# Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals

Yassine Ouzar, Frédéric Bousefsaf, Djamaleddine Djeldjli, Choubeila Maaoui

# Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals

Yassine Ouzar, Frédéric Bousefsaf, Djamaleddine Djeldjli, Choubeila Maaoui
Université de Lorraine, LCOMS, F-57000, Metz, France
{yassine.ouzar, frederic.bousefsaf, djamaleddine.djeldjli, choubeila.maaoui} @univ-lorraine.fr

## Abstract

*Human's affective state recognition remains a challenging topic due to the complexity of emotions, which involves experiential, behavioral, and physiological elements. Since it is difficult to comprehensively describe emotion in terms of single modalities, recent studies have focused on fusion strategy to exploit the complementarity of multimodal signals. In this article, we study the feasibility of fusing facial expressions with physiological cues on human emotion recognition accuracy. The contributions of this work are threefold: 1) We propose a new spatiotemporal network for facial expression recognition using a 3D squeeze and exitation based 3D Xception architecture (squeeze and exitation Xception network). 2) We adopt the first multiple modalities fusion using single input source which, to the best of our knowledge, no existing multimodal emotion recognition system has attempted to identify emotional state from only facial videos using facial expressions and physiological signals features. 3) We compare the performance of the unimodal approach using only facial expressions or physiological data, to multimodal systems fusing facial expressions with video-based physiological cues. In our experiments, physiological signals such as the iPPG signal and features of heart rate variability measured remotely using the imaging photoplethysmography (iPPG) method are used. The preliminary results show that the multimodal fusion model improves the accuracy of emotion recognition, and merging facial expressions features with iPPG signal gives the best accuracy with 71.90 %.*

## 1. Introduction

Human faces are a rich source of information. They are characterized by a great expressive richness to convey emotions, which makes them widely used to identify a person's emotional state through facial expressions. Despite the impressive results achieved by facial expressions recognition systems on acted databases with controlled conditions [31, 49, 54, 55], they are rarely faced with real situations. In a natural environment, reliability cannot be guaranteed and performance degrades considerably [20, 32, 43]. In addition to environmental conditions (camera angles, lighting conditions and occlusion of multiple parts of the face) and the ability to control and fake emotions by people, facial expressions are also more affected by social and cultural differences. Human expressiveness can vary among individuals and can be expressed differently. Additionally, facial expressions can be a mix of different emotion status that occur at the same time or may not be expressed at all. Consequently, using facial expressions to identify person's emotional state can lead to wrong inferences.

Recently, few studies have proposed emotion recognition systems that use physiological cues extracted from the face using the imaging photoplethysmography method [2, 30]. The advantage of using physiological parameters to assess emotion compared to facial expressions is : physiological data are a response to the autonomic nervous system (ANS), which is involuntarily activated and therefore uncontrollable.

Most existing studies have examined the use of facial expressions and physiological cues separately [12, 24, 31, 38, 49]. However, little attention has been paid to a fusion between these two modalities [8, 18, 51]. Combining the two can improve recognition accuracy and provide greater reliability by continuously gathering information about the person's emotional state despite missing acquisition or misleading information that may occur when using a single modality, operating in a noisy environment or in the case of falsified expression. Additionally, fusion of multiple modalities can help to compensate errors and resolve ambiguities by learning useful representations of data of different nature. However, The main limitation is related to asynchrony across modalities, which are usually unaligned. In addition, physiological data are collected through intrusive devices that are psychologically stressful and this can modify the measurement results of physiological signals. Therefore, this will certainly affect the accuracy of emotion scoring [9]. In this work, we propose the first video-based

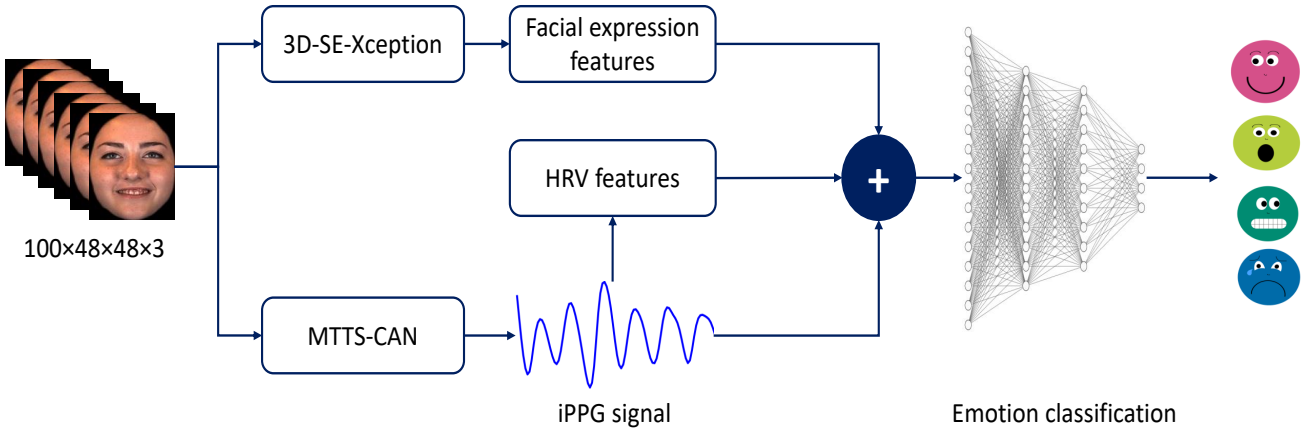



Figure 1. Proposed system for multimodal emotion recognition using facial expressions, iPPG signals and HRV features.

multimodal spontaneous emotion recognition that combines facial expressions with physiological data to derive the advantages of each modality.

In this paper the physiological parameters are measured from facial video recordings based on imaging photoplethysmography principal [6]. While, facial expressions features are extracted using a new spatio-temporal network that combines 3D squeeze and exitation module with 3D Xception architecture. The features vector of facial expressions is then merged with the physiological signals to ultimately estimate the corresponding emotion. In the remainder of this paper, human emotion recognition related works are presented in Section 2. Section 3 details our proposed approach. Then, in Section 4, our method is evaluated. Finally, conclusions and future works are given in Section 5.

## 2. Related works

In literature, various modalities have been used to recognize emotion either in unimodal [12, 36, 45, 47] or multimodal way [5, 18, 39]. Initial research on unimodal emotion recognition systems have focused on the expressiveness of the face because it is visible and it is easier to collect a large set of facial data. The commonly adopted methods for facial expression recognition are either deep learning or handcrafted based approaches [22, 25]. However, deep learning techniques have made a great success due to their high generalizability for new data and their ability to automatically extract robust features and learn complex nonlinear representations. Today, the state of the art deep learning methods allow to achieve a categorization of facial expressions with a reliability of around 98% in controlled situations [23]. Nevertheless, several real environment issues can degrade recognition accuracy such as lighting variations or background appear [28]. Additionally, deep learning algorithms often fail in the case of expressionless faces or falsified expressions.

To address this issue, some attempts have been made to identify emotion through physiological data that are managed by the autonomous nervous system (ANS) which is involuntarily activated and therefore can not be controlled [12]. Physiological signals such as electroencephalography, electrocardiography, skin temperature and electromyography are reliable data for quantifying emotions [10]. However, they are acquired by intrusive contact sensors that can interfere with the subjects and modify their emotional state. Moreover, the complexity of measurement and the sensitivity of the electrodes of these devices strongly limit their scope of application, since they cannot be used outside of the laboratory. Therefore, recent studies have focused on wearable devices that provide various biosignals such as blood volume pulse (BVP) and electrodermal activity and their derivatives to explore new application fields. Going even further, recent works have used heart rate variability measured by the camera to detect emotional state [3, 30]. They rely on imaging photoplethysmography method, which allows non-contact extraction of the blood volume pulse signal from facial video recording, making it more interesting and promising among the other physiological signals that require contact devices and the presence of a specialist to monitor them.

Numerous literature studies show that multimodal emotion recognition systems outperform unimodal approaches [11, 33]. For this reason, several works have merged facial expressions with physiological data to develop reliable systems [8, 18, 21]. Despite the obtained results , they follow a constrained experimental setup under laboratory conditions due to the use of intrusive and sensitive equipment. In addition, dealing with multiple signals of different nature gathered from different sources, may conflict with each other due to asynchrony across modalities and thus lead to misestimation.

## 3. Materials and Methods

### 3.1. Dataset

Although many multimodal emotion databases are available, few of them provide physiological signals. The existing datasets for multimodal emotion recognition from facial expressions and physiological signals are quite limited not only in data size but also in diversity. In this study we explore a new multimodal sponatneous emotions database named BP4D+ [53]. Compared to existing datasets such as MAHNOB [42] and DEAP [19], BP4D+ is a large scale dataset that includes annotated action units (AUs) and discrete emotion categories. In addition, it contains numerous challenging conditions and diversity in terms of significant head motion and ethnic diversity, making it more interesting and challenging. Since its creation, BP4D+ has been widely used in several works related to affective computing and vital signs measurement [26, 46, 50].

This dataset consists of RGB and thermal images, 2D and 3D facial landmarks, actions units and 8 physiological signals collected with contact sensor. 140 subjects (82 females and 58 males) of different ethnic ancestry participated in 10 sessions designed to induce the following emotions : Happiness (T1), Surprise (T2), Sadness (T3), Startle (T4), Skeptical (T5), Embarrassment (T6), Fear (T7), Pain (T8), Anger (T9), and Disgust (T10). 1400 RGB videos lasting 30 seconds to 1 minute were recorded at a frame rate of 25 fps. The resolution of each image is $1040 \times 1392$ pixels. Among the 10 tasks, only four emotions are used in our experiments, corresponding to happiness, embarrassment, fear and pain. These emotion tasks are provided with manually coded action units (33 in total) that were computed only for the most expressive frames of each task.

### 3.2. Data preparation

First, the most expressive frames are extracted from each emotion task using action units code provided in the database. Then, we follow the same protocol used in [37]. A robust face swapping-based segmentation method is used to get rid of non-skin regions that do not hold any color changes associated with cardiac activity [35]. This step improves imaging photoplethysmographic signal extraction from face skin. All the images of the segmented faces are cropped according to the coordinates of the non-zero pixels and then scaled to $48 \times 48 \times 3$. Besides, data augmentation strategy is applied for the training set to create additional and different training instances. Several image transformations such as rotating the image by varying degrees, translating it and flipping it horizontally and vertically, cropping, zooming in, or changing the contrast of the image have been randomly applied on video fragments. It helps to reduce overfitting and improve the generalizability of the model.

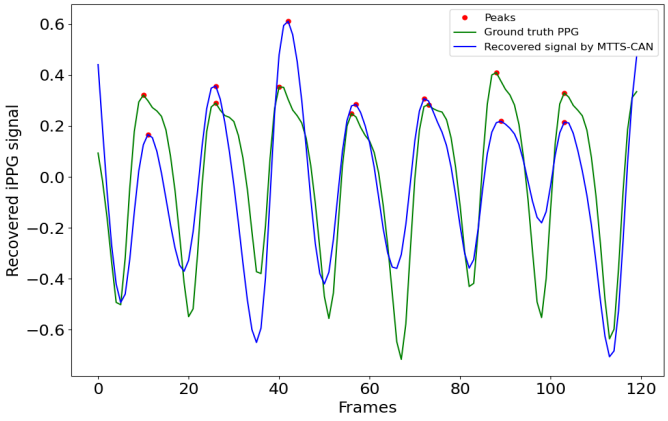



Figure 2. Comparison between a predicted signal by MTTS-CAN and the ground-truth PPG signal taken from BP4D+ dataset.

### 3.3. Video-based physiological signals measurement

In this study, physiological parameters are measured remotely using imaging photoplethysmography method [6]. iPPG is an optical technique for capturing cardiac signals by observing the blood-volume variations on a person's face using a simple camera. The captured light reflected by the skin is translated to a variation of the iPPG signal. Several important vital signs can be derived from the iPPG waveform such as pulse rate, respiration rate and heart rate variability (HRV). However, among these physiological features, only iPPG signal and its derivative HRV features have been used in our experiment. It was reported in several studies that heart rate variability is one of the most important physiological characteristic that reflects affective states of a person [3, 30]. HRV features can be derived from time interval variation between consecutive heartbeats in iPPG signal. [14].

iPPG extraction algorithms can be divided to hand-crafted based algorithms [52] that use signal/image processing steps and deep learning based approaches [34]. In this work, we used a multi-task sequential shift convolutional attention network (MTTS-CAN) proposed by Liu et al. to extract the iPPG signal [29]. MTTS-CAN is one of the recent popular state-of-the-art deep learning based method that provides good performance in terms of heart and respiratory rates measurement. In order to better appreciate the quality of the recovered iPPG signal, we present, in Figure 2, a superposition of a ground truth PPG signal recorded by contact sensor and the iPPG signal predicted by the MTTS-CAN network. It is clear that the estimated iPPG signal is strongly correlated with the ground truth and the location of the peaks is very close.

The core module of MTTS-CAN is a hybrid network that uses the attention mechanism in conjunction with Tem-

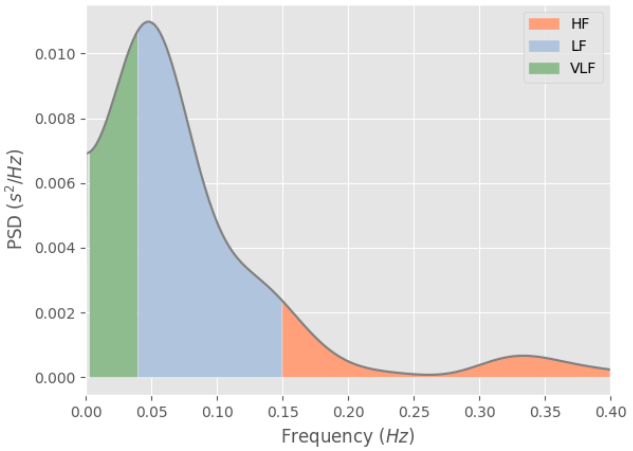



Figure 3. A representative PSD for IBI signal showing The areas of VLF, LF and HF powers of the HRV.

poral Shift Modules [29]. The recovered iPPG signals by MTTS-CAN allow HRV features extraction both in the time-domain and in the frequency-domain. For both time and frequency analysis, peak detection is performed to locate the instant of time at which heartbeat occurs (which allows to compute HRV features).

In time domain, heart rate is calculated as the inverse of the of the interbeat interval (IBI) divided by 60 to get the frequency in beats per minute. From the heart rate variations in the selected window, we computed the mean (meanHR) and standard deviation (stdHR) of the heart rate series. The root mean square of successive interval differences (RMSSD) is also calculated (see Equation 1). This parameter allows assessing vagal activity reflected in heart variability [40].

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(IBI_{i+1} - IBI_i)^2} \quad (1)$$

In frequency domain, the IBI series were interpolated with cubic Hermite and the power spectra were obtained by employing Welch's method [48]. The power spectral density (PSD) of a signal makes it possible to analyze its different oscillatory components such as HRV low frequency (LF) and high frequency (HF) components. The LF component is modulated by baroreflex activity and contains both sympathetic and parasympathetic activity, while the HF component reflects parasympathetic branch of the ANS [1]. The LF and HF powers of the HRV were computed as the area under the PSD curve corresponding to 0.04-0.15Hz and 0.15-0.4Hz respectively (see Figure 3). We also computed the ratio LF/HF, which represents the sympatho-vagal balance [4]. The very low frequency (VLF) components were not employed in our experiments.

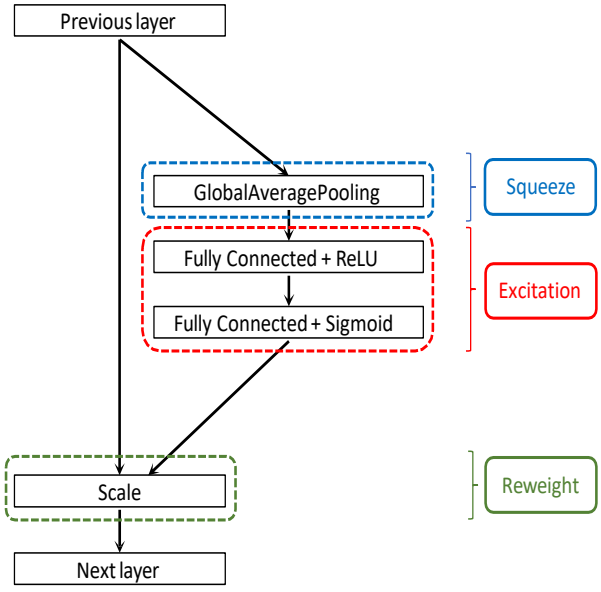



Figure 4. The Squeeze-and-Excitation module consists of global average pooling as a Squeeze operation. The two Fully Connected layers are then used to to learn the feature weights. We first reduce the feature dimension with a shrinkage parameter r, then we recover the dimension with the same r in the next fully connected layer. After the excitation operation, the SE block use the scale operation to re-weight the input layers, by element-wise multiplying the raw input by the excitation output.

### 3.4. Facial expressions recognition network

Xception network is one of the state-of-the-art methods that has proven efficient for general purpose 2D image tasks in terms of accuracy, fast convergence speed and low computational costs [7]. Xception is a derivative of Inception network [44]. It replaces Inception modules with depthwise separable convolution layers and adds residual connections. This modification, compared to Inception architecture, greatly reduces the computational cost and memory requirements, while maintaining similar (or slightly better) performance. The depth-wise separable convolution performs spatial convolution by channel separately without considering the relationship between different channels, while conventional convolution considers all spatial and channel information together. Exploiting channel dependency is an important way to improve convolutional neural network. Therefore, we fuse Xception network with Squeeze and Excitation (SE) [16] module to achieve channel weighting and maintain or improve classification ac-

curacy while reducing the number of parameters and the amount of computation. The SE block aims to explicitly model the interdependency between the channels of the image, in order to recalibrate the channel-wise feature maps in a computationally efficient manner.

The structure of the SE block is depicted in Figure 4. The SE processing blocks are composed of two successive parts: Squeeze and Excitation. The squeeze operation uses a global average pooling layer, while the excitation phase consists of two fully-connected layers that take the rectified linear units and sigmoid activation units as the hidden units respectively. In our implementation, 3D version of Xception network and SE block are used instead of the original implementations that only consider the spatial information. In this way, we simultaneously extract spatio-tempoal features without adding additional layers to take into account the temporal features.

Figure 5 presents the overall architecture of the proposed 3D-SE-XceptionNet which consists of three blocks (entry, middle and exit) as the original architecture of Xception network. However, the model structure is simplified by reducing the number of repetitive depthwise separable convolution layers. Our new mini Xception includes 15 convolution layers instead of 36 compared to the original version. These convolutional layers are structured into 14 modules, all linked with shortcuts as in ResNet architecture [15] except the first and last modules. SE blocks are inserted after the residual connections. The output of the features extraction is flattened and passed to two dense layers with 256 and 4 neurons respectively. The first dense layer takes the rectified linear units as the hidden units while the second takes the softmax activation function to predict the corresponding emotion classes.

# 4. Results and Discussion

The BP4D+ dataset was split to 90 percent training set and 10 percent validation set. Training and validation were performed three times with different samples in order to verify the consistency of the system. Three different experiments were conducted to classify emotions : using (a) facial expressions only, (b) physiological modalities only, and (c) facial expressions and physiological signals together.

## 4.1. Implementation details

The proposed system is implemented with Keras and tensorflow frameworks and ran on Nvidia Quadro P6000s. As BP4D+ is sampled at 25 fps, the length of face video clip is set to Nbframes = 100 frames (corresponding to 4 seconds) while the size of each image frame is $48 \times 48 \times 3$ ($ImHeight \times ImWidth \times Channel$). We used Rectified Adam (RAdam) optimizer [27] to optimize a categorical crossentropy loss function. We trained the network for 50 epochs with batch size = 16, learning rate $10^-4$ and decay = $10^-2$. L1 and L2 regularization strategies with coefficient equal $10^-2$ are employed which help to overcome overfitting issue and improve the model generalizability to new data.

## 4.2. Emotion recognition from facial expressions

5 state-of-the-art networks are compared : 3D-VGG [41], 3D-ResNet [15], 3D-DenseNet [17], 3D-Inception [44] and 3D-Xception [7]. We train these architectures using the BP4D+ dataset and then we compare their performance with our proposed model. As shown in Table 1, our 3D-SE-Xception network outperforms the state-of-the-art deep learning architectures. Note that in the conducted experiments, we do not perform any special preprocessing to the input images except face segmentation (See section 3.2). Compared to other architectures, the accuracy improves to the highest value of 63.40% when the Xception network is fused with the SE block. The proposed framework derives more targeted feature information through the SE module, meanwhile using the Xception network to avoid the vanishing gradient problem through residual connections and reduce the computational cost and memory requirements through the depthwise separable convolutions.

Table 1. Comparison of proposed method to state-of-the-art networks on spontaneous data for facial expression recognition.

| Method | Accuracy |
|---|---|
| 3D-DenseNet [17] | 37.91 |
| 3D-Inception [44] | 42.48 |
| 3D-ResNet [15] | 44.44 |
| 3D-VGG [41] | 49.02 |
| 3D-Xception [7] | 53.59 |
| **3D-SE-Xception (Ours)** | **63.40** |

Figure 6 shows the confusion matrix for the emotion recognition system based on facial expressions. The overall performance of the proposed network was 63.4%. Happiness and pain are the most recognized emotions with an accuracy of 80% and 81% respectively, while fear is missclassified as happiness and pain. This can partially be explained by the multiple behaviors that may occur during the expression of this emotion.

## 4.3. Emotion recognition from Physiological signals

Emotion classification from physiological signals is performed using iPPG signals and HRV features. Three different fusion schemes were conducted for emotion recognition

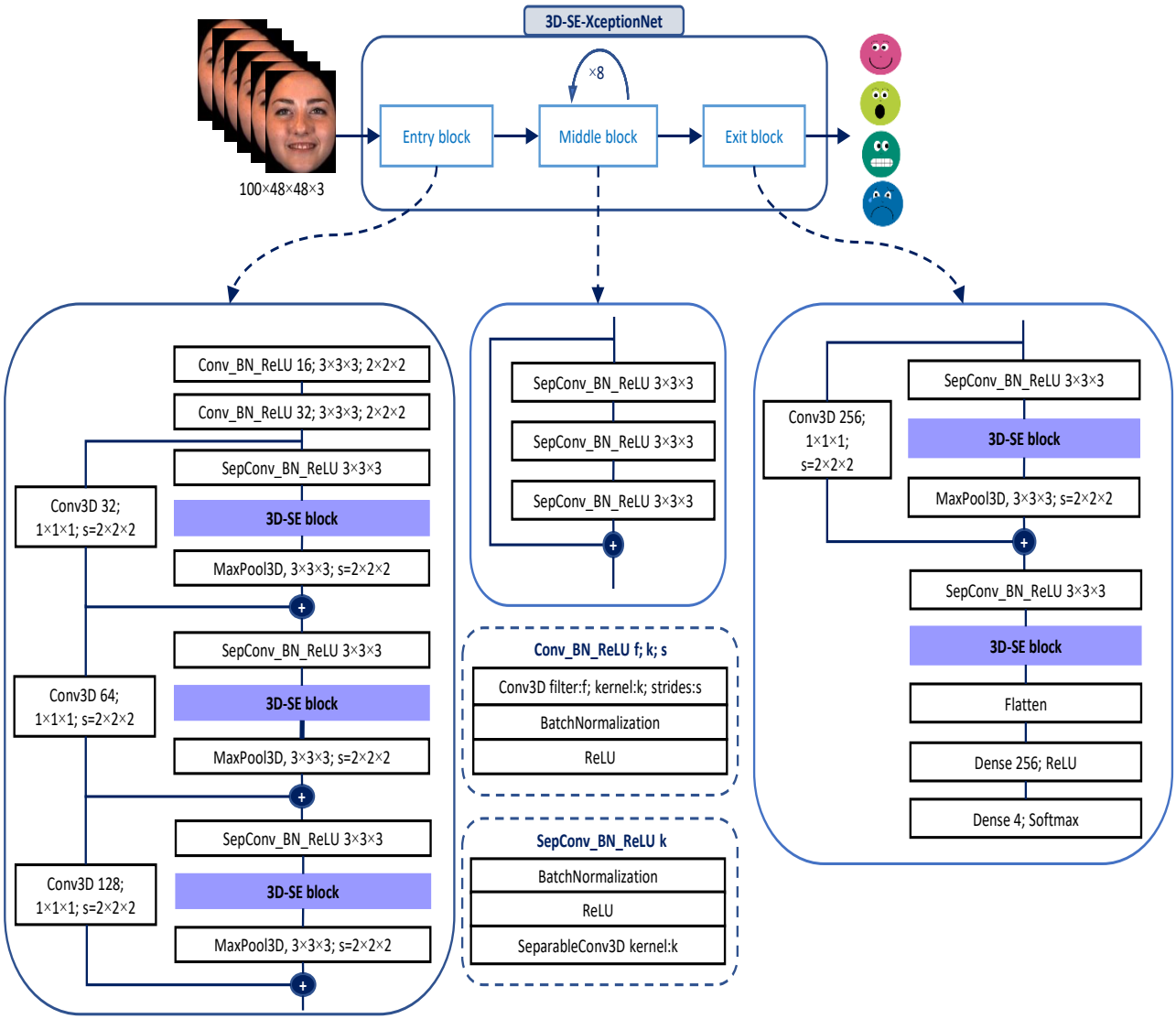



Figure 5. The network structure of 3D-SE-Xception corresponds to a modified version of the Xception network. 2D depthwise separable convolution layers are replaced by 3D depthwise separable convolution to capture both spatial and temporal features across video frames. The SE block was embedded in the model to enhance the useful feature channels and weaken the useless feature channels through channel-wise feature maps recalibration. Two dense layers are used instead of Global Average Pooling.The input video fragment first goes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow which ends in a dense layer with 4 neuron to classify emotions.

using physiological data. First, iPPG signals and HRV features are used separately to classify emotions. Then, we merge them to see which approach gives the best accuracy.

Inspired by the work of Fabiano et al. [13], a feedforward neural network is used in our experiments. It consists of two layers. The input layer has the same number of neurons as the input length (100 for iPPG modality , 6 for HRV), while the output layer includes the same number of neurons as the number of classes of emotion to predict. The activation function for the input layer is ReLU, while the softmax activation function is employed for the output layer.

Table 2 illustrates the recognition accuracy using iPPG signals and HRV features separately and after fusion between them. As can be seen from table 2, whether physiological signals are used separately or combined, the recognition accuracy is low compared to facial expressions. Besides, the performance when using iPPG signals is better than HRV. This can be justified by the short length of the

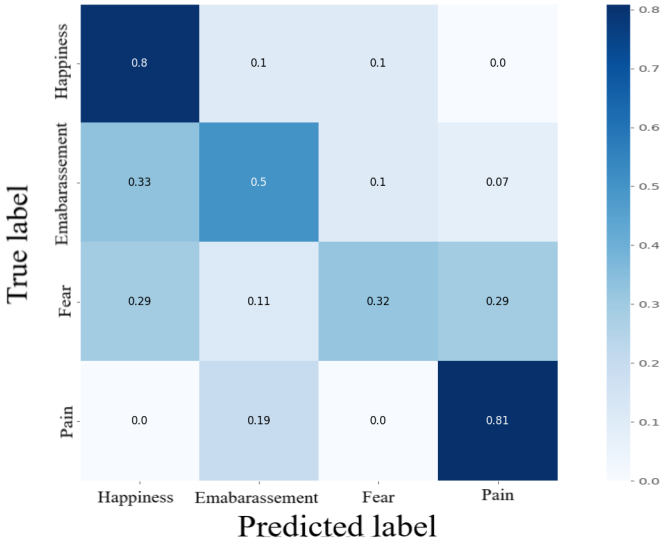



Figure 6. Emotion classification confusion matrix using facial expressions.

iPPG signals used for HRV analysis as well as the signal quality which is prone to noise and artifacts due to movements and lighting conditions. Therefore, it has an impact on the accuracy of HRV characteristics. On the other hand, iPPG and HRV fusion exhibit lower performance. This may be related to the lack of correlation between the iPPG signal and the HRV characteristics.

Table 2. Comparison of emotion recognition accuracy from physiological signals. Abbreviations: (iPPG :Emotion from iPPG signals), (HRV : Emotions from HRV features), (iPPG + HRV : Emotions from the combined iPPG and HRV).

| Method | Accuracy |
|---|---|
| iPPG | 55.33 |
| HRV | 53.59 |
| iPPG + HRV | 44.64 |

### 4.4. Multimodal emotion recognition

The architecture of our multimodal emotion recognition system is shown in Figure 1. Basically, the proposed model consists of two pipelines allowing to extract the features of each modality from video streams (See section 3.3 and 3.4). Each video of BP4D+ is fed to the facial expression network (3D-SE-Xception) and to the iPPG signal network (MTTS-CAN). The first pipeline extracts the features vector after the flatten layer (See Figure 1 using the pre-trained weights of our 3D-SE-Xception model, while the second pipeline returns either the iPPG signal recovered through the MTTS-CAN network or HRV features. Hence, three experiments have been carried out for multimodal emotion recognition. First, facial expressions features are combined with only the iPPG signal, then only with HRV vector. Finally, all modalities are fused. The contactenation result vector is then passed to two dense layers with 256 and 4 neurons respectively. The first dense layer takes the rectified linear units as the hidden units while the second takes the softmax activation function to predict the corresponding emotion class.

The recognition accuracy for each experiment is reported in Table 3. The results show that combining facial expression features with physiological parameters improve the performance compared to unimodal approach either using facial expressions or physiological data separately. This confirms previous studies that have obtained the same results where the precision of the fusion exceeds unimodality systems, and the performance of facial expressions modality is always better compared to physiological signals [8, 18]. Furthermore, the lack of correlation between the iPPG signal and HRV features impacts performance, whether merging just these two modalities or their fusion with facial expressions.

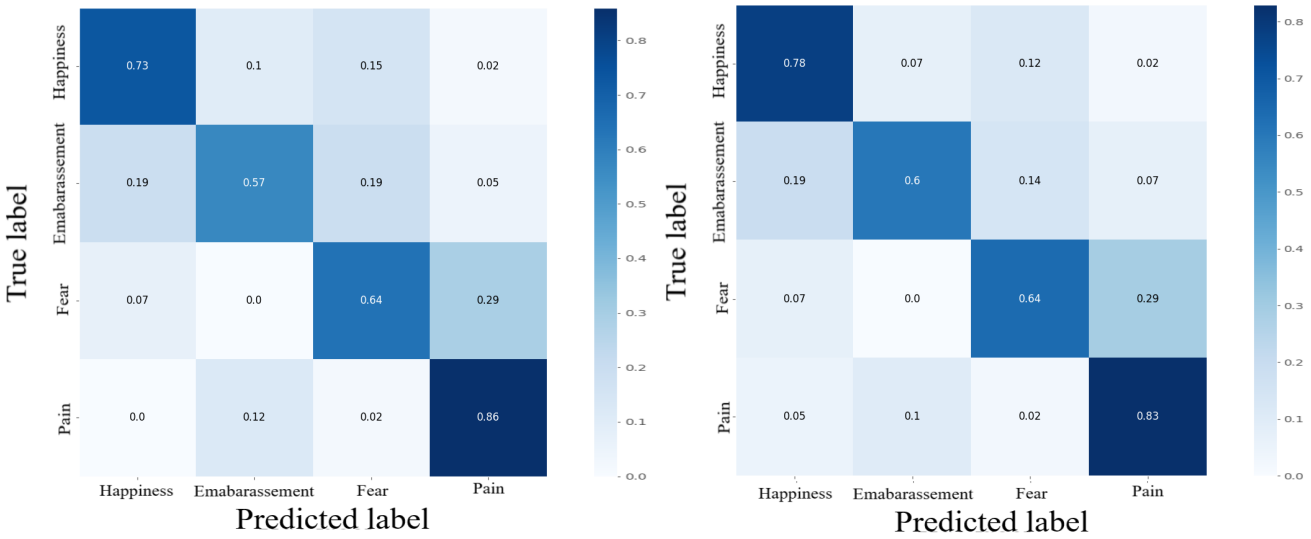
Table 3. Comparison of multimodal emotion recognition accuracy from facial expressions and physiological signals.

| Method | Accuracy |
|---|---|
| Facial expressions + HRV | 70.59 |
| Facial expressions + iPPG | 71.90 |
| Facial expressions + iPPG + HRV | 67.97 |

Figure 7 shows the confusion matrix for the multimodal emotion recognition system based on facial expressions and HRV features fusion, and facial expressions and iPPG fusion. The overall performance of the proposed network is 70.59% and 71.90 % respectively. Compared to using facial expressions only, the fusion with physiological signals improved significantly the accuracy for missclassified emotions. For example, fear accuracy has been doubled from 32% to 64% for each fusion schemes.

### 4.5. Discussion

Facial expressions and physiological signals modalities establish superiority to each other. The combination of facial expression features and iPPG signal achieved the highest accuracy of around 72%. This slightly outperforms the fusion between facial expressions and HRV features. However, merging only the iPPG signal and HRV features, or




(a) Facial expressions and HRV features fusion.

(b) Facial expressions and iPPG signals fusion.

Figure 7. Multimodal emotion classification confusion matrix using facial expressions with HRV Features and iPPG signals.

with facial expressions features, gives the lowest accuracy. We hypothesize that these two modalities may interfere with each other, thus impacting the recognition accuracy. In addition, using multiple modalities considerably improved performance for miss-classified emotions such as Fear. Although facial expressions are visible and easy to categorize compared to physiological cues, incorporating with physiological modalities can provide complementary information and further enhance the performance. On the other hand, the results obtained fit perfectly with existing multimodal systems that use multiple input data sources and demonstrate the possibility of using only facial videos to recognize emotions using human physiological and physical cues.

## 5. Conclusion

This paper proposes a new framework for multimodal emotion recognition through facial expressions and physiological signals. A novel spatiotemporal neural network has been proposed, which fused Squeeze-and-Excitation modules with a 3D Xception network to recalibrate the channel-wise feature maps in a computationally efficient manner. Two physiological parameters were selected, namely the iPPG signal and the HRV features. Unlike existing studies, physiological cues were measured remotely based on imaging photoplethysmography method. This way, only single input source were used to extract features from each modality. It is very interesting and promising to recognize emotions in a multimodal way with a single non-intrusive sensor. using a camera that is integrated on all digital devices used in daily life allows to reduce the cost and to make the system more accessible. Furthermore, video-based physiological signals measurement is more practical and may reduces the discomfort caused by the contact devices. Overall, we have shown that fusion of two modalities (facial expressions with iPPG signals or facial expressions with HRV features) gives significant improvements and offer potential for more accurate recognition of affects and emotions.

As future work, we intend to incorporate other physiological signals and test the performance on other multimodal emotion datasets. We will further explore the complexity of expressions to understand the poor performance of certain emotions.

## References


[1] Bradley M. Appelhans and Linda Luecken. Heart rate variability as an index of regulated emotional responding. *Review of General Psychology*, 10(3):229–240, Sept. 2006. 4

[2] Yannick Benezeth, Peixi Li, Richard Macwan, Keisuke Nakamura, Randy Gomez, and Fan Yang. Remote heart rate variability for emotional state monitoring. In *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 153–156, 2018. 1

[3] Yannick Benezeth, Peixi Li, Richard Macwan, Keisuke Nakamura, Randy Gomez, and Fan Yang. Remote heart rate variability for emotional state monitoring. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 153–156. IEEE, 2018. 2, 3

[4] Robert L. Burr. Interpretation of normalized spectral heart rate variability indices in sleep research: a critical review. *Sleep*, 30 7:913–9, 2007. 4

[5] George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. In *IFIP Interna-*

*tional Conference on Artificial Intelligence Applications and Innovations*, pages 375–388. Springer, 2007. 2

[6] A V J Challoner and C A Ramsay. A photoelectric plethysmograph for the measurement of cutaneous blood flow. *Physics in Medicine and Biology*, 19(3):317–328, may 1974. 2, 3

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017. 4, 5

[8] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access*, 8:168865–168878, 2020. 1, 2, 7

[9] Djamaleddine Djeldjli, Frédéric Bousefsaf, Choubeila Maaoui, and Fethi Bereksi-Reguig. Imaging photoplethysmography: Signal waveform analysis. In *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 2, pages 830–834, 2019. 1

[10] Andrius Dzedzickis, Arturas Kaklauskas, and Vytautas Buinskas. Human emotion recognition: Review of sensors and methods. *Sensors (Basel, Switzerland)*, 20, 2020. 2

[11] Sidney K. D'Mello and Jacqueline Kory Westlund. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47:1 – 36, 2015. 2

[12] Maria Egger, Matthias Ley, and Sten Hanke. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, 343:35–55, 2019. The proceedings of AmI, the 2018 European Conference on Ambient Intelligence. 1, 2

[13] Diego Fabiano and Shaun Canavan. Emotion recognition using fused physiological signals. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 42–48, 2019. 6

[14] Miha Fingar and Primoz Podrzaj. Feasibility of assessing ultra-short-term pulse rate variability from video recordings. *PeerJ*, 8, 2020. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[18] 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. 1, 2, 7

[19] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. 3

[20] Dimitrios Kollias and Stefanos Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, 12(3):595–606, 2020. 1

[21] Jukka Kortelainen, Suvi Tiinanen, Xiaohua Huang, Xiaobai Li, Seppo Laukka, Matti Pietikäinen, and Tapio Seppänen. Multimodal emotion recognition by combining physiological signals and facial expressions: A preliminary study. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5238–5241, 2012. 2

[22] Jyoti Kumari, R. Rajesh, and K.M. Pooja. Facial expression recognition: A survey. *Procedia Computer Science*, 58:486–491, 2015. Second International Symposium on Computer Vision and the Internet (VisionNet'15). 2

[23] James Ren Lee, Linda Wang, and Alexander Wong. Emotionnet nano: An efficient deep convolutional neural network design for real-time facial expression recognition. *Frontiers in Artificial Intelligence*, 3, 2021. 2

[24] Min Seop Lee, Yun Kyu Lee, Dong Sung Pae, Myo Taeg Lim, Dong Won Kim, and Tae-Koo Kang. Fast emotion recognition based on single pulse ppg signal with convolutional neural network. *Applied Sciences*, 2019. 1

[25] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020. 2

[26] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2583–2596, 2018. 3

[27] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, April 2020. 5

[28] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014. 2

[29] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020. 3, 4

[30] Timur Luguev, Dominik Seuß, and Jens-Uwe Garbas. Deep learning based affective sensing with remote photoplethysmography. In *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–4, 2020. 1, 2, 3

[31] Shervin Minaee, Mehdi Minaei, and Amirali Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9):3046, 2021. 1

[32] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 1

[33] Nazmun Nahid, Arafat Rahman, and Md Atiqur Rahman Ahad. Contactless human emotion analysis across different modalities. 2021. 2

[34] Aoxin Ni, Arian Azarang, and Nasser Kehtarnavaz. A review of deep learning-based contactless heart rate measurement methods. *Sensors*, 21:3719, 05 2021. 3

[35] Yuval Nirkin, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gérard G. Medioni. On face segmentation, face swapping, and face perception. *CoRR*, abs/1704.06729, 2017. 3

[36] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 12(2):505–523, 2018. 2

[37] Yassine Ouzar, Djamaleddine Djeldjli, Frédéric Bousefsaf, and Choubeila Maaoui. Lcoms lab's approach to the vision for vitals (v4v) challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2750–2754, 2021. 3

[38] Aasim Raheel, Muhammad Majid, Majdi R. Alnowami, and Syed Muhammad Anwar. Physiological sensors based emotion recognition while experiencing tactile enhanced multimedia. *Sensors (Basel, Switzerland)*, 20, 2020. 1

[39] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. 2

[40] Fred Shaffer and Jay P Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, page 258, 2017. 4

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[42] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. 3

[43] Bo Sun, Liandong Li, Guoyan Zhou, and Jun He. Facial expression recognition in the wild based on multimodal texture features. *Journal of Electronic Imaging*, 25(6):1 – 8, 2016. 1

[44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 4, 5

[45] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz J. Rak. Emotion recognition using facial expressions. *Procedia Computer Science*, 108:1175–1184, 2017. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland. 2

[46] Sergey Tulyakov, Xavier Alameda-Pineda, Elisa Ricci, Lijun Yin, Jeffrey F Cohn, and Nicu Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2396–2404, 2016. 3

[47] Kannan Venkataramanan and Haresh Rengaraj Rajamohan. Emotion recognition from speech. *arXiv preprint arXiv:1912.10458*, 2019. 2

[48] Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967. 4

[49] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018. 1

[50] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[51] Yi Yang, Qiang Gao, Xiaolin Song, Yu Song, Zemin Mao, and Junjie Liu. Facial expression and eeg fusion for investigating continuous emotions of deaf subjects. *IEEE Sensors Journal*, 21(15):16894–16903, 2021. 1

[52] Sebastian Zaunseder, Alexander Trumpp, Daniel Wedekind, and Hagen Malberg. Cardiovascular assessment by imaging photoplethysmography – a review. *Biomedical Engineering / Biomedizinische Technik*, 63:617 – 634, 2018. 3

[53] Zheng Zhang, J. Girard, Yue Wu, X. Zhang, Peng Liu, U. A. Ciftci, S. Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, J. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3438–3446, 2016. 3

[54] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018. 1

[55] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *European conference on computer vision*, pages 425–442. Springer, 2016. 1