, a Multi-Source Learning (MSL) framework, inspired by Multitask Learning, is proposed. MSL extracts useful features for each modality and feeds them to a shared decoder. Parameter sharing enables different modalities to exchange knowledge, extract cross-correlated features, and output recognition results. The final decision is made through voting over all outputs.

The study tested a bi-modality-based method using low-cost vision and WiFi equipment, achieving 83.81% recognition accuracy. Experiments confirmed its effectiveness, surpassing gesture-only and facial-only solutions. The method also demonstrated better robustness against data loss and less computing consumption compared to early-fusion and late-fusion methods.

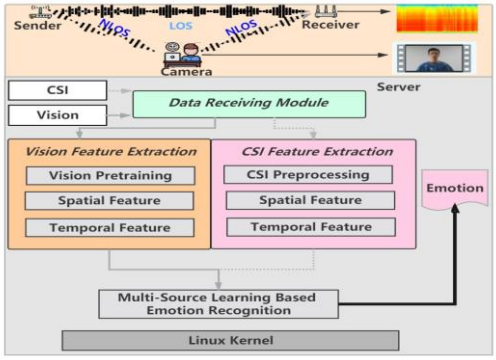


Fig.1: System Overview

In summary, our contributions are summarized as follows:

- We are the first to use WiFi and vision for contactless emotion recognition and have developed 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.
- The Multi-Source Learning framework was designed to enhance emotion recognition by leveraging cross-correlation between modalities, and its effectiveness was verified using our dataset.

II System Design

A. Overview

Figure 1 provides an overview of our emotion recognition system. It is made up of three data processing components: MSL-based emotion recognition, Wi-Fi-based gesture feature extraction, and vision-based facial expression feature extraction.

The Multi-task Cascaded Convolutional Network (MTCNN) are utilized in first section to identify face frames in videos and crop the frame to the appropriate size. Next, we extract the spatial characteristics from these clipped frames using two different types of Densenet. Lastly, we extract the temporal features using the VGG-LSTM network.

The Ricean fading-based model is initially used in the second section to improve CSI and better capture the fine-grained body motions. Next, we create CSI maps out of the data so that convolutional neural network can process it. Lastly, we extract temporal and static features from CSI maps using VGG-LSTM, Densenet169, and Densenet121, respectively.

In the final section, we suggest an MSL framework for bi-modal emotion identification that draws inspiration from multi-task learning. Specifically, MSL examines the relationship between motions and expressions on the face that represent the same emotion by fusing visual and CSI data. We go into detail about our system in the next section.

B. Vision Based Facial Expression Feature Extraction

In this section, we explain method for identifying emotional aspects in recorded videos. First, we go over how to accurately crop faces from video frames. Next, we pre-train the networks we employed in this paper using the FER2013 face dataset, building on earlier work to aid in a faster convergence of the network. In order to extract the emotionally relevant aspects of these videos for bi-modal emotion recognition, we finally make use of the pre-trained networks.

1) Face detection and Alignment: For vision-based emotion identification, stable face tracking is essential. Using the Multi-task Cascaded Convolutional Network, we crop the faces in this study. Compared to the dlib detector (a widely used picture tool for face identification), MTCNN can recognize more faces and also modify the head position

2) Perception Pre-training: Pre-training is frequently utilized in computer vision applications due to the abundance of well-structured and well-labelled picture data sets. Pretraining is comparable, on the one hand, to the initial parameter setting process that is necessary for the later training process.

3) Vision Feature Extraction: In accordance with earlier studies [11], [12], we use two depth-based methods—Densenet and VGGLSTM—to implement vision-based feature extraction; Densenet is used for static data and VGG-LSTM for temporal features.

C. *Wi-Fi-based Gesture Extraction*

1) CSI Collection: A type of fine-grained physical layer (PHY) data, channel state information (CSI) details the signal's attenuation factors on each transmission line, including power decay of distance, multipath fading or shadowing fading, scattering, and other data.

2) CSI Pre-processing: Capturing emotional expressions, particularly the subtle ones like a barely noticeable nod, depends heavily on CSI's sensitivity to gesture. This section presents a CSI enhancement model that uses Rician fading to reduce gesture-unrelated information on the channel response and highlight gesture-induced information.

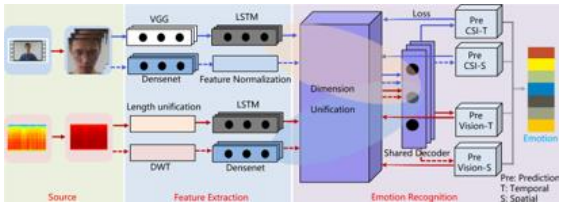


Fig 2: MSL Based Emotion Recognition

3) CSI Feature Extraction: The CSI data needs to be processed in order to eliminate background noise following CSI augmentation. Due to electromagnetic and ambient noise interference, the recorded CSI data has a significant quantity of Gaussian white noise. We filter the CSI data prior to recognition in order to retrieve pertinent information while eliminating unnecessary information.

III. DATASET CONSTRUCTION AND PERFORMANCE EVALUATIONS

In this part, we first provide the WiFi-Vision dataset and then assess our system using it.

A. WiFi-Vision Dataset Construction

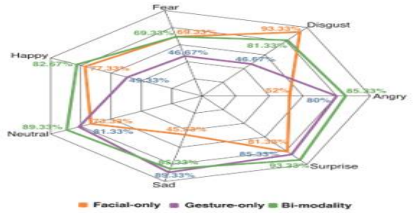
As far as we are aware, there isn't a WiFi-Vision bi-modal dataset. In order to assess our system, we create the first WiFi-Vision emotion dataset. We used the Acted Facial Expressions in the Wild (AFEW) [17] dataset to model the behaviour of individuals with various emotional states. Our WiFi-Vision dataset's body gesture is calibrated using the AFEW dataset as a guide. A preview of our bi-modal system and a few samples of our soon-to-be-public dataset are displayed.

We employ a laptop to capture the video data during the dataset collection process, and two Mini PCs (Intel 5300 NIC) with four antennas are used to retrieve the CSI, as depicted in Figure 4. Our dataset includes 35 different gesture and facial expression templates, as well as seven different emotion categories (i.e., angry, disgust, fear, happy, neutral, sad, and surprised) (5 templates are selected for each emotion). Ten participants, ages 23 to 25, seven of them male and three of them female, repeat each template five times. In the end, 1750 video and the associated CSI sequences are gathered. The public will have access to our dataset.

C. Overall Performance

We use ten-fold cross-validation to rigorously assess our system on our dataset. The accuracy associated with the gesture-only, vision-only, and

gesture-vision bi-modal setups is shown in Fig in that order. To begin with, the bi-modal



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

This section presents a comparison between the currently popular early-fusion and late-fusion systems and our suggested MSL-based approach. Related studies state that in early fusion, we connect all of the information from every encoder to use a decoder to determine the outcome . Additionally, after getting the recognition results from each encoder, we employ weighted voting to determine the final results for late-fusion and MSL. The decoder structure utilized in all three schemes is the same. Table I displays the final recognition results. First, we can verify that, when compared to early-fusion and late-fusion techniques, MSL gets the best performance. It suggests that the shared decoder's information-sharing mechanism helps with emotion identification. Second, successful outcomes are obtained by both early-fusion and late-fusion techniques.

demonstrates how well early-fusion, late-fusion, and MSL-based methods can be recognized. The recognition outcomes for early-fusion are not even greater than seven emotions. Certain emotions, like melancholy, are difficult to identify, whereas other emotions, like surprise, are highly accurate. This is because the decoder is trained using an early-fusion based approach that integrates all features in an attempt to achieve the maximum possible overall accuracy, which results in an unbalanced performance across various emotions. For many reasons, late-fusion and MSL are superior to early-fusion. In the case of late-fusion, fusion occurs at the decision-making level and is independent of modality. The decoder's knowledge exchange guarantees that each encoder is optimized for MSL by utilizing blended characteristics.

IV. CONCLUSION AND FUTURE WORK

A hybrid vision and CSI-assisted emotion recognition system that makes use of body gesture and facial expression modalities was presented in this study. While investigating the WiFi signal for contactless gesture identification, we decided for vision-based facial expression recognition. We suggested a Multi-Source Learning (MSL) framework and Rician fading theory-based CSI enhancement technique to mine correlations between bimodal data for improved emotion identification. In order to access the suggested approach, we created the first Wi-Fi-Vision emotion dataset using only inexpensive commodity vision and Wi-Fi sensors. The empirical findings demonstrated our system's superiority.

In order to improve emotion detection, we will investigate the modalities involved in emotion recognition in more detail as well as any possible linked between them in future research. We will investigate MSL's potential further, research dynamic weight updating techniques, and make use of the attention mechanism to enable the framework to enhance MSL's performance even more.

Reference

[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 wi-fi 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, Chanshu wu, and Zheng Yang, "Zero-effort cross-domain gesture recognize 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 Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Mul-ˇ timodal machine learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423–443, 2018.