Wearable-Based Human Emotion Inference System

Zirui Zhao(✉) and Canlin Zheng

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
2362534742@qq.com

Abstract. Many emotion recognition methods which have been proposed today have different shortcomings. For example, some methods use expensive and cumbersome special-purpose hardware, such as EEG/ECG helmets, while others based on cameras and speech caused risk of privacy leakage. With the prosperous development and popularization of wearable devices, people tend to be equipped with multiple smart devices, which provides potential opportunities for lightweight emotional perception. Based on this, we take actions on developing universal portable system and multi-source wearable sensing technology devices.

This paper designs an emotion recognition framework called MW-Emotion (Multi-source Wearable emotion recognition) featuring on low cost, universality, and portable commercial wearable devices to perceive multi-source sensing data and implement a system. It takes four basic emotions as the research object and implement an emotion recognition to explore deep context innovatively. The experimental results show that MW-Emotion has a recognition accuracy of 85.1% for person-dependent mode. The framework uses the method that different types of data are effectively fused through signal processing. We call it multimodal data fusion technology, which reduces the energy waste caused by data redundancy and effectively resists interference.

Keywords: Wearable devices · Emotion recognition · Multimodal data · Multiple perception

1 Introduction

Nowadays, the fast-paced life has brought increasing psychological pressure to modern humans. More and more people have symptoms such as emotional instability and long-term depression, and even suffer from psychological diseases such as anxiety and depression. According to Report on National Mental Health Development in China [11] from 2019 to 2020 released in March 2021, the results show that up to 31.1% of college students have depression or anxiety tendencies. The report points out that China's national mental health counseling services are in short supply and there is a big gap in service level. In the choice of psychological service methods, more than half of people tend to choose "self-regulation". The first step of self-regulation is to know your emotional state in time. To solve these problems, in recent years, researchers capture emotions

K. Wu et al. (Eds.): ICECI 2021, LNICST 437, pp. 128–135, 2022.
https://doi.org/10.1007/978-3-031-04231-7_12

through various techniques, such as capturing facial expressions, body postures, walking gait, and other limb body language of the human trunk for emotional inference, using microphones to record voice to extract semantic information or speech prosody information, and physiological signals, like Electroardioram (ECG), Electroencephalogram (EEG), Galvanic Skin Response apparatus (GSR) *etc*. However, these methods have the risk of privacy disclosure, and need to wear expensive and bulky professional sensor equipment.

To overcome these shortcomings, wearable devices have gradually become popular and diversified, and modern people even use a variety of wearable devices at the same time. Commercial wearable devices have gradually improved in terms of capacity, performance, and intelligence, providing new possibilities and huge opportunities for reliable and lightweight emotional exploration. Wearable devices are embedded with many physical sensors. These body language sensors can detect the user's movement, monitor physiological signals, and capture images, sounds, and videos. With the further development of the identification, including sensor technology and artificial intelligence technology, and with the help of low-cost commercial portable wearable devices, not only can physical data such as the user's heartbeat, steps, and gestures be obtained, it can also make deep Hierarchical emotional calculations possible. Different types of multi-modal emotional related data can be collected in an unobtrusive manner at the same time, which greatly enhances the portability and ease of use for users. To sum up, compared with previous related work, this new emotion recognition method based on low-cost, universal and portable wearable device perception data has the following three outstanding advantages: Equipment universality, Privacy security and Emotional validity.

2 Related Work

Emotion recognition has always been a hot research issue in the field of artificial intelligence. Its multi-faceted application value and academic value have caused extensive research by scholars and companies at home and abroad.

Body Language. Body language is a basic type of human social behavior. It is an important indicator of a person's emotions and inner state. It contains different types of non-verbal expressions, such as facial expressions, eye movements, body posture, walk gait, *etc*. However, this method has the following shortcomings: (1) expensive and requires huge calculations; (2) easy to cause privacy leakage; (3) The problem of emotion hiding (4) requires continuously track.

Speech Signals. There are two aspects. One is based on the acoustic characteristics of speech and the other is based on the semantic content of speech, features in the speech signal such as frequency spectrum, Mel cepstrum coefficient (MFCC), *etc*., and makes emotional judgment by analyzing the polarity of emotional characteristic words. However, there is a risk of leaking the content of the user's voice, and it is greatly affected by the difference in the habits of individual expressing emotions. It is also easy to pretend that the true internal emotions cannot be measured.

Physiological Signals. Common physiological signals for emotion recognition embraces EEG, ECG, GSR, Radar Signal Processor (RSP), *etc.* These physiological signals are more related to people's internal emotional state, but they are too complicated and bulky to carry (Figs. 1 and 2).

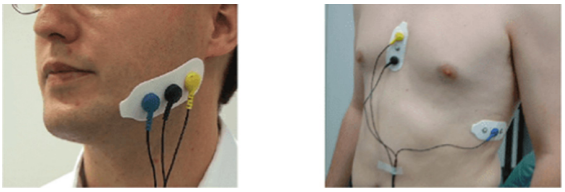

Fig. 1. EMG sensor applied to the jaw and ECG sensor applied to chest [1, 6].

Fig. 2. Respiration sensor and Skin conductivity, temperature and Finger blood volume pulse (BVP) sensor [1, 6].

Usage Pattern Perception. This emotion recognition method mainly analyzes and models the usage of sensors such as accelerometers, gyroscopes, microphones, GPS, light sensors, compasses, screen pressure sensors, and various applications in the mobile terminal to perform emotion recognition. However, the accuracy and reliability of this method still need to be improved.

Multimodal Perceptual Signals. The method is better and comprehensive than previous methods by integrating two or more different signals of the above technologies, and uses the information obtained from multiple modalities to cooperate to improve the performance of the model. Different modalities can be expressed in different sources or forms of information, such as sound modalities, image modalities, text modalities, EEG, EMG, ECG, and other domain modalities. People live in such an environment where multiple modes blend with each other, and integrate with what the machine hears, sees, hears, and feels, to give the machine and model better performance. Although this method of synthesizing multimodal sensing signals has the advantages of accuracy, it also has their disadvantages.

3 System Design

In this experiment, we use Multimodal Perceptual Signals as our experimental signals, then we design a system called MW-Emotion to sense the emotions and process them.

3.1 Multi-modal Sensing Signal Design

Compared with the above related work, the advantage of MW-Emotion system is that it can be non-invasive and portable to sense human emotions. The cost of the whole system is much lower than professional emotion sensing devices. It only needs headphones, watches, and smart glasses. What's more, it can be carried around and tested for a long time. More specifically, the MW emotion sensing system uses five mainstream portable micro wearable intelligent data sensing devices to collect multi-source and multi-modal data. These five devices are smart wristbands, smart headphones, smart glasses, smart necklaces and smart phones, which are worn on the user's head, ears, wrist, neck and nose. These conventional low-cost sensors on these intelligent devices, including microphone sensor and optical capacitance pulse sensor PPG, are used to collect the user's pulse signal and body sound signal respectively. The body sound signals include heart beating sound, nasal breathing sound, laryngeal trachea sound and other body sound signals. After sensing the data acquisition, the collected signals are integrated and sent to the mobile terminal in the form of data frames for processing. The mobile terminal further automatically and in real time infers emotions.

The wearing position and perceived data type are as follows. There are 5 main positions. Position 1 is an in-ear headset with a built-in micro microphone. When wearing, it forms a resonant cavity with the ear canal to amplify the sound of heartbeat; Positions 2 and 3 are microphones, which are respectively used to collect the tracheal sound and nasal sound of the throat. Breathing sound, tracheal sound and other body non speech sounds can be captured by these two microphone sensors. Positions 4 and 5 are optical capacitance pulse sensors, which can monitor the wearer's heartbeat. These physiological signals are closely related to human emotion and have the potential to reflect emotion.

At present, many researchers have explored and proved the correlation between respiration, heartbeat and other body sound signals and emotion from a biological point of view. For example, Ikuo Homma [3], a Japanese physiologist, discussed the relationship between breathing and emotion from the perspective of brain neurology, and pointed out that autonomous breathing is not only controlled by metabolic needs, but also constantly responds to emotional changes.

3.2 Multi-modal Signal Preprocessing

After collecting the multi-modal original signal, we first subtract the average value of each channel signal to remove the DC drift. In addition, due to the interference of nearby power lines, the power supply is 50 Hz AC current, so the collected signal contains 50 Hz component and corresponding odd harmonics. Nevertheless, there is obvious power frequency interference here. To reduce this, we first FFT the signals of each channel, set the corresponding coefficients of these interference components to zero, more specifically, use the suppression bands of [49,51] Hz, [149152] Hz and [249251] Hz to cut off the frequencies of the five channels, and then use the inverse fast Fourier transform (IFFT) to transform the signal back into a time domain signal sequence.

Preprocessing for Different Signals: For the five signals, different signal processing methods are adopted due to different signal types and acquisition positions. The following five locations are the five positions described above.

- Location 1: Since the main frequency of heartbeat signal is in [1, 6] Hz, we filter this channel using band-pass FIR filter with [0.1, 10] Hz to ensure information redundancy.
- Location 2/3: The throat and nose channels mainly detect breathing information, including the respiratory rate and depth. They also record non-speech body sounds such as laughter and crying.
- Location 4/5: For PPG signals, we first filter it with [0.1,10] Hz band-pass FIR filter. However, the PPG signals are affected by sensor and skin contact and whether sweating. Long time experiments lead to signal stability and low-frequency drifts. To eliminate such interference, we directly band stop the components less than 0.5 Hz after FFT transform, and then recover the signal through IFFT.

Normalization: Due to variations in quantization standards for different signals, normalization is necessary. Each channel signal is soft-normalized using the 5th and 95th quantiles.

3.3 Experiment Setup

In the experiment, we invited 30 users to participate in data collection, and used video or audio materials with different emotional colors to stimulate users' emotions. To get enough data, we let users self report their emotions during this period after observing each stimulus material.

3.4 Emotion Recognition Model

After the above steps, we use DenseNet as the recognition model. To improve its classification ability, we make full use of each layer of information. We use short-time Fourier transform (STFT) to convert each channel signal into time-frequency spectrum, and then feed them into the model.

STFT: The short-time Fourier transform (STFT) is to select a time-frequency localized window function, assuming that the analysis window function g(t) is stationary (pseudo stationary) in a short time interval. Move the window function so that f(t)g(t) is a stationary signal in different finite time widths, so as to calculate the power spectrum at different times. It uses a fixed window function. Once the window function is determined, its shape will not change, and the resolution of the short-time Fourier transform will be determined. STFT is a good choice for analyzing piecewise stationary signals or nearly stationary signals.

DenseNet121 Model: A milestone in the history of CNN is the emergence of ResNet model. ResNet can train deeper CNN models in higher accuracy using "skip connection". DenseNet achieves better performance than ResNet with fewer parameters and

computational costs. The core of DenseNet model is to establish "dense connection" between all front layers. Another superiority of DenseNet is achieving feature reuse through connections between features on the channel.

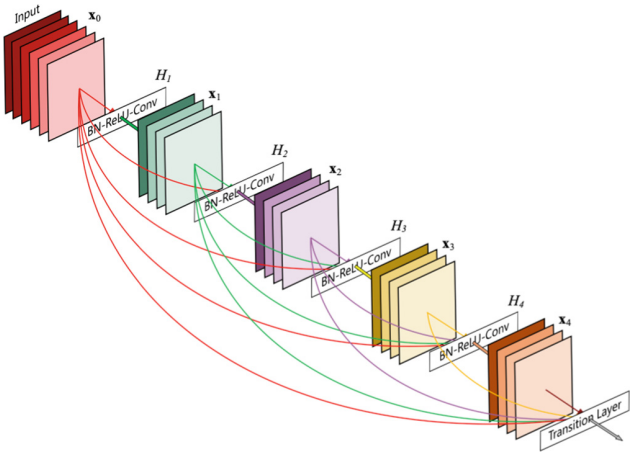

Fig. 3. A 5-layer dense block with a growth rate of k = 4. Each layer takes all preceding feature-maps as input [9].

For a L layer network, DenseNet contains $\frac{L(L+1)}{2}$ connections, which is a dense connection compared with ResNet. All previous layers are connected as input:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{1}$$

Among them, the above $H_L(\cdot)$ represents non-linear transformation. It is a combined operation, which may include a series of BN, ReLU, Pooling and Conv operations. There may be multiple convolution layers between the L layer and the $L-1$ layer.

DenseNet have many advantages. Due to the dense connection mode, DenseNet improves the gradient back propagation, making the network easier to train; It also realizes short-circuit connection through concat feature, realizes feature reuse, and adopts a smaller growth rate. The Characteristic Map of each layer is smaller, and the implicit "deep supervision" is realized.

4 Experiment

In our experiment, we used four public emotional stimulation data sets, namely Stim Film [4], EMDB [5], IADS [2] and SEED [7], which included video and audio widely used to induce specific types of emotions in psychological experiments to collect users' emotions. The data are collected by five wearable devices on different locations as shown in Fig. 3.

4.1 Results and Evaluation

We have tested DenseNet's learning model. Figure 4 shows the confusion matrix of the DenseNet implementation.

We used four basic emotions in the psychological sense, including Happiness, Neutrality, Sadness and Mixed emotions. Mixed emotions (Mix) refer to fear, nausea and anger. According to the existing emotion category paradigm in psychology, especially the two-dimensional emotion paradigm, the definitions of these three kinds of emotions are very similar in the emotional paradigm of psychology. During the experiment, by asking the experimenters, it is found that sometimes users cannot effectively distinguish these three kinds of emotions, because these three emotions often occur at the same time. Therefore, these emotions are mixed here.

We observe that other emotions are easily confused with neutrality, which may be caused by the actual aroused emotional intensity. In addition, the recognition accuracy of MW emotion is slightly lower than that of EQ radio [8] (*i.e.*, 87%), and EQ radio uses the data collected from the same subject to train and test the model. However, its performance is still better than memento [10], which [10] identifies six fitness types with an accuracy of 57%.

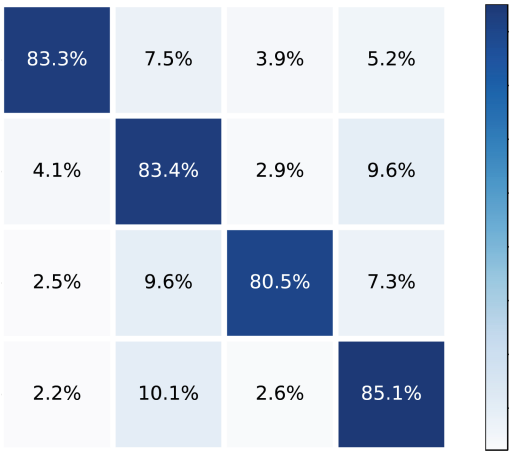

Fig. 4. The confusion matrix achieved by the DenseNet

Under the current setting, MW emotion can still achieve considerable recognition performance.

5 Conclusion

We propose a new low-cost, popular, and portable multi-mode wearable emotion recognition system, called MW emotion. The system can recognize emotional states in order to protect privacy, reliable detection and easy to use in daily life. Our experiments show that MW emotion can use multi-mode wearable devices to identify four basic emotional

states with high precision, thus showing the potential in the future, that is, using low-cost, popular and portable wearable devices can perceive everyone's emotions.

This study reveals the potential correlation between shallow data and deep data perceived by low-cost, universal and portable wearable devices. It is a novel emotion inference technology and a new exploration to realize the perceptual interaction between machines and people. Its novelty is that it can noninvasively detect the potential relationship between emotion and more hidden and imperceptible physiological activities (such as breathing, heartbeat and other types of body sound). We handle these by designing appropriate data quality enhancement techniques and deep learning network, and we conduct real-world experiments to evaluate MW-Emotion and proves its favorable performance. To the best of our knowledge, MW-Emotion is the first work to demonstrate the feasibility to recognize emotion state using low-cost, pervasive and portable wearable devices.

With the increasing development of wearable devices, emotion recognition will serve people in a reliable, lightweight and simple way in the future. This technology has broad application prospects, such as intelligent recommendation based on emotional state, timely and continuous monitoring of emotion and emotional relief.

References

1. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: first steps towards an automatic system. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 36–48. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24842-2_4
2. Bradley, M.M., Lang, P.J.: The International Affective Digitized Sounds (IADS-2): Affective ratings of sounds and instruction manual. University of Florida, Gainesville, FL, Technical report B-3 (2007)
3. Homma, I., Masaoka, Y.: Breathing rhythms and emotions. Exp. Physiol. **93**(9), 1011–1021 (2008)
4. Schaefer, A., Nils, F., Sanchez, X., Philippot, P.: Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers. Cogn. Emot. **24**(7), 1153–1172 (2010)
5. Carvalho, S., Leite, J., Galdo-Álvarez, S., Gonçalves, O.F.: The Emotional Movie Database (EMDB): A self-report and psychophysiological study. Appl. Psychophysiol. Biofeedback **37**(4), 279–294 (2012). https://doi.org/10.1007/s10484-012-9201-6
6. Koelstra, S., et al.: Deap: a database for emotion analysis; using physiological signals. IEEE Trans. Affective Comput. **3**(1), 18–31 (2011)
7. Duan, R.N., Zhu, J.Y., Lu, B.L.: Differential entropy feature for EEG-based emotion classification. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 81–84. IEEE (2013)
8. Zhao, M., Adib, F., Katabi, D.: Emotion recognition using wireless signals. In: Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, pp. 95–108 (2016)
9. Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., Weinberger, K.: Convolutional networks with dense connectivity. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
10. Jiang, S., Li, Z., Zhou, P., Li, M.: Memento: an emotion-driven lifelogging system with wearables. ACM Trans. Sens. Netw. (TOSN) **15**(1), 1–23 (2019)
11. Zhang, X., Lewis, S., Firth, J., Chen, X., Bucci, S.: Digital mental health in China: a systematic review. Psychol. Med. 1–19 (2021)