IEICE TRANS. INF. & SYST., VOL.E104–D, NO.12 DECEMBER 2021

PAPER

Multimodal-Based Stream Integrated Neural Networks for Pain Assessment

Ruicong ZHI[†,††a)], **Caixia ZHOU**[†,††], **Junwei YU**[†,††], ***Nonmembers***, **Tingting LI**[†,††], ***Member***, ***and*** **Ghada ZAMZMI**[†††], ***Nonmember***

SUMMARY Pain is an essential physiological phenomenon of human beings. Accurate assessment of pain is important to develop proper treatment. Although self-report method is the gold standard in pain assessment, it is not applicable to individuals with communicative impairment. Non-verbal pain indicators such as pain related facial expressions and changes in physiological parameters could provide valuable insights for pain assessment. In this paper, we propose a multimodal-based Stream Integrated Neural Network with Different Frame Rates (SINN) that combines facial expression and biomedical signals for automatic pain assessment. The main contributions of this research are threefold. (1) There are four-stream inputs of the SINN for facial expression feature extraction. The variant facial features are integrated with biomedical features, and the joint features are utilized for pain assessment. (2) The dynamic facial features are learned in both implicit and explicit manners to better represent the facial changes that occur during pain experience. (3) Multiple modalities are utilized to identify various pain states, including facial expression and biomedical signals. The experiments are conducted on publicly available pain datasets, and the performance is compared with several deep learning models. The experimental results illustrate the superiority of the proposed model, and it achieves the highest accuracy of 68.2%, which is up to 5% higher than the basic deep learning models on pain assessment with binary classification.

key words: *multi-modality, pain assessment, dynamic facial feature, biomedical feature, stream integrated neural networks*

1. Introduction

Physical pain is a complex and subjective experience that is often caused by noxious stimuli damaging tissue. It can be defined as a protective mechanism that alerts us about the damage that is occurring or potentially occurring [1]. Accurate pain assessment is vital for understanding patients' medical conditions and developing suitable treatments. Globally, self-report method is considered as the gold standard in pain assessment, and has been applied successfully in pain management. However, self-reporting results would provide inconsistent and unreliable information in cases where an individual is suffering from a form of cognitive impairment [34]. Moreover, the self-report manner is not applicable to individuals with communicative impairment.

The observational measures based on behavioral indicators and physiological indicators could be considered as an effective way for pain assessment. The pain related behavioral indicators include facial expression, body movement, and sound signals. Facial expression is the most specific pain behavioral indicator, which is more salient and consistent than other behavioral indicators [3]–[5]. Painful facial expressions are defined as the movement and distortion of facial muscles associated with painful stimuli, which can be described by the action units (AUs). Prkachin and Solomon [2] found that four AUs on faces – brow lowering, orbital tightening, levator contraction, and eye closure – carried the bulk of information about pain. In addition, biomedical signals are effective objective measurements which provide a lot of clues for pain assessment.

In practice, the evaluation of indicator-based manner requires professional caregivers with plenty of training, and the manual pain assessment is time-consuming and laboring for long-term continuous pain monitoring. Therefore, an automatic multi-modal based pain assessment system is desired to objective pain assessment. Various studies have investigated the feasibility and relevancy of automatic pain assessment systems based on measurable visual and physiological parameters (see Sect. 2.4). These studies show that non-verbal pain indicators such as pain related facial expressions and changes in physiological parameters could provide valuable insights for pain detection and intensity estimation.

In this paper, we propose a multimodal-based Stream Integrated Neural Network with Different Frame Rates (SINN) for automatic pain assessment. The proposed network combines facial expression and biomedical signals. The main contributions can be summarized as follows:

(1) We propose four-stream inputs to SINN for facial expression feature extraction. These inputs include the spatial information (original image sequences), the temporal information (optical flow sequences), the static information (slow pathway), and the dynamic information (fast pathway). These variant facial features are then integrated with biomedical features, and the joint features are utilized for pain assessment.

(2) We propose to learn the dynamic facial features in both implicit and explicit manners to better represent the facial changes. The optical flow ConvNet3D deals with facial image sequences and image sequences with different frame

Manuscript received March 25, 2021.
Manuscript revised June 18, 2021.
Manuscript publicized September 10, 2021.

[†]The authors are with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, P.R. China.
[††]The authors are with Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, PR China, 100083.
[†††]The author is with the Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620, USA.
a) E-mail: zhirc_research@126.com

DOI: 10.1587/transinf.2021EDP7065

rates are employed to learn dynamic facial features explicitly and implicitly.

(3) We propose to integrate facial expression, skin conductance level, electrocardiogram, and electrical muscle activity, to identify various pain states. Our experimental results prove that the multimodal scheme greatly enhances the performance of pain assessment.

2. Related Work

Automatic pain assessment is a difficult task due to the variances between different subjects. In this section, we introduce the state-of-the-art automated pain assessment methods for facial expression, body movement, and biomedical signals.

2.1 Facial Expression

Facial expression is the most well-applied indicator for pain assessment in practice as it is non-invasive and easily acquired by video recording techniques [6]–[10]. Extracting the best set of facial features is critical to obtain accurate pain assessment. Furthermore, excellent facial features can reduce the dependence on the selection of classifiers. Existing works for pain assessment based on facial expression analysis can be divided into two categories: frame-level feature extraction and sequence-level feature extraction. For frame-level facial expression feature extraction, researchers tried both appearance features and geometry features for facial representation. Appearance features focus on capturing information that represents facial texture, which reflects the magnitude and direction of facial surface displacement. Examples of the texture descriptors that have been applied successfully to pain assessment include Local Binary Pattern (LBP) [11]–[13], Gabor Transform [14], and Discrete Cosine Transform (DCT) [15]. Geometry-based features describe the changes in the facial geometry using a set of fiducial points or a connected face mesh, such as the Active Appearance Model (AAM) [16], [17]. The main limitation of frame-level feature extraction is that they deal with static images and ignore the dynamic pattern of pain.

In the last decades, several researchers reported the importance of using temporal facial representation, since pain is a dynamic event and it evolves in a specific pattern over time. For example, Multiple Instance Learning (MIL) [18], [19] was used to create instance labels inside the bag, which benefited from weakly supervised pain intensity estimation tasks. Other works, which used Three Orthogonal Planes (TOP) [20], [21], were proposed to extract spatio-temporal information for pain detection, and they were well applied to several texture descriptors such as LBP and histogram of oriented gradients (HOG). Werner et al. [22] proposed a novel feature set, called facial activity descriptors, to describe facial actions for pain detection and pain intensity estimation. Bourou et al. [23] calculated several distances (e.g., mean and median) from ROIs to classify pain expressions.

Recently, self-learning features, extracted by deep learning algorithms, are exploited in automatic pain assessment to extract facial representations through a joint feature learning and classification/regression pipeline. For instance, Kharghanian et al. [24] utilized Convolutional Deep Belief Network (CDBN) to extract facial features in an unsupervised manner. Convolutional Neural Networks (CNN) were utilized in [25]–[27] to learn facial features for pain recognition because of their powerful feature learning ability. For example, Egede et al. [26] combined three kinds of facial features, including deep learning features for regions of interest, geometric features, and texture features. To integrate temporal pain analysis, Long Short-Term Memory (LSTM) [25] has been commonly used for temporal feature extraction. For example, Bellantonio et al. [27] employed a combination of CNN and Recurrent Neural Networks (RNN) to set up a deep hybrid pain detection framework to analyze both spatial and temporal pain information from videos of faces. Similarly, a bidirectional LSTM-RNN was used to automatically estimate Prkachin and Solomon Pain Intensity (PSPI) levels from face images. Moreover, Zhou et al. [28] designed a real-time regression framework based on recurrent CNN for automatic frame-level continuous pain intensity estimation.

2.2 Body Movement

Body motion is an effective indicator of pain, especially for patients with chronic diseases and infants. Walsh et al. [29] found that pain was communicated through averted head and trunk, hand touches to various sites, knee bending, and shoulder to front movements. Olugbade et al. [30] explored pain-body movements by conducting experiments to discriminate subjects with low-level and high-level pain. Wang et al. [31] utilized LSTM to detect events of protective behavior captured from healthy people and people with chronic back pain.

2.3 Biomedical Signals

Pain could cause changes in biomedical signals such as Electromyography (EMG), Electrocardiogram (ECG), and skin conductance level (SCL) signals. For example, Kachele et al. [32] extracted EMG, ECG, and SCL signals and used them to detect different levels of pain. Walter et al. [33] obtained the amplitude and change of galvanic skin reaction (GSR), EMG, ECG, and SCL. In [34], several parameters were extracted temporally from biomedical signals. Although the overall performance was significantly improved, the authors found that pain assessment based on biomedical signals was not performed well for low pain intensities.

2.4 Fusion of Multiple Pain Signals

Pain causes both behavioral and physiological changes, multimodal-based pain assessment attracted increasing attention [35]. Walter et al. [36] combined both facial

expressions and biomedical channels to obtain pain-related information. Haque et al. [37] used early feature fusion by generating a new input that merged RGB, depth, and thermal images into one matrix, and the outputs of several classifiers were combined through late fusion. In the case of deep learning, Thiam et al. [38] proposed a deep learning-based method that merged several kinds of information, such as speech information, geometry descriptor, head pose, and biomedical signals, to detect different pain states. These studies showed that both the behavioral and biomedical signals were critical for pain assessment and the superiority of multimodal-based pain recognition.

3. Proposed Method

This paper presents a multimodal-based Stream Integrated Neural Network with Different Frame Rates (SINN). SINN utilizes facial expression and biomedical signals. For facial expressions, a four-stream structure is used to extract four kinds of facial features. As for the biomedical signals, the LSTM structure is utilized and then the biomedical channel is merged with the four-stream structure of a facial expression to generate the final SINN structure. Figure 1 depicts the flowchart of our proposed SINN network.

As shown in the Fig. 1, the original video sequence is normalized by interpolation to a 9-frame sequence to form the high frame rates sample, and a 4-frame sequence to form the low frame rates sample. Then the optical flow stream information is extracted for both the low frame rate image sequence and high frame rate image sequence, respectively, to generate a four-stream input of facial expression data. These inputs are convolved to extract facial features through the ResNet3D module, then merge with the biomedical signal feature extracted by LSTM. Finally, Softmax is used to output the probability of each pain level. The proposed method is described in detail next.

3.1 Preprocessing

All the facial images of the sequences are preprocessed by face alignment and face frontalization to get the exact face region and eliminate noise. We use Procrustes Analysis [39] to perform face alignment, and face frontalization was referred to [40].

3.2 SINN for Facial Expression Analysis

3.2.1 3DConvNet

The Two-dimensional Convolutional Neural Network (2DCNN) is one of the most successfully applied deep learning methods that are used to learn the image texture information effectively. However, it is far not enough to lean facial features by 2D convolution from spatial dimensions when applied to video analysis tasks. Facial image sequences contain plenty of spatial and temporal information that is very helpful for identifying different pain states. In this paper, we use 3DCovNet for feature learning so that it can deal with the image sequences conveniently. Moreover, an optical scheme of squeezing is utilized to decrease the complexity of 3DConvNet.

3DConvNet works by convolving a 3D kernel to the cube formed by stacking multiple frames together. The multiple adjacent frames are connected to form the input and the feature maps of 3DCNN. The formula of 3DConvNet is expressed as:

$$x_{3d}^{l} = \sigma\left(z^{l}\right) = \sigma\left(x_{3d}^{l-1} * W_{3d}^{l} + b^{l}\right) \tag{1}$$

where x_3d is a four-dimension array. This four-dimension array is [num_of_frames, width, height, channel] for inputs, and [first dimension of feature maps, width, height, num_of_feature_maps] for feature maps. W_3d is the parameters of the 3D convolutional kernel, and b is the biases

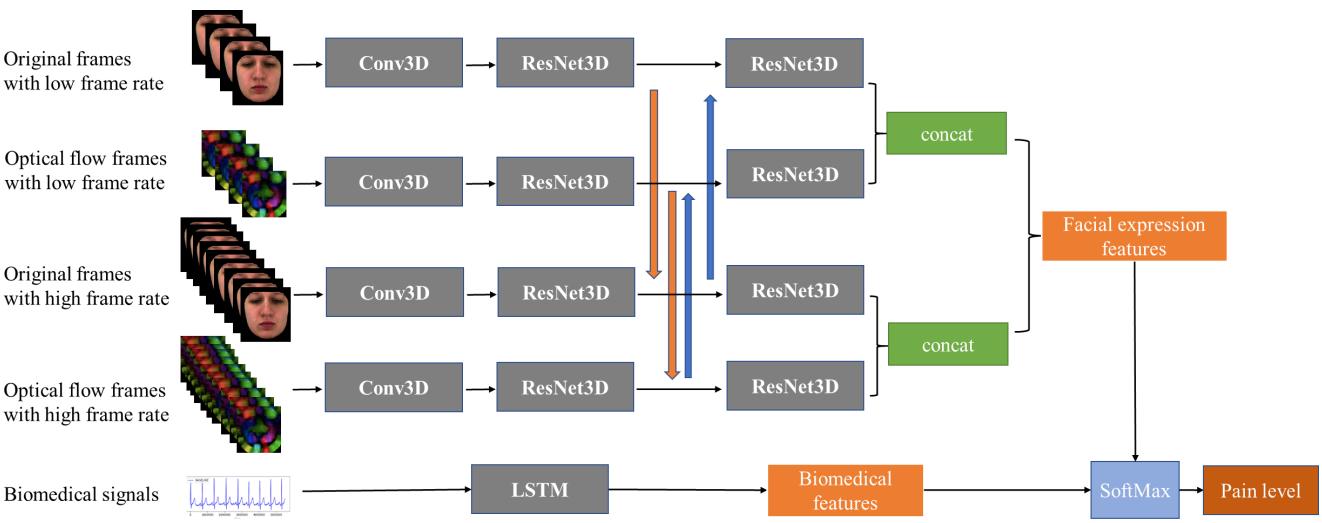


Fig. 1 The skeleton of the proposed SINN. The structure of ResNet3D can be seen in Fig. 3 (b).

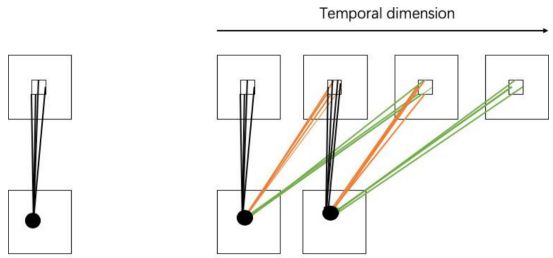


Fig. 2 Comparison of 2D and 3D convolution operation (a) 2D convolution operation (b) 3D convolution operation.

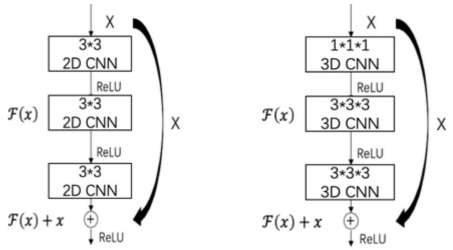


Fig. 3 Shortcut connections (a) original ResNet (b) proposed ResNet3D.

parameter.

The 3DConvNet can extract both spatial information and motion information. Figure 2 (a) and Fig. 2 (b) present the 2D convolutional operation and 3D convolutional operation, respectively.

In this paper, we use 3DCovNet for feature learning so that it can deal with the image sequences conveniently. Moreover, an optical scheme of squeezing is utilized to decrease the complexity of 3DConvNet.

Although 3DConvNet can extract both spatial and temporal information with a single convolutional kernel, the network has a significantly high computational complexity. Further, this network is deep with a high number of layers to enhance performance. When deeper networks start converging, a degradation problem occurs. This degradation problem could be somehow addressed by the deep residual neural networks.

The stacked layers are expected to fit a residual mapping, instead of a direct desired underlying mapping. Formally, the desired underlying mapping is denoted as $\mathcal{H}(x)$, x is the input of the Resnet block. Let the stacked nonlinear layers fit another mapping of $\mathcal{F}(x) := \mathcal{H}(x) - x$, and then the original mapping is recast into $\mathcal{F}(x) + x$. The hypotheses are that it is easier to optimize the residual mapping than the original unreferenced mapping. To the extreme, if an identity mapping is optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers. The formulation of $\mathcal{F}(x) + x$ is realized by feedforward neural networks with "shortcut connections."

We change the bottleneck layer in the ResNet to make it more suitable for our task. First, the two-dimensional convolution in the ResNet structure is replaced by the three-dimensional convolution to deal with the dynamic feature extraction for facial expression image sequences. Second, the idea of SqueezeNet [41] is integrated into the ResNet architecture. The feature map is reduced by $1 \times 1 \times 1$ convolution kernel, and then $3 \times 3 \times 3$ convolution is performed, so that the network parameters can be compressed. The structure of the proposed ResNet3D is shown in Fig. 3.

3.2.2 Optical Flow 3DConvNet

As the 3DConvNet can't estimate the motion implicitly, we use the optical flow to estimate the motion between video frames explicitly [42]. The optical flow-based 3DConvNet for motion information extraction is implemented through a scheme of feeding the 3DConvNet by stacking optical flow displacement fields between several consecutive facial frames.

Optical flow is the instantaneous velocity of the moving motion of a spatially moving object on an imaging plane. The optical flow method explores the change of pixels in the time domain and the correlation between adjacent frames to find the correspondence between the previous frame and the current frame. The computation of the optical flow vector from an image sequence requires constant brightness, continuous-time, and small motion constraints.

Consider the brightness of a pixel I(x, y, t) in the first frame, where t represents the time dimension, x and y represent the spatial coordinates in which it is located. It moves the distance of (dx, dy) to the next frame by dt time. According to the first constraint of optical flow, the brightness of the pixel before and after the motion is constant, i.e.

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{2}$$

The right-side expression can be rewritten by Taylor expansion as follows:

$$I(x, y, t) = I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \varepsilon \tag{3}$$

where ε represents the second-order infinitesimal term and can be neglect. Merge Eq. (3) and Eq. (4), and divide by dt for both side:

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t}\frac{dt}{dt} = 0 \tag{4}$$

Let u, v be the velocity vectors of the optical flow along X and Y axes, respectively:

$$u = \frac{dx}{dt}, \; v = \frac{dy}{dt} \tag{5}$$

The velocity vector of the pixel is calculated using the spatio-temporal differentiation of the time-varying image grayscale.

The optical flow means that each pixel in the image has a displacement in the x-direction and y-direction, so the

size of the corresponding optical flow image is the same as that of the original image. It could be displayed with the Munsell color system as shown in Fig. 4. Figure 4 (c) shows the optical flow image we obtained from the facial images (Fig. 4 (a) and Fig. 4 (b)).

We calculate the optical flow images by the adjacent images of original facial image sequences. The optical flow image sequences, together with the original pain facial expression image sequences, are fed to the neural network for further facial feature learning to obtain temporal information for classification.

3.2.3 Stream Integration for Facial Expression of Pain

Our strategy for combining the four-stream inputs is shown in Fig. 5. Two-stream image sequences with different frame rates are designed to process the facial frames individually. Low frame rates are fed to the network to analyze static parts of facial frames, while image sequences with high frame rates are fed to the network to process dynamic parts of facial frames. Our idea is partly inspired by the retinal ganglion of primates [43], [44]. Also, both the high frame rate pathway and slow frame rate pathway can interact bidirectionally. A lateral connection scheme is utilized to fuse the information of two pathways (low frame rate and high frame rate), which is a popular technique for merging different levels of spatial resolution and semantics [45] and has been used in two-stream networks [46]. The fast and slow pathways are connected to make facial features rich. The feature map sizes of the high frame rate stream and low frame rate stream are the same after the ResNet3D.

In this paper, we firstly use the multi-stream integrated neural network for detecting facial expressions of pain. As shown in Fig. 5, two kinds of dynamic information extraction manners are utilized to represent the dynamic facial features of pain states. The network has four-input streams: original image sequences of low frame rate by 3DConvNet (input stream 1), optical flow image sequences of low frame rate by optical flow 3DConvNet (input stream 2), original image sequences of high frame rate by 3DConvNet (input stream 3), and optical flow image sequences of high frame rate by optical flow 3DConvNet (input stream 4). The facial features extracted by each stream are fused and fed to Softmax for pain recognition.

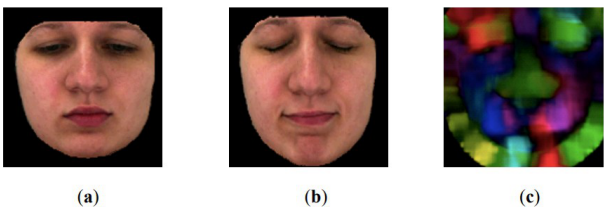


Fig. 4 Example of optical flow image for pain facial expression (a) and (b) are original images, and (c) is the corresponding optical flow image.

3.3 Multimodal-Based SINN

Several studies reported that biomedical signals were good indicators of pain. In this work, SCL (Skin conductance level), ECG (electrocardiogram), and EMG (electrical muscle activity) are utilized and fused with the facial information extracted by SINN.

The EMG signal is filtered by a Butterworth bandpass filter (20–250Hz). The filtered signal is further denoised by Andrade [7], which is based on Empirical Mode Decomposition. The ECG signal is filtered with a Butterworth bandpass filter (0.1–250Hz) [38]. The biomedical signals are recorded together with the videos of the face simultaneously. The time window of the biomedical signals and face videos are the same.

Examples of those biomedical signals for baseline (BL) and pain intensity 4 (PA4) are shown in Fig. 6. The x-axis is time and the y-axis represents the amplitude of signals. The figure shows that the biomedical signal is a vector concatenated based on the time window. Hence, LSTM is utilized to extract features of the biomedical signals. After that, the

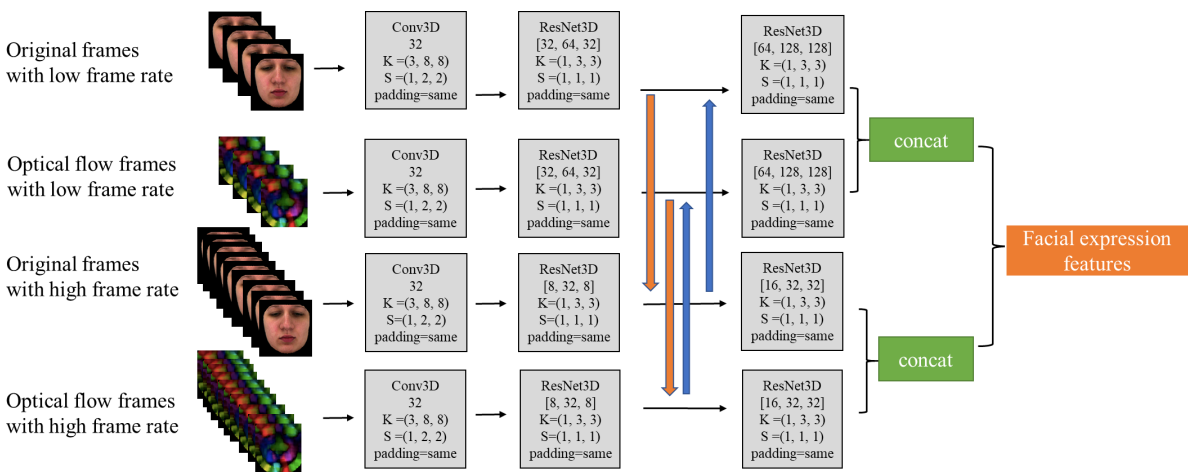
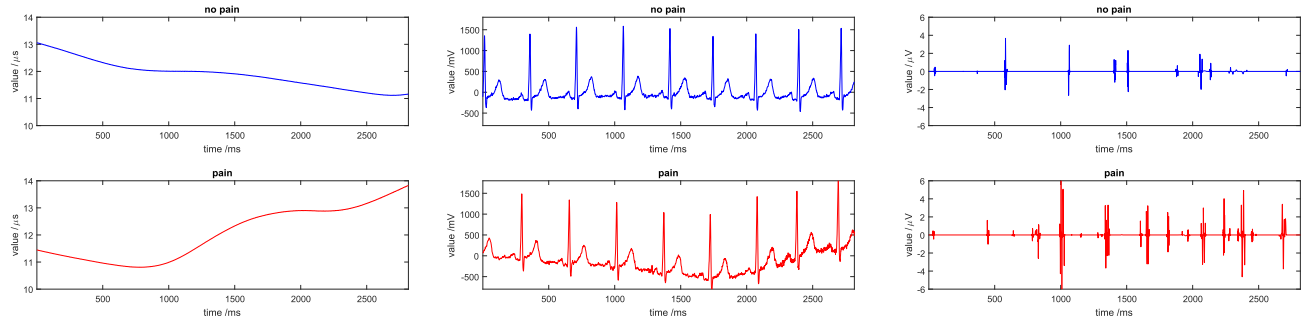


Fig. 5 The facial feature learning structure in SINN network. K means kernel size and S means strides. The number in [ ] means the channel.



(a) SCL (b) ECG (c) EMG

Fig. 6 Biomedical signal comparison (a) SCL (b) ECG (c) EMG. Blue line denotes the signal of baseline (BL) and red line denotes the signal of pain intensity 4 (PA4).

extracted feature vector is fed to a fully connected layer for dimensionality reduction, then merged with the facial expression features. The final feature vector, which contains both the facial and biomedical features, is then fed to two Softmax layers. Finally, the probability of the pain category is obtained.

In summary, the proposed SINN receives five different inputs, including four streams for facial expression and one stream for biomedical signals. In the following section, we use the proposed multimodal-based SINN for automatic pain assessment, and evaluate the performance on two publicly available pain datasets, i.e. BioVid and MIntPAIN. We perform two levels of pain classification: binary classification and multi-level classification. We also investigate subject-independent and subject-dependent pain assessment by using different cross-validation schemes.

4. Experiments and Discussion

4.1 Dataset and Evaluation Protocol

We evaluate the proposed SINN network on the BioVid and MIntPAIN pain datasets. BioVid datase (http://www.iikt.ovgu.de/BioVid.html) [5] has 8700 facial videos recorded from eighty-seven participants. Three kinds of biomedical signals are collected, i.e., SCL, EMG, and ECG. EEG signals are not collected to prevent the occlusion of facial expressions by EEG acquisition devices. There are five kinds of stimulus intensity labels, i.e., Baseline (BL), pain intensity 1 (PA1), pain intensity 2 (PA2), pain intensity 3 (PA3), and pain intensity 4 (PA4). To enlarge this dataset, we augment the videos ten times by random cropping.

MIntPAIN dataset (https://vap.aau.dk/mintpain-database/) [37] has data for 20 subjects captured during the electrical muscle pain simulation. Each subject exhibit two trials during the data capturing session, in which each trial has 40 sweeps of pain stimulation. In each sweep, two kinds of data are captured: one for no pain (BL) and the other one for four kinds of different pain levels (i.e., PA1, PA2, PA3, and PA4). In total, each trial has 80 folders for 40 sweeps. To enlarge this dataset, we augment the videos twenty times using random cropping. Since the BL samples are four times more than other types of data, 1/4 of the data is randomly selected from the BL to maintain the balance with other categories.

For the BioVid dataset, the subject-independent experiment is conducted by randomly selecting five subjects (5.75%) for validation, five subjects (5.75%) for testing, and the rest (88.5%) for training. As for the MIntPAIN dataset, we perform subject-dependent pain recognition and subject-independent pain recognition separately. In the case of subject-independent, we divide the 20 subjects into five groups, and then the five-fold cross-validation is conducted. In the case of subject-dependent, we use 300 samples from each pain level category as the verification set, 300 samples are used as the testing set, and the rest as the training set.

4.2 Results Facial Expression Based SINN

The proposed SINN is utilized on BioVid and MIntPAIN datasets for pain assessment. For fairness, several models based on deep learning are selected for comparison. Up until now, rarely researches utilize the deep learning method to deal with pain recognition problems mainly due to the limitation of appropriate datasets. Daniel and Rosalind [47], [48] employed the feasibility of using physiological signals to detect the presence of pain by RNNs through regression on BioVid. They explored the effectiveness of biomedical signals for pain assessment. To the best of our knowledge, we are the first to propose a deep learning network that combines facial expression video and biomedical signals for assessing pain. Therefore, the well-applied deep learning models which deal with the video processing problem successfully are chosen to be compared, including two-stream neural networks [42] with a low frame rate (Two-stream NN low) and high frame rate (Two-stream NN high), and slow-fast neural networks [46].

The two-stream neural networks have two kinds of input (original image sequences and optical flow sequences), and the facial features are combined after convolution layers. Slow-fast is a network structure that combines the two streams of original facial sequences with a slow frame rate

Table 1 The results for binary classification on BioVid (BL and PA4).

methods	accuracy	precision	recall	F1-score
SINN	0.65	0.673	0.554	0.608
SINN+StF	0.603	0.577	0.687	0.627
SINN+FtS	0.62	0.558	0.666	0.607
Two-stream NN low	0.62	0.599	0.658	0.627
Two-stream NN high	0.626	0.623	0.582	0.602
Slowfast	0.6	0.597	0.518	0.555

Table 2 The accuracy comparison of facial expression based pain classification on the BioVid and MIntPAIN datasets.

	Binary classification		Multi-level classification	
Method	BioVid	MIntPAIN	BioVid	MIntPAIN
SINN	0.65	0.636	0.281	0.252
SINN+StF	0.603	0.601	0.262	0.235
SINN+FtS	0.62	0.613	0.249	0.227
Two-stream NN low	0.62	0.598	0.242	0.231
Two-stream NN high	0.626	0.601	0.243	0.232
Slowfast	0.6	0.588	0.233	0.219

and fast frame rate, respectively. Moreover, there are three types of variances of the facial expression-based SINN: the low frame rate to high frame rate fusion (SINN+StF), the high frame rate to low frame rate fusion (SINN+TtS), and non-fusion (SINN). The difference is the manner of interaction between the low stream and fast stream. The six deep learning models are conducted on the BioVid dataset based on facial expression image sequences for pain detection, which is a binary classification task with two categories (pain and no pain).

The experimental results are compared in Table 1 using several metrics, including accuracy, precision, recall, and F1-score. To assess the statistical significance of these methods, we use McNemar's test [49] with $p < 0.05$. As shown in Table 1, SINN achieves the highest accuracy and precision, which are 2.4% (accuracy) and 5% (precision) significantly higher ($p < 0.05$) than the second-best model. These results prove that the information extracted by variant streams fusion plays a positive role in pain classification. SINN+StF achieves the highest recall rate and F1-score, i.e. the recall of SINN+StF is 2.9% higher than the second-best model. Similar results are achieved on the MIntPAIN dataset for binary classification. The SINN model achieves an accuracy of 0.636, which was around 3% higher than that of two-stream NN, and almost 5% higher than that of the SlowFast network ($p < 0.05$).

Assessing multiple levels of pain is more challenging than the binary assessment. Pain intensities assessment is a typical multi-level classification task (PA1, PA2, PA3, PA4, and no pain). Table 2 present the overall performance of binary and multi-class assessment on BioVid and MIntPAIN datasets. In most cases, SINN achieves the highest performance and outperforms other models in multi-level pain assessment. The result of SINN with StF is slightly higher than that of SINN with FtS. We think this might be attributed to the over-fitting of SINN with the FtS model. All the

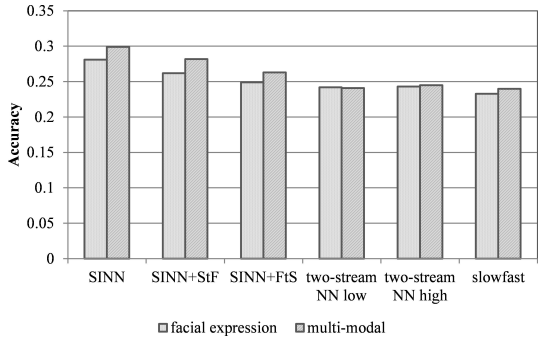


Fig. 7 Comparison of facial expression-based and multimodal-based multi-level classification pain assessment on BioVid dataset.

SINN-based models achieve higher accuracy compared to the two-stream model and the slow-fast model ($p < 0.05$), which indicates the effectiveness of the proposed SINN method. The best result obtained by SINN on the BioVid dataset is 28.1%, and on the MIntPAIN dataset is 25.2% for five-category pain intensity classification. The accuracy of multi-level classification is much lower than that of binary classification. This is attributed to the fact that multi-level classification is more complicated, and the pain label is annotated by the stimuli level instead of the subject report, i.e. the pain tolerance of different subjects is different, leading to variant respondences among subjects. There are a certain amount of high pain level videos with no obvious pain facial expression.

The proposed SINN model obtain the highest accuracy and significantly outperform the compared deep learning models by up to 4% for multi-level classification pain assessment on the MIntPAIN dataset. Moreover, the proposed SINN model outperforms the baseline model of the MIntPAIN dataset [37], which applied convolutional neural networks using only RGB image sequences, by up to 6.6% higher accuracy (SINN: 25.2% accuracy and baseline: 18.6% accuracy). These results suggest that the proposed network enhances the performance of the deep learning model significantly, and it can obtain the spatial and motion information implicitly (three-dimensional convolution) and explicitly (optical flow sequences).

In addition, we also conduct a subject-dependent pain assessment on MIntPAIN for the multi-level classification task. The proposed network achieves 87.1% accuracy, which is much higher than the subject-independent assessment. These results indicate that the SINN model could obtain rich facial information to present pain states and the rigid facial change influences pain recognition performances. Figure 8 presents the confusion matrix, where BL denotes the no-pain label, and PA1 to PA4 are four levels of pain intensities. As shown in the table, our method has better pain detection than no-pain detection. This might be attributed to the fact that some subjects made some kind of relaxation action during no pain stimulation, which confuses the identification between pain and no pain.

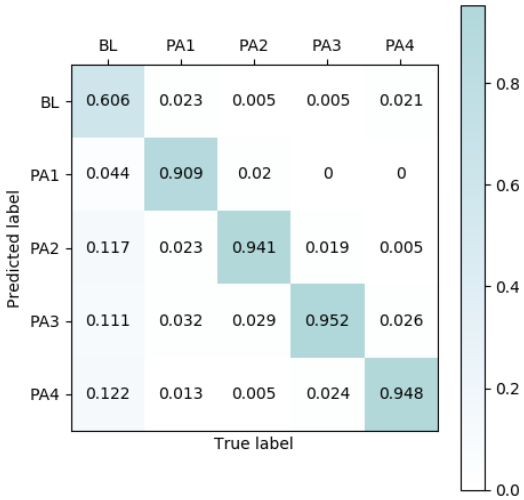


Fig. 8 The confusion matrix of SINN for the subject-dependent experiment on MIntPAIN dataset.

Table 3 The accuracy comparison of multimodal pain classification on BioVid dataset.

	Binary classification		Multi-level classification	
Method	Facial expression	Multimodal	Facial expression	Multimodal
SINN	0.65	0.682	0.281	0.299
SINN+StF	0.603	0.668	0.262	0.282
SINN+FtS	0.62	0.656	0.249	0.263
Two-stream NN low	0.62	0.649	0.242	0.241
Two-stream NN high	0.626	0.638	0.243	0.245
Slowfast	0.6	0.638	0.233	0.24

4.3 Results of Multimodal-Based SINN

In this section, we present the results of assessing pain using both facial expression and biomedical signal. Specifically, the facial expression-based pain assessment models evaluated in Sect. 3.3 are fused with biomedical signal features learned by LSTM networks. The multimodal-based SINN is compared with facial expression-based SINN and several deep learning models.

The comparison of binary classification accuracy is illustrated in Table 3 with single-model and multi-modal. McNemar's test is used to measure statistical significance with $p < 0.05$. It can be seen from the table that joint features can promote pain assessment performance by more than 3%, and there is a significant difference between multimodal-based and facial expression-based models. According to the comparisons of multimodal-based deep learning methods, the best results are obtained by multimodal-based SINN with an accuracy of 68.2%, which is up to 5% higher than that of the SlowFast method. The difference analysis shows that there are significant differences between SINN and Two-stream NN high and SlowFast schemes ($p < 0.05$). Despite no significant difference between SINN and Two-stream NN low in the statistical difference analysis, the accuracy of SINN is approximately more than 3% higher than that of Two-stream NN low.

Table 4 The multi-level classification confusion matrix of the multimodal-based SINN on the BioVid dataset.

	BL	PA1	PA2	PA3	PA4
BL	0.679	0.197	0.037	0.014	0.073
PA1	0.549	0.272	0.051	0.046	0.082
PA2	0.543	0.146	0.135	0.035	0.141
PA3	0.567	0.181	0.07	0.042	0.14
PA4	0.353	0.191	0.04	0.058	0.358

Figure 7 visually illustrates the comparison between facial expression-based and multimodal-based multi-level classification pain assessment, where the left bar denotes the results of facial expression-based pain assessment, and the right bar denoted the results of multimodal-based pain assessment. As can be seen in the figure, the multimodal-based SINN achieves 29.9% accuracy, which is almost 2% higher than facial expression-based SINN. Moreover, the multimodal-based SINN outperforms two-stream-based neural networks and slow-fast neural networks by up to 5.4% and 5.9%, respectively. Significant differences are found ($p < 0.05$) between the accuracy of SINN and Two-stream NN and Slowfast. Therefore, we can conclude that combining different features can enhance pain assessment performance, and the biomedical signal is helpful for pain recognition. The performance of multimodal-based SINN is acceptable (overall accuracy of binary classification was 68.2%) but lower than the performance of the traditional machine learning-based pain assessment. For example, the method proposed by Markus et al. [36] which combined facial features with biomedical signals achieved 83.1% binary classification accuracy when evaluated on BioVid dataset with leave-one-subject-out cross-validation. In [50], the authors achieved 80.6% accuracy using a Random Forest (RF) with bio-physiological, facial expression, and head movement features.

We believe that these methods achieve better overall performance due to two main reasons. First, the traditional machine learning methods extract hand-crafted features from facial expression and several biomedical signals respectively, which can explore the intrinsic information of multi-modal through different manners. Our proposed deep learning scheme utilizes unified architecture to deal with the sequence signals, and there is no need to design the feature extraction and classifier beforehand. Second, the five-fold cross-validation (subject-independent) utilized in our experiments is more challenging than leave-one-subject-out cross-validation, as there are fewer training samples and more testing samples, which can lead to lower classification accuracy.

The confusion matrix of the multimodal-based SINN for multi-level classification on the BioVid dataset is shown in Table 4. The columns of the table are the predicted labels, and the rows are the ground truth labels. As can be

seen, almost half of the pain samples are misclassified to the no pain category. The case is not entirely due to the network structure and self-learning scheme, but mostly because of the instance variation for different pain stimuli. A great number of video samples with pain labels have no obvious facial changes, leading to the deviation of the information extracted in the facial feature learning. It is even worse for PA2 and PA3. Although integrating physiological features can alleviate this problem to a degree, the misclassification problem is not improved.

5. Conclusion

This paper presents a novel network that utilizes different pain indicators to deeply exploit pain-related information. The scheme is implemented by a multimodal-based stream integrated neural network with different frame rates (SINN), which employs both facial expression and biomedical signals for pain assessment. In addition, the network learns dynamic facial features in both implicit manner and explicit manner, which is conducted by different frame rate operation and optical flow image sequence processing, respectively. Multi-streams can reflect the spatial information (original image sequences), the temporal information (optical flow sequences), the static information (slow pathway), and the dynamic information (fast pathway), which enrich the ability for characteristic facial description by facial features. Experimental results on public BioVid and MIntPAIN pain datasets illustrated that the proposed SINN model performed well for pain detection and pain intensity recognition, especially for the binary classification task. The results also showed that the joint feature from facial expression and biomedical signals could promote the accuracy of the automatic pain recognition system. The multimodal-based SINN achieves a better accuracy which is up to 5% higher than facial expression-based SINN in binary pain classification. And in multi-level pain classification, the accuracy is enhanced for almost 2% comparing the multimodal-based SINN to single facial expression modal. The deep learning method needs a large scale of training samples to enhance the classification performance, and we will try a larger dataset to evaluate the proposed SINN in the future.

Acknowledgements

This research was funded by the National Natural Science Foundation of China, grant number 61673052, the National Research and Development Major Project, grant numbers 2017YFD0400100, the Fundamental Research Fund for the Central Universities of China, grant numbers FRF-TP-20-10B, FRF-GF-19-010A, FRF-IDRY-19-011. The computing work is supported by USTB MatCom of Beijing Advanced Innovation Center for Materials Genome Engineering.

References

[1] B. Gholami, W.M. Haddad, and A.R. Tannenbaum, "Relevance vector machine learning for neonate pain intensity assessment using digital imaging," IEEE Trans. Biomed. Eng., vol.57, no.6, pp.1457–1466, 2010.

[2] K.M. Prkachin and P.E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," Pain, vol.139, no.2, pp.267–274, 2008.

[3] C.C. Johnston and M.E. Strada, "Acute pain response in infants: a multidimensional description," Pain, vol.24, no.3, pp.373–382, 1986.

[4] M.H. Willis, S.I. Merkel, T. Voepel-Lewis, and S. Malviya, "Flacc behavioral pain assessment scale: a comparison with the child's self-report," Pediatric nursing, vol.29, no.3, p.195, 2003.

[5] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H.C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A.O. Andrade, and G.M. da Silva, "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," 2013 IEEE international conference on cybernetics (CYBCO), pp.128–131, IEEE, 2013.

[6] G. Zamzmi, R. Kasturi, D. Goldgof, R. Zhi, T. Ashmeade, and Y. Sun, "A review of automated pain assessment in infants: features, classification tasks, and databases," IEEE Rev. Biomed. Eng., vol.11, pp.77–96, 2017.

[7] A.d.O. Andrade, Decomposition and analysis of electromyographic signals, Ph.D. thesis, University of Reading, 2005.

[8] M. Tavakolian and A. Hadid, "A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics," International Journal of Computer Vision, vol.127, no.10, pp.1413–1425, 2019.

[9] M. Tavakolian, M.B. Lopez, and L. Liu, "Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation," Pattern Recognition Letters, vol.140, pp.26–33, 2020.

[10] Y. Huang, L. Qing, S. Xu, L. Wang, and Y. Peng, "Hybnet: a hybrid network structure for pain intensity estimation," The Visual Computer, pp.1–12, 2021.

[11] L. Nanni, S. Brahnam, and A. Lumini, "A local approach based on a local binary patterns variant texture descriptor for classifying pain states," Expert Systems with Applications, vol.37, no.12, pp.7888–7894, 2010.

[12] M.S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A.C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P.J. Watson, A.C. de C. Williams, M. Pantic, and N. Bianchi-Berthouze, "The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset," IEEE transactions on affective computing, vol.7, no.4, pp.435–451, 2015.

[13] S. Agrawal and P. Khatri, "Facial expression detection techniques: based on viola and jones algorithm and principal component analysis," 2015 Fifth International Conference on Advanced Computing & Communication Technologies, pp.108–112, IEEE, 2015.

[14] G. Lu, X. Li, and H. Li, "Research on recognition for facial expression of pain in neonates," Acta Optica Sinica, vol.28, no.11, pp.2109–2114, 2008.

[15] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," International Symposium on Visual Computing, pp.368–377, Springer, 2012.

[16] A.B. Ashraf, S. Lucey, J.F. Cohn, T. Chen, Z. Ambadar, K.M. Prkachin, and P.E. Solomon, "The painful face–pain expression recognition using active appearance models," Image and vision computing, vol.27, no.12, pp.1788–1796, 2009.

[17] M. Rupenga and H.B. Vadapalli, "Automatic spontaneous pain recognition using supervised classification learning algorithms," 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), pp.1–6, IEEE, 2016.

[18] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, and H.C. Traue, "Head movements and postures as pain behavior," PloS one, vol.13, no.2, p.e0192767, 2018.

[19] A. Ruiz, O. Rudovic, X. Binefa, and M. Pantic, "Multi-instance dynamic ordinal random fields for weakly-supervised pain intensity estimation," Asian Conference on Computer Vision, pp.171–186, Springer, 2016.
[20] R. Yang, S. Tong, M. Bordallo, E. Boutellaa, J. Peng, X. Feng, and A. Hadid, "On pain assessment from facial videos using spatio-temporal local descriptors," 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), pp.1–6, IEEE, 2016.
[21] J. Chen, Z. Chi, and H. Fu, "A new framework with multiple tasks for detecting and locating pain events in video," Computer Vision and Image Understanding, vol.155, pp.113–123, 2017.
[22] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H.C. Traue, "Automatic pain assessment with facial activity descriptors," IEEE Transactions on Affective Computing, vol.8, no.3, pp.286–299, 2016.
[23] D. Bourou, A. Pampouchidou, M. Tsiknakis, K. Marias, and P. Simos, "Video-based pain level assessment: Feature selection and inter-subject variability modeling," 2018 41st International Conference on Telecommunications and Signal Processing (TSP), pp.1–6, IEEE, 2018.
[24] R. Kharghanian, A. Peiravi, and F. Moradi, "Pain detection from facial images using unsupervised feature learning approach," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.419–422, IEEE, 2016.
[25] P. Rodriguez, G. Cucurull, J. Gonzàlez, J.M. Gonfaus, K. Nasrollahi, T.B. Moeslund, and F.X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," IEEE Trans. Cybern., 2017.
[26] J. Egede, M. Valstar, and B. Martinez, "Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation," 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), pp.689–696, IEEE, 2017.
[27] M. Bellantonio, M.A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T.B. Moeslund, P. Rasti, and G. Anbarjafari, "Spatio-temporal pain recognition in cnn-based super-resolved facial images," Video Analytics. Face and Facial Expression Recognition and Audience Measurement, pp.151–162, Springer, 2016.
[28] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent convolutional neural network regression for continuous pain intensity estimation in video," Proc. IEEE conference on computer vision and pattern recognition workshops, pp.84–92, 2016.
[29] J. Walsh, C. Eccleston, and E. Keogh, "Pain communication through body posture: The development and validation of a stimulus set," PAIN®, vol.155, no.11, pp.2282–2290, 2014.
[30] T.A. Olugbade, N. Bianchi-Berthouze, N. Marquardt, and A.C. Williams, "Pain level recognition using kinematics and muscle activity for physical rehabilitation in chronic pain," 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp.243–249, IEEE, 2015.
[31] C. Wang, T.A. Olugbade, A. Mathur, A.C.D.C. Williams, N.D. Lane, and N. Bianchi-Berthouze, "Automatic detection of protective behavior in chronic pain physical rehabilitation: A recurrent neural network approach," arXiv preprint arXiv:1902.08990, 2019.
[32] M. Kächele, M. Amirian, P. Thiam, P. Werner, S. Walter, G. Palm, and F. Schwenker, "Adaptive confidence learning for the personalization of pain intensity estimation systems," Evolving Systems, vol.8, no.1, pp.71–83, 2017.
[33] S. Walter, S. Gruss, K. Limbrecht-Ecklundt, H.C. Traue, P. Werner, A. Al-Hamadi, N. Diniz, G.M. da Silva, and A.O. Andrade, "Automatic pain quantification using autonomic parameters," Psychology & Neuroscience, vol.7, no.3, pp.363–380, 2014.
[34] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H.C. Traue, "Automatic pain recognition from video and biomedical signals," 2014 22nd International Conference on Pattern Recognition, pp.4582–4587, IEEE, 2014.
[35] S.D. Subramaniam and B. Dass, "Automated nociceptive pain assessment using physiological signals and a hybrid deep learning network," IEEE Sensors J., 2020.
[36] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, "Multimodal data fusion for person-independent, continuous estimation of pain intensity," International Conference on Engineering Applications of Neural Networks, pp.275–285, Springer, 2015.
[37] M.A. Haque, R.B. Bautista, F. Noroozi, K. Kulkarni, C.B. Laursen, R. Irani, M. Bellantonio, S. Escalera, G. Anbarjafari, K. Nasrollahi, O.K. Andersen, E.G. Spaich, and T.B. Moeslund, "Deep multimodal pain recognition: a database and comparison of spatio-temporal visual modalities," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp.250–257, IEEE, 2018.
[38] P. Thiam, V. Kessler, M. Amirian, P. Bellmann, G. Layher, Y. Zhang, M. Velana, S. Gruss, S. Walter, H.C. Traue, et al., "Multi-modal pain intensity recognition based on the senseemotion database," IEEE Transactions on Affective Computing, 2019.
[39] J.C. Gower, "Generalized procrustes analysis," Psychometrika, vol.40, no.1, pp.33–51, 1975.
[40] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," Proc. IEEE conference on computer vision and pattern recognition, pp.4295–4304, 2015.
[41] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.
[42] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," arXiv preprint arXiv:1406.2199, 2014.
[43] D.H. Hubel and T.N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," Journal of neurophysiology, vol.28, no.2, pp.229–289, 1965.
[44] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth: anatomy, physiology, and perception," Science, vol.240, no.4853, pp.740–749, 1988.
[45] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," Proc. IEEE conference on computer vision and pattern recognition, pp.2117–2125, 2017.
[46] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," Proc. IEEE/CVF International Conference on Computer Vision, pp.6202–6211, 2019.
[47] D. Lopez-Martinez and R. Picard, "Multi-task neural networks for personalized pain recognition from physiological signals," 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp.181–184, IEEE, 2017.
[48] D. Lopez-Martinez and R. Picard, "Continuous pain intensity estimation from autonomic signals with recurrent neural networks," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.5624–5627, IEEE, 2018.
[49] G.M. Foody, "Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority," Remote Sensing of Environment, vol.113, no.8, pp.1658–1663, 2009.
[50] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H.C. Traue, "Automatic pain recognition from video and biomedical signals," 2014 22nd International Conference on Pattern Recognition, pp.4582–4587, IEEE, 2014.

Ruicong Zhi received the Ph.D. degree in signal and information processing from Beijing Jiaotong University in 2010. From 2016 2017, she visited the University of South Florida as a visiting scholar. She visited the Royal Institute of Technology (KTH) in 2008 as a joint Ph.D. She is currently a full professor in the School of Computer and Communication Engineering, University of Science and Technology Beijing. She has published more than 60 papers, and more than twenty patents. She has been the recipient of more than ten awards, including the National Excellent Doctoral Dissertation Award nomination. Her research interests include facial and behavior analysis, artificial intelligence, and pattern recognition.

Caixia Zhou received the Bachelor degree in Anhui University of Finance and Economics in 2018. She is currently pursuing the MS degree at the School of Computer and Communication Engineering, University of Science and Technology Beijing. Her research interest includes computer vision and emotion analysis.

Junwei Yu received the Bachelor degree in Jinan University in 2017. He is currently pursuing the MS degree at the School of Computer and Communication Engineering, University of Science and Technology Beijing. His research interest includes computer vision and signal processing.

Tingting Li received the MS degree at School of Computer and Communication Engineering from the University of Science and Technology Beijing in 2019 and received the Bachelor degree in Computer Science from the University of Science and Technology Beijing in 2016. Her research interest includes computer vision and emotion analysis.

Ghada Zamzmi received the MS degree in Computer Science from the University of South Florida, 2015. She is currently working toward the Ph.D. degree at Computer Science and Engineering, University of South Florida. Her research interest includes computer vision, image/video analysis, and emotion recognition for healthcare and human-computer interface.