IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 22, 2020, accepted June 2, 2020, date of publication June 4, 2020, date of current version June 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3000066

Spatial–Spectral–Temporal Framework for Emotion Recognition

KAN HONG
School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang 330000, China
e-mail: 179739527@qq.com

This work was supported in part by the National Natural Science Foundation of China under Grant 61866015, and in part by the Science and Technology Project Foundation of the Education Department of Jiangxi Province under Grant GJJ180598.

ABSTRACT An emotion recognition method based on multispectral imaging technology and tissue oxygen saturation (StO2) is proposed in this study. This method is called spatial–spectral–temporal adjustment convolutional neural network (SACNN). First, we use the algorithm to extract the StO2 content of an emotionally sensitive nose area through real-time multispectral imaging technology. Compared with facial expression data, StO2 data are more objective and cannot be controlled and changed artificially. Second, we construct a clustering algorithm based on the emotional state by extracting the spectral, StO2, and spatial features of the nose image to obtain accurate signals of emotionally sensitive areas. To utilize the correlation between spectral and spatial signals, we propose an adjustment-based CNN module, which reorganizes the relationship between all previous layers of the feature map, thereby making the relationship among layers close and highly quantitative. The features extracted through this method are consistent with spatial–spectral features. Third, we incorporate the extracted temporal feature signal into the long short-term memory module and finally complete the correlation between the spatial–spectral–temporal features. Experimental results show that the accuracy of the SACNN algorithm in emotional recognition reaches 90%, and the proposed method is more competitive than state-of-the-art approaches. To the best of our knowledge, this study is the first to use time-series StO2 signals for emotion recognition.

INDEX TERMS Multispectral imaging, oxygen saturation, spatial–spectral–temporal adjustment convolutional neural network.

I. INTRODUCTION

As the basis of human–computer interaction (HCI), emotion recognition affects the continuous development of machine intelligence. Many mental diseases are relevant to emotions [1], [2]. Therefore, research on emotion recognition technology has a great development prospect and academic value. Emotional recognition is essentially pattern recognition, and increased focus has been devoted to developing emotional artificial intelligence in HCI.

The methods of emotion recognition have achieved notable performance, but improvements are still necessary.

(1) Researchers have attempted to use spectral signals to construct a model for single emotion assessment, such as stress. However, using spectral imaging technology to recognize multiple emotions remains to be an undeveloped area.

(2) Features of emotion (e.g., facial expression and breath rate) are easily controlled by humans. Hence, data objectivity is affected.

(3) Deep learning algorithms have been applied in learning MSI-based psychological feature. However, the deep learning algorithm is still unable to implement the corresponding process to consider the spectral and spatial characteristics of emotional features.

To address these problems, we developed an emotion recognition algorithm called spatial–spectral–temporal adjustment convolutional neural network (SACNN) based on nose tissue oxygen saturation (StO2) information and multispectral signals. We applied MSI to determine the stress state in the past few years and discovered that StO2 is a sensitive and important parameter for stress assessment [47]–[52]. We believe that multiple emotion recognition features can be identified through spectral vision and StO2 signals. Therefore, the proposed algorithm obtains a multiband face image via MSI technology and extracts the StO2 signal at the

The associate editor coordinating the review of this manuscript and approving it for publication was Szidónia Lefkovits.

nose through multiband information. To obtain emotionally sensitive signals accurately, we perform clustering on the StO2 and spectral signals to initialize the zones with the same emotional properties or spectral signal base in the same data cube for the next step. Such a base can lay a foundation for the subsequent feature extraction. We discard the original clustering method that relies solely on calculating ''distance'' and use the correlation mechanism of the target node and background context instead to construct an inter-band correlation between node and intervals. Accordingly, the correlation between different bands is re-clustered, and a model is established through the strength of correlation to achieve a clustering method for inter-band cross-time domains. Then, we use StO2 and the spectral signal cube after clustering as input signals for the next deep learning investigation.

The proposed SACNN model fully uses the correlation between spectral and spatial information. CNNs are well-established techniques for image processing applications [53]–[61]. Although CNNs have been popular for many years already, their actual application has become successful only in recent years. DenseNet, which was proposed in CVPR2017, reduces the possibility of gradient disappearance by linking the feature in the current layer to those of all previous ones in the training process [61]. However, no training learning model, especially for the characteristics of multispectral data among current deep learning algorithms, is available for emotion recognition. The SACNN algorithm module uses spectral and spatial correlations to make the relationship between convolution layers close and quantitative. Our algorithm module highlights the connection and interaction between layers and the multispectral data cube, particularly for the feature training set, thereby making feature extraction targeted. In our applications, we aim to determine the differences among various emotions by using StO2 and spectral signals and enable these differences to be reflected in the layers of image learning. We believe that from the relationship and variance between layers, hints can be obtained regarding the correlation and difference between spectral bands and StO2 signals. The main contributions of this study are as follows:

(1) StO2 signals (i.e., at the human nose) and MSI technology are used for the first time as the signal source and multiple emotion recognition tool, respectively. Our method is objective because the human body's StO2 constitutes information that cannot be changed artificially.

(2) A clustering algorithm based on the spectral domain and relationships is proposed, and StO2 and spectral signals are used to achieve target area clustering based on the correlation between spectral and spatial information.

(3) The SACNN algorithm framework is proposed for multispectral signals and StO2 characteristics. This algorithm further excavates the spatial–spectral correlation and difference through an adjustment method. The technique also helps us reconstruct the relations and weights between all feature maps (FMs), thereby making the feature extraction process targeted and accurate.

The rest of this paper is organized as follows. Section II gives literature review, and section III comprehensively describes the proposed SACNN method for emotion recognition. Section IV presents the details of experiment and discusses the experimental results to evaluate the proposed method. Section V provides the conclusions of this study.

II. LITERATURE REVIEW

Current methods of emotion recognition mainly involve facial expression recognition [3]–[6], speech emotion recognition [7]–[9], gesture expression recognition [10], text recognition [11], physiological pattern recognition, and multimodal emotion recognition [12]–[15]. In practical applications, the non-contact method of extracting physiological parameters for face imaging has attracted special attention. This method typically extracts physiological features from an image through face imaging, and the correlation between features and the ground truth are jointly modeled, thereby allowing a computer to ''read'' the human emotional state.

Many studies in the past 20 years have used infrared thermal imaging to recognize fear and anxiety among emotions [16]–[24]. Face imaging signals have likewise been applied to extract relevant physiological information, and these signals also play an important role in emotion recognition. When the human body is in a special emotional state, a series of physiological reactions occur on their own, and corresponding recognition models can be constructed using physiological parameters. For example, effective extraction and extensive research have been conducted on the physiological parameters of the human body (e.g., sweating [25]–[27], blood flow velocity [20], [22], heart rate [21], [28], and breathing [25]). Emotion-sensitive facial muscles and regions (e.g., supraorbital, cheek, and perinasal areas) have also been extracted and studied [28]–[34]. Subtle changes that occur in the face, such as head motion (shaking), head pose, yawning, eye blink rate, and eye closure duration, have been utilized to detect emotions and fatigue [35]–[41]. Deep learning algorithms have also been widely applied in emotion recognition [42], [43].

In addition to RGB and infrared thermal imaging, spectral vision technology has received increased attention in the field of biomedical information in recent years. This technology combines traditional imaging and spectral technology and simultaneously acquires the spatial and spectral information of an object to determine its material characteristics [43]–[46]. Our research is the first to apply multispectral imaging (MSI) technology to assess the stress state [47]–[52]. We realize that the StO2 is a sensitive and important information in stress recognition [48]–[50].

III. SACNN MODEL FOR EMOTION RECOGNITION

Previous studies have shown that the nose is a sensitive and important position in emotion recognition [23], [48], [49], [67]. In this work, we take the nose as our region of interest (ROI). We extract the StO2 signal at the nose through multispectral data. To process emotion data specifically, we adopt

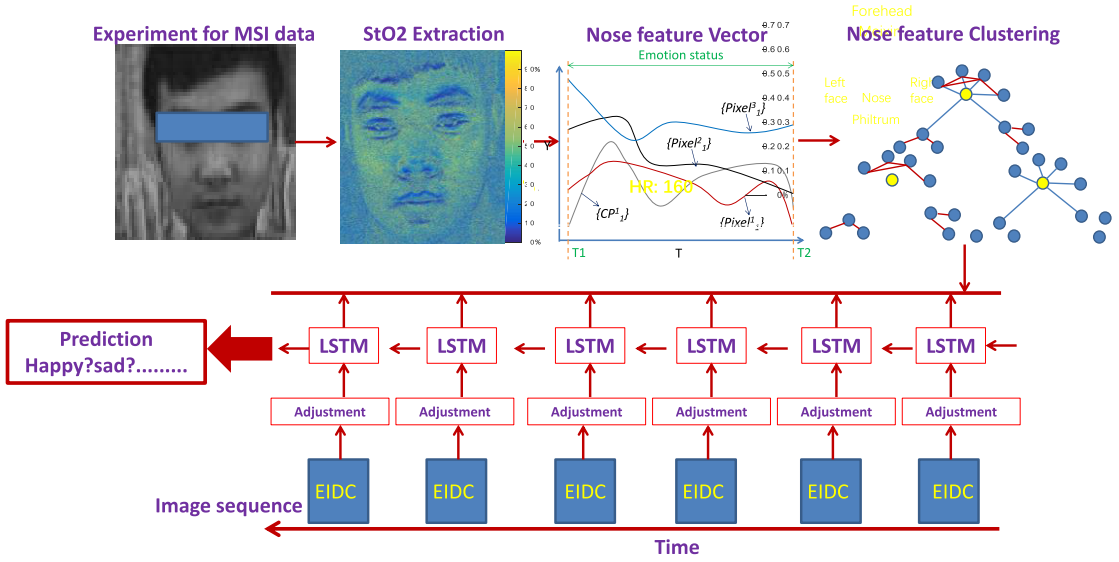

FIGURE 1. SACNN flowchart.

the spectral–spectral correlation between the StO2 signal and the multispectral data for clustering the nose signals. We call these clustered data cubes (i.e., four spectral wavelength layers and one StO2 data layer) an emotion-induced data cube (EIDC). Each data block we obtained involves 5D data after clustering. A data block also contains temporal data because these data are image sequences. Thus, we face a complex training object, that is, we need to extract features from the spectral–spatial–temporal signals. To extract the features of different emotional states, we propose the SACNN model by fully using the data feature. We add the adjustment model to the algorithm, and the adjustment is processed in accordance with the spectral and spatial correlation. The signal after the adjustment is processed as the input of the next layer. All front layer features are pretreated to ensure that the features of the current layer input can be optimized. The spectral and spatial characteristics are further integrated into the deep learning model. We expect that the adjustment can re-arrange the relevance and importance of all the features of the previous layer and influence the subsequent training; hence, our algorithm model is targeted and can fully use spectral–spatial characteristics. Afterward, the trained features are combined with the long short-term memory (LSTM) temporal treatment specialty to complete the emotion recognition. Figure 1 presents a flowchart of the algorithm model to describe the proposed SACNN model clearly.

A. StO2 SIGNAL EXTRACTION

The algorithm for extracting StO2 using MSI technology is based on the effect of light absorption on the human skin. The four main chromophores of the absorption spectrum of the human skin are deoxyhemoglobin (Hb), oxyhemoglobin (HbO2), scattering effect, and melanin. StO2 extraction is based on the Beer–Lambert (BL) model [64]–[66]. The BL model describes the relationship between light absorption and the thickness of the absorbing medium. The formula is expressed as follows:

$$A = \varepsilon_{HbO_2} C_{effHbO_2} + \varepsilon_{Hb} C_{effHb} + \varepsilon_{melanin} C_{effmelanin} + G' \quad (1)$$

where ε is the molar absorptivity of HbO2, Hb, and melanin; c is the effective concentration; and G' represents all parameters (e.g., skin mirror reflection and regression error) that are unrelated to the absorption rate of tissues. The effective concentrations of Hb, HbO2, and StO2 can be obtained through simultaneous decomposition of the selected spectral band data equations. In terms of wavelength selection, we fully consider the connection of the extinction coefficient of Hb, HbO2, and melanin in the human skin. The extinction coefficient of melanin declines rapidly with the increase in wavelengths. The weight of light absorption is low when the light infrared range is near. The algorithm does not use the ultraviolet band in reducing the weight of melanin, although Hb and HbO2 have several absorption peaks in the ultraviolet range. The extinction coefficient of Hb and HbO2 has a steep drop at 600 nm, a value that is nearly at the same order of magnitude as that for melanin. Therefore, this study obtains the corresponding real-time StO2 data from the four-band MSI data in the 500–600 nm interval.

B. CLUSTERING METHOD FOR THE REGION OF INTEREST

In this work, we take the nose as our ROI. However, directly using the nose as a signal source would result in a signal that is too rough. The pattern presented by different persons' StO2 images is inconsistent under different emotional states. To investigate the characteristics of the nose signal further and initialize the region with the same emotional attributes or the same spectral signal basis, we propose a clustering model based on spectral characteristics and StO2 signals to

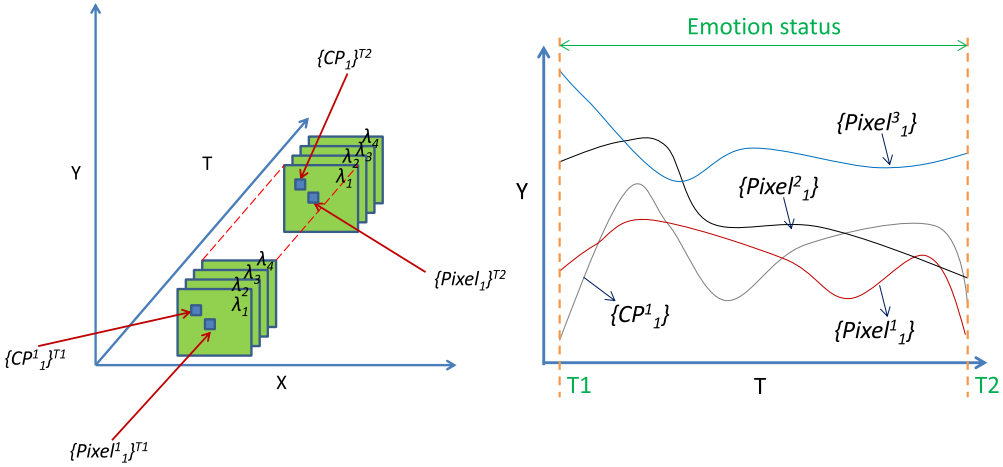

FIGURE 2. (a) Case with a spatial relationship between the cluster center $\left\{CP_1^1\right\}$ in the first band and the neighbor scan point under the T1 to T2 time series. (b) Case with the trend of the spectral signals in the first wavelength of the first cluster center and the three neighbor scanning points under different time series. We must identify the correlation between the time-series data at the cluster central point and all points of the neighbor scanning area in different wavelength spaces.

provide a foundation for feature extraction in the same data cube. We divide the signal region with spectral similarity into a cluster by using the clustering model. We use the data after clustering as the input of the next training and learning. This technique provides a recognizable signal basis for emotion recognition.

The most commonly used clustering method calculates the ''distance'' between spatial data [68]–[70]. The structure of the proposed clustering model uses the correlation structure between the spectral and StO2 signals. Our data are time-domain data. Thus, we must consider the current image cube and the data correlation in the entire time.

In accordance with the number of cluster sets, the initial cluster center is evenly distributed in the image. Given our time-series data, we expect that the spectral characteristics in the time domain will be considered together to obtain clustering relevance. Accordingly, we use a neighbor correlation structure to calculate the spectral correlation between StO2 and the spectral data of four wavelengths. We consider the relevance of the neighbor field and find that we only need to compute the linkage likelihood between a cluster central point (CP) and its k nearest neighbors. This approach can produce clustering results. At T1, we define the clustering center of the StO2 layer $\left\{\mathrm{CP}_{\mathrm{StO2}}^{\mathrm{j}}\right\}^{\mathrm{T1}}$ and the CP of the four wavelength cluster centers at the same spatial position as $\left\{\mathrm{CP}_{\mathrm{i}}^{\mathrm{j}}\right\}^{\mathrm{T1}}$ (Figure 2). Here, i stands for the wavelength, and j stands for the Nth cluster centers. The scanning object of the neighbor target is defined as $\left\{\mathrm{Pixel}_{\mathrm{StO2}}^{\mathrm{j}}\right\}^{\mathrm{T1}}$ and $\left\{\mathrm{Pixel}_{\mathrm{i}}^{\mathrm{j}}\right\}^{\mathrm{T1}}$, where i stands for the wavelength and j stands for the Nth point. The initialization is based on StO2 images, and the clustering center and scanning point are consistent in the four wavelengths. From Figure 2, we need to identify the correlation between CP and scanning points. Figure 2(a) illustrates the spatial relationship between a cluster center $\left\{\mathrm{CP}_{\mathrm{i}}^{\mathrm{j}}\right\}$ in the first wavelength and the neighbor scanning point from T1 to T2 time series. Figure 2(b) plots the trend of the CP spectral signals and the trend of spectral signals for the three scanning points at the different emotional states and time series. The changes in the spectral signals in the emotional state are evident, but no obvious fixed law can be observed, that is, the change is nonlinear. Thus, directly and linearly correlating the emotion ground truth with the image signals is difficult. Therefore, we need the correlation between the clustering CP and the spectral signal of all the scanned points. For this spatial domain, we propose a spectral template (ST) to calculate the spectral correlation.

We introduce the time domain into the algorithm and define the correlation matrix $CM\ (t)$ under the same time series. We start from the first cluster CP of the StO2 layer and define signal S_k between the scanning point and the cluster center (CP and Pixel). K represents the number of all the time-series signals, where we have four wavelengths and StO2 layers. K is set to 5. The normalized spectral and spatial signal is expressed as follows:

$$\overrightarrow{S_k(t)} = (S_k(t) + \overline{S_k(t)}/\sigma(t) \tag{2}$$

where $\sigma(t)$ is the standard deviation and $\overline{S_k(t)}$ is the mean. Subsequently, we use the Pearson correlation coefficient to analyze the parameters. The correlation matrix is defined as follows:

$$CM_{mn}(t) = \frac{1}{T}\sum\nolimits_{t=1}^{T} \overrightarrow{S_m(t)}\ \overrightarrow{S_n(t)} = \langle \overrightarrow{S_m(t)}\ \overrightarrow{S_n(t)} \rangle_t \tag{3}$$

where m and n stand for the number of time-series signals. The correlation in the structure of CP and the scanning object is included in the bivariate measures. The independent coefficient matrix $CM\ (t)$ can explain the cross correlation

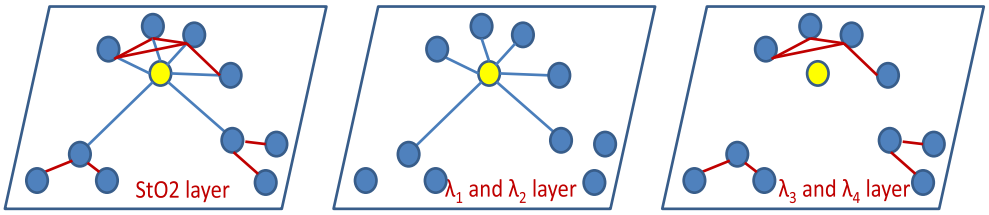

FIGURE 3. ST structure. We draw 1- and 2-hop neighbors of several CPs in StO2 in (a), and (b) presents the 1-hop correlation of the λ1 and λ2 wavelengths and the 2-hop correlation of the λ3 and λ4 wavelengths in (c) at the same spatial position. Thus, we can fully connect the correlation of all wavelengths. The blue line connection is the 1-hop connection, and the red line is the 2-hop connection.

between the clustering center and all other target points in the searching range.

The next step determines the relevant points by considering the relationship between different wavelengths. We must subdivide our scanning points further through these relationships. Accordingly, we define a multidimensional spatial relationship map ST. In ST, we define the StO2 layer as the top layer, and the four-wavelength signals are defined as two, three, four, and five layers. Then, we set the proximity point based on the cluster center KNNs in ST, and we use its high-order neighbors as the scanning point for ST. We thus define the high orders up to 2-hop of CP, and we set the 2-hop neighbor on the StO2 layer. We define the 1-hop node neighbor as 6 and the 2-hop neighbor as 2. Figure 3 shows our spectral feature structure. The CP neighbors provide us additional StO2 infrastructure and spectral structure data. Figure 3(a) plots the 1- and 2-hop neighbor points of several CPs in the StO2 layer. The blue line represents the 1-hop connection, and the red line signifies the 2-hop connection. Figure 3(b) depicts the 1-hop correlation in the λ_1 and λ_2 wavelengths, and Figure 3(c) demonstrates the 2-hop correlation in the λ_3 and λ_4 wavelengths at the same spatial position. When we use the neighbor point for seeking relevance, we are not completely placed on the StO2 layer but evenly reset on the four other wavelengths. This feature is important because we must fully connect the relevance of all wavelengths. Therefore, the 1-hop correlation in λ_1 and λ_2 and the 2-hop correlation in λ_3 and λ_4 will construct the ''distance'' equation for us.

We then calculate the correlation ''distance'' equation of the search point and CP and indirectly reflect the correlation of the four wavelengths on the 1- and 2-hop neighbor points. The target equation of correlation matrix of $\overrightarrow{CM}$ can be written as follows:

$$\min \left\| 1 - a\overline{CM_{1-hop}} - b\overline{CM_{2-hop}} - \overline{CM_{StO2}} \right\|^2 \quad (4)$$

Here, we set integers a and b as ambiguous. We ignore the integer constraint of fuzziness a. Solving the minimum value can be regarded as a standard least-squares estimation problem. This solution is frequently referred to as the floating solution, and its estimated value $\hat{a}$ and the related covariance matrix can be expressed as

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} \begin{bmatrix} Q_{\hat{a}} Q_{\hat{a}\hat{b}} \\ Q_{\hat{b}\hat{a}} Q_{\hat{b}} \end{bmatrix} \quad (5)$$

The fuzziness floating solution and its covariance matrix are used to calculate the fuzziness integer solution á as follows:

$$\min_{a} (\hat{a}-a)^T Q_{\hat{a}}^{-1} (\hat{a}-a) \quad (6)$$

Conversely, using the integer properties of fuzziness further improves the accuracy of the equation estimate.

$$\acute{b} = \hat{b} - Q_{\hat{b}\hat{a}} Q_{\hat{a}}^{-1} (\hat{a} - \acute{a}) \quad (7)$$

$$Q_{\acute{b}} = Q_{\hat{b}} - Q_{\hat{b}\hat{a}} Q_{\hat{a}}^{-1} Q_{\hat{a}\hat{b}} \quad (8)$$

where $\acute{b}$ is the estimated target fixed solution. The accuracy of the fixed solution and floating solution of the target equation is improved, and á is the fixed solution of integer ambiguity. Constant iteration of these steps involves seeking the minimum solution to the target equation, that is, the optimization process in which we find the maximum correlation. The optimal correlation point obtained using this equation is the basis of clustering, and we finally obtain the clustering signal space we require. Here, we complete the pretreatment of the data, and the data cube (including four spectral bands and one StO2 layer) after clustering is called EIDC.

C. SPECTRAL–SPATIAL FEATURE EXTRACTION

In this study, we aim to learn joint spectral–spatial features in the nose by using the proposed SACNN model for emotion detection. This algorithm is a combination of a CNN (DenseNet) and an adjustment model. Previous studies have revealed that the nose is sensitive to emotional states, and we must find the corresponding recognition feature. After completing the image clustering, we input all EIDCs into the training model. We identify the change features in nose StO2 and spectral signals between different emotional states and the differences among these changes. Our application aims to determine the pre- and post-difference between StO2 and spectral signals, and we expect the difference in performance to be reflected in the layers of image training and learning. We use DenseNet as our basic framework for training and learning. In the dense block of DenseNet,

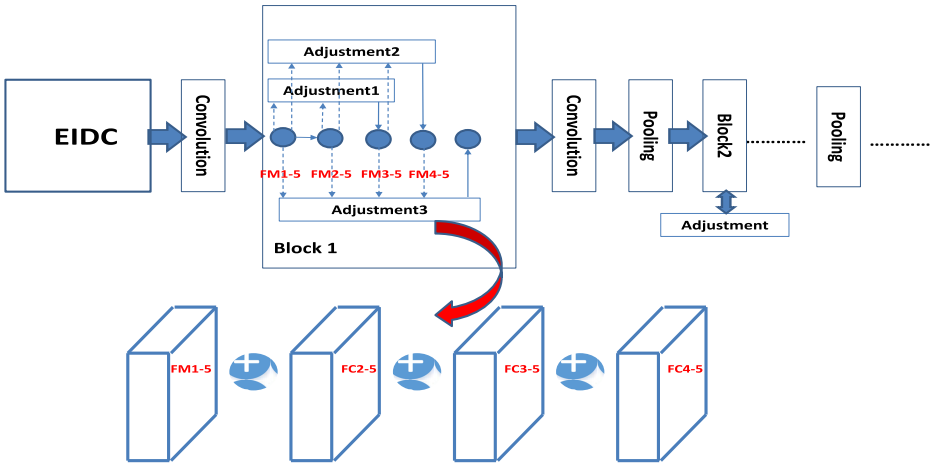

FIGURE 4. Structure diagram of the learning process in SACNN. The dotted and solid lines are the input and output features, respectively. In Block 1, the adjustment structure is described in detail. In Block 2, the adjustment process is repeated. The input of adjustment3 involves all FMs in the previous layer, and the output after adjustment is used as the input of the current layer. FM1–5 refers to the input FM from the first layer to the fifth layer. The other FMs are analogized similarly.

the current layer is directly linked to the FM of all previous ones, thus improving gradient elimination well. However, we believe that considering the layers' relationship and differences will generate additional possibilities for learning various features. We can obtain several hints from the correlation and difference between the spectral wavelength and StO2 signals. Therefore, in addition to the clustering operation that we completed using the correlation of the multispectral information outside deep learning, we begin by deep learning the interlayers to explore the relevance and difference of the spatial–spectral signals, with the connection of layers as the cutting point.

The adjustment model has been widely used in data processing and error analyses [71]–[73]. In data processing, adjustment utilizes excess observation data to eliminate contradictions and errors of observation. The correction or weight structure given in this process is used to reconstruct the relationship and weight of all previous FMs (that is, the input of the current layer). We expect to further excavate the spatial–spectral relevance and difference by using this method. Moreover, the degree of relevance between the current and previous layers is considered by SACNN.

In our model, the weight of the current layer when it is linked with all previous ones is distributed and handled properly with regard to the relevance of all previous layers. Thus, our algorithm structure can obtain the new weight of an FM when the current layer is linked with all previous ones. Such a structure is conducive to the close connection of spectral–spatial features between layers. The DenseNet structure takes all of the output of the previous layers as the input to maintain the continuity of the feature and prevent gradient disappearance. However, we address all FMs of the previous layers through adjustment to ensure a thorough excavation of the spectral–spatial feature. Subsequently, the data features change after each convolution, and this change may influence the current spectral features. We directly take the input of all the previous layers' features as the input of the current layer, which cannot highlight the correlation and characteristics of the spectra. The best way to improve this is to pretreat the features of all the previous layers to guarantee that the input features of the current layer can be optimized. We use the adjustment method to complete this process. Adjustment is a theoretical and computational method for handling various observation results. Its purpose is to eliminate contradictions between observations, obtain the most reliable results, and improve the accuracy of the evaluated results. The adjustment in our case encounters the FMs of the previous layers. Through the adjustment, the spectral and spatial correlation and the importance of the features of all previous layers can be rearranged and can influence the subsequent training.

Figure 4 displays the learning process in the SACNN algorithm. We still use DenseNet's block basic framework and utilize Block 1 as an example. The dotted and solid lines represent the output and input features, respectively. The block has five layers, and three adjustments are performed. The output of Block 1 continues to complete the convolution and pooling processes. In the next block, the adjustment process is also conducted. Each block has three adjustment processes because a block has five layers, and all the previous layers are used as the input for three times (Block 1, Figure 4). Inputting from the previous layers to the current layers occurs four times in Block 1. However, only the first layer's features are inputted for the second layer. Thus, this process does not require adjustment. Figure 4 illustrates the specific structure of adjustment3 in Block 1. The input of adjustment3 involves all FMs of the previous layers. Moreover, all the FMs' output after processing by adjustment3 are used as the input of the current layer (the fifth layer).

We take adjustment3 in Block 1 as an example (Figure 4) to explain our model. We employ Helmut's estimation

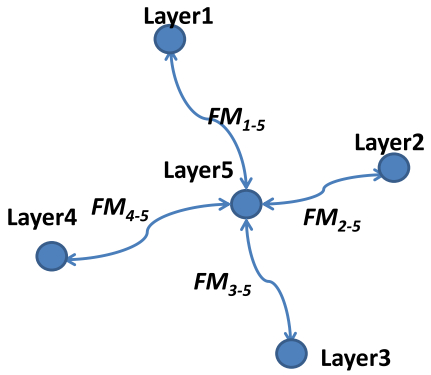

FIGURE 5. Adjustment structure in the dense block for adjustment3. Similar to the DenseNet structure, the FMs of all the previous layers are used as the input for adjustment3, and the FM after the adjustment is used as the input for the fifth layer.

formula [73] as the basis for our adjustment model. As the input of the fifth layer, FMs are generated from all previous layers to the fifth layer. We make adjustments for these FMs. The principle of adjustment is to allow reasonable modifications to the calculation results (adding corrected values or adjusting weights to reduce the error), thereby improving the data accuracy. The properties of the characteristic features are adjusted and extracted during FM adjustment because the adjustment itself is reconstructed through the characteristics of the features. In this manner, the following FM is obtained (Figure 5).

$$Layer1 \rightarrow Layer5 \rightarrow FeatureMap_{1-5} \quad (FM_{1-5}) \tag{9}$$

$$\cdots\cdots$$

$$Layer4 \rightarrow Layer5 \rightarrow FeatureMap_{4-5} \quad (FM_{4-5}) \tag{10}$$

FM_{1-5}, FM_{2-5}, FM_{3-5}, and FM_{4-5} are then acquired. With regard to the last layer FM, the four FMs are assumed to be independently obtained for an improved distribution of the weight and adjustment. Then, adjustment disposal is executed for the four obtained FMs in order for such disposal to be beneficial in optimizing the spatial–spectral structure. The initial weight value is set up in accordance with distance. Figure 5 illustrates the adjustment structure in the dense block for adjustment3. An FM is linked with Level 5 in each layer. Relevant adjustment disposal is performed for every link to reach the optimum state.

In this step, adjustment disposal is conducted for FMs generated in the four routes, and the output is the result of the adjustment disposal of FM_{1-5}, FM_{2-5}, FM_{3-5}, and FM_{4-5}.

$$FM_{1-5} + v_{1-5} = \widehat{X}_1 \tag{11}$$

$$FM_{2-5} + v_{2-5} = \widehat{X}_2 \tag{12}$$

$$FM_{3-5} + v_{3-5} = \widehat{X}_3 \tag{13}$$

$$FM_{4-5} + v_{4-5} = \widehat{X}_4 \tag{14}$$

$$f\left(FM_{1-5}, FM_{2-5}, FM_{3-5}, FM_{4-5}\right) = FM_5 \tag{15}$$

where $\widehat{X}$ is the theoretical value after adjustment, v is the correction error value, and FM is defined as L for the subject of adjustment. L is obtained as follows:

$$L_{1-5} = \widehat{X}_1 - v_{1-5} \tag{16}$$

$$L_{2-5} = \widehat{X}_2 - v_{2-5} \tag{17}$$

$$L_{3-5} = \widehat{X}_3 - v_{3-5} \tag{18}$$

$$L_{4-5} = \widehat{X}_4 - v_{4-5} \tag{19}$$

$$L_5 = f\left(FM_{1-5}, FM_{2-5}, FM_{3-5}, FM_{4-5}\right) - v_5 \tag{20}$$

Many choices are available for obtaining L_5. If we only use FM_{3-5} and FM_{4-5} for calculation and the weight parameters for the other FMs are set to 0, then we will obtain a typical ResNet structure. If all FMs are set to the non-zero weight parameters, then they will comply with the DenseNet framework. Different weight configurations inevitably generate different calculation results, which are called redundant observations in the adjustment. The error equation is obtained as

$$V = B\widehat{X} - L \tag{21}$$

where B is the coefficient matrix and V is the corrected positive matrix. To re-optimize the FMs, our goal is to expect $\sum v_i^2 = min$ and $V^T PV = min$. Here, P is the weight. Adjustment disposal is then performed by using the Helmut variance estimation method. We expect that

$$E\left(L\right) = B\tilde{X}, \quad E\left(\Delta\right) = 0 \tag{22}$$

$$D\left(L\right) = \sigma_0^2 P^{-1}, \quad D\left(\Delta\right) = D\left(L\right) = \sigma_0^2 P^{-1} \tag{23}$$

where D represents variance, P is the weight, σ_0^2 is the standard deviation, and Δ is the error. The error equations are expressed as follows:

$$V = B\widehat{X} - L \tag{24}$$

$$N = B^{\mathrm{T}} PB, \quad W = B^{\mathrm{T}} PL \tag{25}$$

$$N\widehat{X} = \mathrm{W}, \quad \widehat{X} = N^{-1}\mathrm{W} \tag{26}$$

where P is the weight. The following process is Helmut's estimation formula [73] for our SACNN model. We suppose that the two categories of observed values, $\underset{n_1\times 1}{L_1}$ and $\underset{n_2\times 1}{L_2}$ in L, are present; the weight matrices of the two categories are P_1 and P_2, respectively, and $P_{12} = 0$. Then, we obtain

$$V_1 = B_1\widehat{X} - L_1 \tag{27}$$

$$V_2 = B_2\widehat{X} - L_2 \tag{28}$$

We set

$$L = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}, \quad V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} \tag{29}$$

$$N = N_1 + N_2 \tag{30}$$

$$W = W_1 + W_2 \tag{31}$$

We correct the expectation of positive number V as 0, such that

$$E\left(V_1\right) = 0 \tag{32}$$

$$E\left(V_1^{\mathrm{T}} P_1 V_1\right) = tr(P_1 D(V_1)) \tag{33}$$

Algorithm: Iterative steps for the estimation of the variance component

Step 1: Initialize the weight of *FM* (according to the distance) and estimate it before the test, namely, confirm the initial value of each *FM* weight $\{P_i\}_{i=1,\ldots m}$. Perform first-time adjustment and update $V_i^T P_i V_i$.

Step 2: Carry out first-time estimation of the variance component and obtain the first-time estimation value of variance of the unit weight of various observed values. Then, determine the weight according to the following equation.

$$\hat{P}_i = \frac{c}{\hat{\sigma}_{0i}^2 P_i^{-1}}$$

where *C* is an arbitrary constant.

Step 3: Repeat the two steps above, namely, adjustment – variance component estimation – readjustment after determining the weight until $\hat{\sigma}_{0_1}^2 = \hat{\sigma}_{0_2}^2 = \ldots = \hat{\sigma}_{0_m}^2$

FIGURE 6. Iterative computation steps for the estimation of the variance component.

The mathematical expectation symbols are selected and changed to be the calculated values $V_1^{\mathsf{T}} P_1 V_1$ and $V_2^{\mathsf{T}} P_2 V_2$ obtained by adjustment. Subsequently, the matrix form is obtained as (34), (35), (36), and (37) shown at the bottom of this page.

The abovementioned equation is Helmut's estimation formula [73] for our SACNN model. The method of estimating various types of pre-test observation by adopting the correction value of pre-adjustments was first proposed by Helmut. A variety of computational quantity is independent of each other, that is, the variance matrix of the calculation quantity is a quasi-diagonal matrix, which is the variance estimate or the variance volume estimate. The formula solution for the estimated parameters is defined as follows:

$$\widehat{\theta} = S^{-1} W_\theta \tag{38}$$

Accordingly, we can obtain the estimation formula with m types of data. Thus, the number of estimated parameters is m [73]. If the abovementioned formula is changed into the matrix type, then an estimation equation with m types of data can be obtained as

$$S_{m\times m}\widehat{\theta}_{m\times 1} = W_{\theta m\times 1} \tag{39}$$

$$\widehat{\theta} = [\widehat{\sigma}_{0_1}^2 \widehat{\sigma}_{0_2}^2 \ldots\ldots \widehat{\sigma}_{0_m}^2]^T \tag{40}$$

$$W_\theta = [V_1^{\mathsf{T}} P_1 V_1 \; V_2^{\mathsf{T}} P_2 V_{2\ldots\ldots} \; V_m^{\mathsf{T}} P_m V_m]^{\mathsf{T}} \tag{41}$$

$$\widehat{\theta} = S^{-1} W_\theta \tag{42}$$

In our situation, m is set to 5. We summarize the iterative computation steps for the estimation of the variance component in Figure 6.

After completing the adjustment process, all FMs from the previous layer to the current layer are re-planned and reconstructed to comply with the adjustment principle for the emotion data. The algorithm further excavates the spatial–spectral relevance and difference and is conducive to a close connection of the spectral–spatial features between layers. Moreover, the degree of relevance between the current and previous layers is considered through SACNN. After extracting the feature of StO2 and spectral signals through SACNN (Figure 1), we place the features obtained from each time point in the timing process into the LSTM model and generate an output prediction for all the features as a whole. Therefore, we can obtain the complete SACNN emotion recognition structure.

IV. EXPERIMENTS

A visible and near-infrared MSI system that covered spectral wavelengths of 450–800 nm and a homemade tunable snapshot MSI system were used. The imaging system consisted of a Tamron lens, an acoustic optic tunable filter imaging spectrograph, and a computer. The area CCD array detector of the camera contained 1392 (h) × 1040 (v) active pixels with a spectral resolution of 2 nm. The frame rate of the MSI system was set to 30 Hz. The participants were recruited by posting an advertisement in a newspaper. A total of 250 healthy volunteers of both genders (52% male and 48% female) participated in the experimental trials. The participants were aged 20–60 years, with a mean age of 34.5 years and a standard deviation of 16.7.

$$S\widehat{\theta} = W_\theta \tag{34}$$

$$S = \begin{bmatrix} n_1 - 2tr\left(N^{-1}N_1\right) + \left(tr\left(N^{-1}N_1\right)^2\right) & tr(N^{-1}N_1N^{-1}N_2) \\ tr(N^{-1}N_1N^{-1}N_2) & n_2 - 2tr(N^{-1}N_2) + tr(N^{-1}N_2)^2 \end{bmatrix} \tag{35}$$

$$\widehat{\theta} = [\widehat{\sigma}_{0_1}^2, \widehat{\sigma}_{0_2}^2]^{\mathsf{T}} \tag{36}$$

$$W_\theta = [V_1^{\mathsf{T}} P_1 V_1, V_2^{\mathsf{T}} P_2 V_2]^{\mathsf{T}} \tag{37}$$

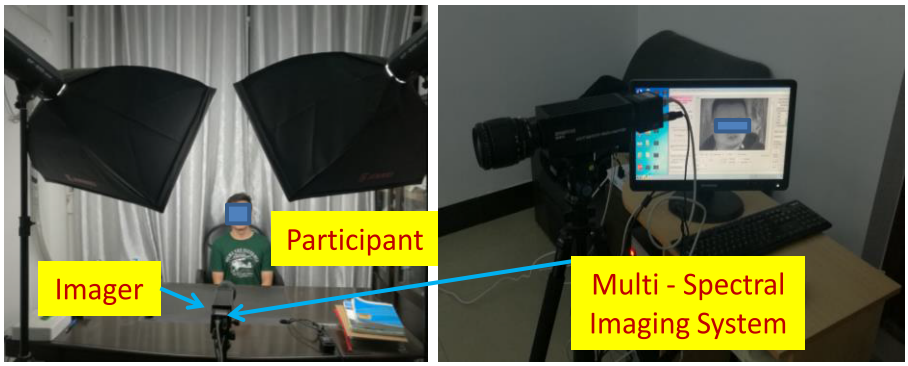

FIGURE 7. Trial environment. Experimental condition of the participants (left) and our MSI system (right).

Various kinds of stimuli have been used in emotion research. Existing studies have evaluated the reliability and efficiency of film clips in eliciting emotions [62], [63]. Emotional films contain scenes and audio and can expose subjects to real-life scenarios and elicit strong subjective and physiological changes. Each emotional film lasts for about 25 minutes. Our experiment involved five categories of emotions of the 250 subjects, and such emotions were elicited by showing emotional film clips to the participants. The captured MSI data cube included 6000 image sequences with five emotion labels, namely, anger, fear, happiness, sadness, and surprise. In this database, each sequence starts with a neutral emotion, followed by the peak of the emotion, and ends with a baseline.

The experiments were conducted through the following steps. The participants were asked to wear a heart rate monitor (Garmin) and led to a well-lit room where they comfortably sat down. A resting period of approximately 5 min was provided to allow the participants to settle in their environment. The participants were then asked to watch a film to elicit emotions. The MSI images of the participants were simultaneously recorded using an imaging system. Figure 7 illustrates the experimental environment. Seventy percent of the image sequence in data collection was used for training, and the remaining 30% was used for testing.

The data for the training and testing were randomly obtained from the data collections. For our SACNN model, we first used a clustering method to obtain the EIDC dataset. In a later study, we still used the block framework (Figure 5). Each block maintains a five-layer structure, and the growth rate (GR) of the FM was set to 24 and 32. The depths were set to 40, 100, 190, and 250. For convolutional layers with a kernel size of 3×3, each side of the input was zero-padded by one pixel to fix the feature map size.

The experimental results were compared in many ways. First, as shown in Table 1, we divided the input features of our emotion recognition model into the following situations: MSI, MSI + StO2, and MSI + StO2 + Clustering. This setting was designed to demonstrate the influence of the input feature on the results of our algorithm. We placed different input features into our SACNN model for emotion recognition calculation (Table 1). The sole input MSI feature can obtain a correct rate of only 69.3%, but when the emotionally sensitive StO2 was introduced, the rate increased to 84.7%. The EIDC achieved the highest accuracy rate of 90.04%. Our data preprocessing model showed a positive response to the experimental results.

TABLE 1. Comparison of the average accuracy of different input features calculated using the SACNN model.

Input Feature	Method	Accuracy (%)
MSI	SACNN	69.30
MSI+StO2	SACNN	84.70
MSI+StO2+Clustering (EIDC)	SACNN	90.04

These experiments reveal the positive effect of the pretreatment steps in our SACNN structure on the experimental results. Our model is an improvement based on DenseNet. To determine the effect of our model, we provided both models with the same input and parameter setting to test the correction rate. We then evaluated our model with different depths and growth rates on the emotion recognition task and compared it with state-of-the-art DenseNet architectures. We experimented on the SACNN structure with different depths and GR. SACNN and DenseNet were trained through stochastic gradient descent. The EIDC signal was used as the input for the DenseNet and SACNN models. Tables 2 and 3 summarize the experimental results. In comparison with DenseNet, our SACNN algorithm progressed in terms of accuracy. Regardless of the parameter setup, the accuracy rate of the SACNN model increased by 2%–3% compared with that of the traditional DenseNet algorithm. When used to classify emotion, the accuracy of our algorithm exceeded 90%. Although our algorithm has achieved progress in terms of recognition rate, the computation time increased by around 7% compared with that of DenseNet. Our next work will focus on reducing our algorithm's computation time.

TABLE 2. Comparison of the average accuracy obtained using the SACNN model and the DenseNet algorithm corresponding to different depth settings at a GR of 24. The EIDC signal is used as the input feature.

Depth	GR	DenseNet Accuracy (%)	SACNN Accuracy (%)
40	24	81.95	85.21
100	24	83.91	86.77
190	24	85.63	88.04
250	24	86.48	89.77

TABLE 3. Comparison of the average accuracy obtained using the SACNN model and the DenseNet algorithm corresponding to different depth settings at a GR of 32. The EIDC signal is used as the input feature.

Depth	GR	DenseNet Accuracy (%)	SACNN Accuracy (%)
40	32	82.38	85.73
100	32	84.22	87.30
190	32	85.96	88.37
250	32	87.11	90.04

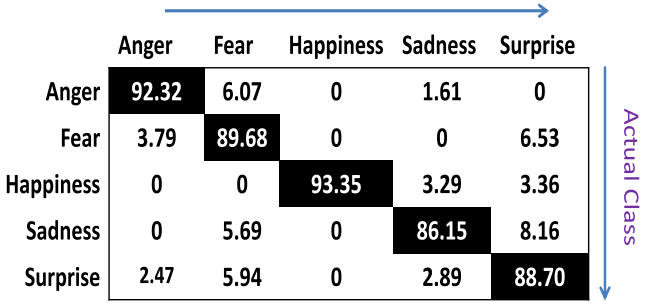

FIGURE 8. Experimental results on emotion recognition. The average recognition rate is 90.04%.

Figure 8 depicts the confusion matrix of the experimental results for emotion recognition. In general, our algorithm performed well in recognizing all types of emotion because the accuracy of each emotion was approximately 90%. Two kinds of emotions, namely, anger and happiness, were relatively easy to recognize and had recognition accuracies of 92.3% and 93.4%, respectively. Fear, sadness, and surprise obtained correction rates of 89.7%, 86.1%, and 88.7%, respectively. Relatively high confusion appeared among three pairs of emotions: fear versus anger, fear versus sadness, and surprise versus sadness. The average accuracy of the entire algorithm exceeded 90%.

Finally, we compared our algorithm structure with recognition algorithms from a similar category to verify the merits of our algorithm. In this study, the StO2 signal was applied for the first time as a feature to recognize emotion. Thus, no specific similar algorithm was found. Accordingly, we used several extensively applied famous algorithms as comparison objects and employed the MSI-extracted StO2 feature as the input [74]–[81]. Our SACNN model achieved better accuracy than other famous algorithms (Table 4), thereby further illustrating the advantages of our algorithm.

TABLE 4. Comparison of the experimental results of SACNN and other famous algorithms.

Methods	Accuracy (%)
SVM	67.13
KNN	60.50
BP	63.90
Bayes	70.22
Decision Tree	58.94
GAN	79.37
Ensemble Learning	66.89
Probabilistic NN	71.42
SACNN	90.04

V. CONCLUSION

This study is the first to use the SACNN algorithm in recognizing human emotional states. Compared with other methods (especially facial recognition), our technical method does not involve judging the emotional state based on facial expressions, although we also take images of faces. We use real-time MSI technology to extract the content of StO2 at the human nose as an ROI feature and recognize human emotions. Thus, our method has considerable advantages over other methods in terms of data source because our data are not static but real-time, and the content of StO2 involves objective data and cannot be changed artificially. This feature differs considerably from an expression because an expression is easy to disguise.

For our SACNN algorithm structure, we use spatial and spectral signal features for MSI to extract StO2 signals and integrate the spectral features into the cluster algorithm structure. Clustering initializes zones with the same emotional properties or the same spectral signal base to provide a basis for the next training in the same data cube. The most common clustering method entails calculating the ''distance'' between spatial data. The proposed clustering model structure can further use the correlation structure of the spectral and StO2 signals. We also consider the data correlation in the entire process. The algorithm divides the ROI (nose) into multiple data cubes (EIDCs) in accordance with spatial and spectral features. Then, the signal pretreating process is completed from multispectral images to EIDCs. Subsequently, we incorporate the timing EIDC sequence into our learning models and LSTM to extract the emotion feature.

Compared with the traditional DenseNet structure, the SACNN structure focuses on the data's spectral correlation and spatial–structural correlation. We reconstruct and strengthen the correlation between layers in SACNN through

the adjustment method. The features extracted through our algorithm are consistent with our demand for spectral and spatial features. In this study, we use DenseNet as our basic framework in training and learning. In the dense block of DenseNet, the current layer is directly linked to the FM of all previous ones, and this structure can efficiently improve gradient disappearance. However, we assume that considering the layers' relationship and differences will generate additional possibilities for extracting recognition features. We posit that for the layers' relationship and difference, several clues can be obtained from the correlation and difference between the spectral wavelength and StO2 signals.

We start with our deep learning model at different layers to explore the correlation and difference of the spatial–spectral signals, with the connection between layers as our cutting point. The correction value or weight structure given in the process of adjustment is utilized to reconstruct the relationship and weight of all previous FMs (i.e., the input of the current layer). We handle all FMs in the previous layers through the adjustment to ensure that the spectral–spatial features are fully explored. In our case, the adjustment involves FMs from previous layers. Through the adjustment, we can rearrange the spectral and spatial correlations, the importance of all previous layers' features, and the impact on the subsequent training. Afterward, LSTM is introduced to further strengthen the connection and feature extraction of time-series data. Finally, our SACNN algorithm is used to extract the corresponding emotion recognition features in the spectral–spatial–temporal signals. The algorithm achieves an encouraging accuracy rate in emotional recognition.

Meanwhile, our research has a methodological limitation that must be addressed. Our algorithm is based on DenseNet but is more complex than DenseNet. A refined model for reducing the computation time, in combination with the application DEMO in the HCI field, should be developed in future research.

In summary, this study is the first to use objective data and signals that cannot be artificially altered, such as StO2 content, to recognize human emotions. The accuracy of our algorithm exceeds 90%. Our future study will focus on improving the accuracy of the algorithm, and we will exert extra efforts to eliminate external interference parameters and reduce the computation time. Furthermore, we will consider a small-scale, inexpensive industrialization of our study. Thus, the customized recognition system could be applied in the industry in the future.

REFERENCES

[1] A. M. Al-Kaysi, A. Al-Ani, C. K. Loo, T. Y. Powell, D. M. Martin, M. Breakspear, and T. W. Boonstra, "Predicting tDCS treatment outcomes of patients with major depressive disorder using automated EEG classification," *J. Affect. Disorders*, vol. 208, pp. 597–603, Jan. 2017.

[2] A. V. Bocharov, G. G. Knyazev, and A. N. Savostyanov, "Depression and implicit emotion processing: An EEG study," *J. Clin. Exp. Neuropsychol.*, vol. 1, no. 47, pp. 225–230, 2017.

[3] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 529–545, Mar. 2017.

[4] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[5] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[6] L. Chen, M. Wu, M. Zhou, Z. Liu, J. She, and K. Hirota, "Dynamic emotion understanding in human-robot interaction based on two-layer fuzzy SVR-TS model," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 50, no. 2, pp. 490–501, Feb. 2020.

[7] S. Deb and S. Dandapat, "Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 802–815, Mar. 2019.

[8] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.

[9] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, Apr. 2019.

[10] Z. Yang and S. S. Narayanan, "Modeling dynamics of expressive body gestures in dyadic interactions," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 369–381, Jul. 2017.

[11] Z. Wang, S. Y. M. Lee, S. Li, and G. Zhou, "Emotion analysis in code-switching text with joint factor graph model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 469–480, Mar. 2017.

[12] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[13] R. Martinez, A. Salazar-Ramirez, A. Arruti, E. Irigoyen, J. I. Martin, and J. Muguerza, "A self-paced relaxation response detection system based on galvanic skin response analysis," *IEEE Access*, vol. 7, pp. 43730–43741, 2019.

[14] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Long short term memory hyperparameter optimization for a neural network based emotion recognition framework," *IEEE Access*, vol. 6, pp. 49325–49338, 2018.

[15] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: Taxonomy, review and current challenges," *Sensors*, vol. 20, no. 8, pp. 2384–2394, 2020.

[16] I. Pavidis, N. L. Eberhardt, and J. A. Levine, "Human behavior: Seeing through the face of deception," *Nature*, vol. 415, no. 6867, pp. 35–36, 2002.

[17] I. Pavlidis, J. Levine, and P. Baukol, "Thermal image analysis for anxiety detection," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, no. 1, Oct. 2001, pp. 315–318.

[18] I. Pavlidis, I. Garza, P. Tsiamyrtzis, M. Dcosta, J. W. Swanson, T. Krouskop, and J. A. Levine, "Dynamic quantification of migrainous thermal facial patterns—A pilot study," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1225–1233, May 2019.

[19] S. Ioannou, S. Ebisch, T. Aureli, D. Bafunno, H. A. Ioannides, D. Cardone, B. Manini, G. L. Romani, V. Gallese, and A. Merla, "The autonomic signature of guilt in children: A thermal infrared imaging study," *PLoS One*, vol. 8, no. 11, Nov. 2013, Art. no. e79440.

[20] C. Puri, L. Olson, I. Pavlidis, J. Levine, and J. Starren, "StressCam: Non-contact measurement of users' emotional states through thermal imaging," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst. (CHI)*, vol. 2, no. 1, 2005, pp. 1725–1728.

[21] M. Garbey, N. Sun, A. Merla, and I. Pavlidis, "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 8, pp. 1418–1426, Aug. 2007.

[22] D. Shastri, A. Merla, P. Tsiamyrtzis, and I. Pavlidis, "Imaging facial signs of neurophysiological responses," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 2, pp. 477–484, Feb. 2009.

[23] K. Hong and S. Hong, "Real-time stress assessment using thermal imaging," *Vis. Comput.*, vol. 32, no. 1, pp. 1369–1377, 2015.

[24] K. Hong, G. Liu, W. Chen, and S. Hong, "Classification of the emotional stress and physical stress using signal magnification and canonical correlation analysis," *Pattern Recognit.*, vol. 77, pp. 140–149, May 2018.

[25] I. Pavlidis, J. Dowdall, N. Sun, C. Puri, J. Fei, and M. Garbey, "Interacting with human physiology," *Comput. Vis. Image Understand.*, vol. 108, nos. 1–2, pp. 150–170, Oct. 2007.

[26] S. J. Ebisch, T. Aureli, D. Bafunno, D. Cardone, G. L. Romani, and A. Merla, "Mother and child in synchrony: Thermal facial imprints of autonomic contagion," *Biol. Psychol.*, vol. 89, no. 1, pp. 123–129, Jan. 2012.
[27] S. Ioannou, S. Ebisch, T. Aureli, D. Bafunno, H. A. Ioannides, D. Cardone, B. Manini, G. L. Romani, V. Gallese, and A. Merla, "The autonomic signature of guilt in children: A thermal infrared imaging study," *PloS One* vol. 8, no. 3, pp. 1–11, 2013.
[28] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–8, Aug. 2012.
[29] S. J. Ebisch, T. Aureli, D. Bafunno, D. Cardone, B. Manini, S. Ioannou, G. L. Romani, V. Gallese, and A. Merla, "Facial imprints of autonomic contagion in mother and child: A thermal imaging study," in *Proc. Appendix 1 Thermol. Int. 22/3 EAT Book*, 2012.
[30] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.
[31] K. L. Calvin and V. G. Duffy, "Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement," *Occupational Ergonom.*, vol. 7, no. 2, pp. 83–94, 2007.
[32] A. Savran, H. Cao, A. Nenkova, and R. Verma, "Temporal Bayesian fusion for affect sensing: Combining video, audio, and lexical modalities," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1927–1941, Sep. 2015.
[33] S. Jarlier, D. Grandjean, S. Delplanque, K. N'Diaye, I. Cayeux, M. I. Velazco, D. Sander, P. Vuilleumier, and K. R. Scherer, "Thermal analysis of facial muscles contractions," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 2–9, Jan. 2011.
[34] B. Manini, D. Cardone, S. J. H. Ebisch, D. Bafunno, T. Aureli, and A. Merla, "Mom feels what her child feels: Thermal signatures of vicarious autonomic response while watching children in a stressful situation," *Frontiers Hum. Neurosci.*, vol. 7, pp. 1–10, Jun. 2013.
[35] N. Alioua, A. Amine, and M. Rziza, "Driver's fatigue detection based on yawning extraction," *Int. J. Veh. Technol.*, vol. 10, no. 8, pp. 1–7, 2014.
[36] M. Sacco and R. A. Farrugia, "Driver fatigue monitoring system using support vector machines," in *Proc. 5th Int. Symp. Commun., Control Signal Process.*, May 2012, pp. 1–5.
[37] A. Liu, Z. Li, L. Wang, and Y. Zhao, "A practical driver fatigue detection algorithm based on eye state," in *Proc. Asia Pacific Conf. Postgraduate Res. Microelectron. Electron. (PrimeAsia)*, Shanghai, China, Sep. 2010, pp. 163–172.
[38] D. Liu, P. Sun, Y. Xiao, and Y. Yin, "Drowsiness detection based on eyelid movement," in *Proc. 2nd Int. Workshop Edu. Technol. Comput. Sci.*, Wuhan, China, 2010, pp. 276–277.
[39] J. Jimenez-Pinto and M. Torres-Torriti, "Face salient points and eyes tracking for robust drowsiness detection," *Robotica*, vol. 30, no. 5, pp. 731–741, 2012.
[40] R. Irani, K. Nasrollahi, and T. B. Moeslund, "Contactless measurement of muscles fatigue by tracking facial feature points in a video," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 127–135.
[41] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund, "Facial video-based detection of physical fatigue for maximal muscle activity," *IET Comput. Vis.*, vol. 10, no. 4, pp. 323–330, 2016.
[42] H.-R. Kim, Y.-S. Kim, S. J. Kim, and I.-K. Lee, "Building emotional machines: Recognizing image emotions through deep neural networks," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2980–2992, Nov. 2018.
[43] P. M. Ferreira, F. Marques, J. S. Cardoso, and A. Rebelo, "Physiological inspired deep neural networks for emotion recognition," *IEEE Access*, vol. 6, pp. 53930–53943, 2018.
[44] N. Neittaanmäki-Perttu, M. Grönroos, T. Tani, I. Pölönen, A. Ranki, O. Saksela, and E. Snellman, "Detecting field cancerization using a hyperspectral imaging system," *Lasers Surg. Med.*, vol. 45, no. 7, pp. 410–417, Sep. 2013.
[45] H. Fabelo *et al.*, "*In-vivo* hyperspectral human brain image database for brain cancer detection," *IEEE Access*, vol. 7, pp. 39098–39116, 2019.
[46] R. Pike, G. Lu, D. Wang, Z. G. Chen, and B. Fei, "A minimum spanning forest-based method for noninvasive cancer detection with hyperspectral imaging," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 653–663, Mar. 2016.
[47] T. Chen, P. W. T. Yuen, K. Hong, I. Ibrahim, A. Tsitiridis, U. Soori, J. Jackman, D. James, and M. Richardson, "Assessment of tissue blood perfusion in-vitro using hyperspectral and thermal imaging techniques," in *Proc. 5th Int. Conf. Bioinf. Biomed. Eng.*, May 2011, pp. 166–178.
[48] T. Chen, P. Yuen, K. Hong, A. Tsitiridis, F. Kam, J. Jackman, D. James, M. Richardson, W. Oxford, J. Piper, F. Thomas, and S. Lightman, "Remote sensing of stress using electro-optics imaging technique," *Proc. SPIE*, vol. 7486, pp. 601–612, Sep. 2009.
[49] P. Yuen, T. Chen, K. Hong, A. Tsitiridis, F. Kam, J. Jackman, D. James, M. Richardson, L. Williams, W. Oxford, J. Piper, F. Thomas, and S. Lightman, "Remote detection of stress using hyperspectral imaging technique," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention (ICDP)*, 2009, pp. 1221–1222.
[50] K. Hong, X. Liu, G. Liu, and W. Chen, "Detection of physical stress using multispectral imaging," *Neurocomputing*, vol. 329, pp. 116–128, Feb. 2019.
[51] X. Li, K. Hong, and G. Liu, "Detection of physical stress using facial muscle activity," *J. Opt. Technol.*, vol. 85, no. 9, pp. 562–569, 2018.
[52] K. Hong, "Classification of emotional stress and physical stress using facial imaging features," *J. Opt. Technol.*, vol. 83, no. 8, pp. 508–512, 2016.
[53] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[54] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[55] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[56] Z. Wang, L. Zheng, Y. Li, and S. Wang, "Linkage based face clustering via graph convolution network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2312–2322.
[57] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. NIPS*, 2015, pp. 2377–2385.
[58] G. Larsson, M. Maire, and G. Shakhnarovich, "FractalNet: Ultra-deep neural networks without residuals," in *Proc. ICLR*, 2016, pp. 1605–1615.
[59] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. ECCV*, 2016, pp. 646–661.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 278–289.
[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1562–1572.
[62] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition Emotion*, vol. 9, no. 1, pp. 87–108, 1995.
[63] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition Emotion*, vol. 24, no. 7, pp. 1153–1172, Nov. 2010.
[64] K. J. Zuzak, M. D. Schaeberle, E. N. Lewis, and I. W. Levin, "Visible reflectance hyperspectral imaging: Characterization of a noninvasive, *in vivo* system for determining tissue perfusion," *Anal. Chem.*, vol. 74, no. 9, pp. 2021–2028, May 2002.
[65] D. Yudovsky, L. Pilon, A. Nouvong, and K. Schomacker, "Assessing diabetic foot ulcer development risk with hyperspectral tissue oximetry," *J. Biomed. Opt.*, vol. 16, no. 2, pp. 260–269, 2011.
[66] L. C. Cancio, A. I. Batchinsky, J. R. Mansfield, S. Panasyuk, K. Hetz, D. Martini, B. S. Jordan, B. Tracey, and J. E. Freeman, "Hyperspectral imaging: A new approach to the diagnosis of hemorrhagic shock," *J. Trauma, Injury, Infection, Crit. Care*, vol. 60, no. 5, pp. 1087–1095, May 2006.
[67] D. Shastri, M. Papadakis, P. Tsiamyrtzis, B. Bass, and I. Pavlidis, "Perinasal imaging of physiological stress and its affective potential," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 366–378, Jul. 2012.
[68] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[69] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Stat. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, Mar. 2011.
[70] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[71] D. R. Li, G. Zhang, W. Jiang, and X. Yuan, "SPOT-5 HRS satellite imagery block adjustment without GCPS or with single GCP," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 31, no. 5, pp. 377–381, 2006.

[72] M. Wang, B. Yang, D. R. Li, J. Y. Gong, and Y. D. Pi, "Technologies and applications of block adjustment without control for ZY-3 images covering China," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 42, no. 4, pp. 427–433, 2017.
[73] C. Xizhang, "Generalized surveying adjustment," 2nd ed., Wuhan Univ. Surveying Mapping Press, Wuhan, China, Tech. Rep., 2000.
[74] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27–37, 2011.
[75] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Aug. 1985.
[76] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
[77] Z. Li and B. D'Ambrosio, "Efficient inference in bayes networks as a combinatorial optimization problem," *Int. J. Approx. Reasoning*, vol. 11, no. 1, pp. 55–81, Jul. 1994.
[78] B. Chandra, S. Mazumdar, V. Arena, and N. Parimi, "Elegant decision tree algorithm for classification in data mining," in *Proc. 3rd Int. Conf. Web Inf. Syst. Eng. (Workshops)*, Dec. 2002, pp. 160–169.
[79] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 2, no. 1, 2014, pp. 2672–2680.
[80] I. H. Witten, *Data Mining*, 4th ed. San Francisco, CA, USA: Margan Kaufmann, 2017, pp. 479–501.
[81] A. Zaknich, C. J. S. deSilva, and Y. Attikiouzel, "A modified probabilistic neural network (PNN) for nonlinear time series analysis," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Nov. 1991, pp. 227–237.

KAN HONG was born in China. He received the Ph.D. degree in computer vision, in 2013. He is currently a Lecturer with the Jiangxi University of Finance and Economics. His research interests include spectral vision, machine learning, signal processing, and image processing.