Electronic Research Archive

ERA, 32(5): 3477−3521.
DOI: 10.3934/era.2024161
Received: 07 February 2024
Revised: 19 March 2024
Accepted: 15 May 2024
Published: 29 May 2024

https://www.aimspress.com/journal/ERA

Review

Trusted emotion recognition based on multiple signals captured from video and its application in intelligent education

Junjie Zhang[1], Cheng Fei[1,*], Yaqian Zheng[2], Kun Zheng[3,*], Mazhar Sarah[3,4] and Yu Li[3]

[1] Smart Learning Institute, Beijing Normal University, Beijing 100875, China
[2] School of Educational Technology, Beijing Normal University, Beijing 100875, China
[3] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[4] Faculty of Engineering and Computing, National University of Modern Languages, Islamabad 44000, Pakistan

* **Correspondence:** Email: feicheng@bnu.edu.cn, zhengkun@bjut.edu.cn; Tel: +8618311053727, +8618911299525.

Abstract: The emotional variation can reflect shifts in mental and emotional states. It plays an important role in the field of intelligent education. Emotion recognition can be used as cues for teachers to evaluate the learning state, analyze learning motivation, interest, and efficiency. Although research on emotion recognition has been ongoing for a long time, there has been a restricted emphasis on analyzing the credibility of the recognized emotions. In this paper, the origin, development, and application of emotion recognition were introduced. Then, multiple signals captured from video that could reflect emotion changes were described in detail and their advantages and disadvantages were discussed. Moreover, a comprehensive summary of the pertinent applications and research endeavors of emotion recognition technology in the field of education was provided. Last, the trend of emotion recognition in the field of education was given.

Keywords: artificial intelligence; intelligent education; trusted emotion recognition; multiple signals

1. Introduction

Compared to traditional offline education, e-learning offers greater flexibility and convenience. Students can access learning resources anytime and anywhere through the internet. They are not

constrained by time or place, enabling them to study at their own pace and convenience. In addition, e-learning breaks down geographical barriers, which provides students with global learning opportunities. They can participate in international online courses, interact with other students and access educational resources from different cultures. The combination of e-learning and offline education is an effective way to achieve education equity [1]. However, it is essential to acknowledge that e-learning encounters challenges, including technological demands and the lack of face-to-face interaction. More importantly, teachers find it challenging to ascertain the genuineness of the emotional states conveyed by students [2].

Studies have demonstrated that emotions not only provide teachers with valuable insights into students' learning status but also significantly enhance their learning outcomes [3]. According to Yerkes-Dodson's law, solving complex algebra problems becomes feasible when emotions are maintained at a calm level. When emotions are at a moderately agitated level, they help to solve simple mathematical arithmetic problems. When emotions are at a higher level of agitation, such as anger, it is difficult to solve difficult problems. Therefore, it is important to recognize students' trusted emotions. In contrast to traditional offline education, the primary challenge in e-learning lies in the absence of emotional communication, potentially resulting in less favorable learning outcomes.Emotion arises from the dynamic interplay between internal feelings and the external environment. Physiological signals, such as EEG [4], ECG [5], and EMG [6], can truly reflect emotional changes. Yet, utilizing professional equipment for collecting physiological signals in e-learning isn't feasible due to its exorbitant cost, limiting its widespread applicability. However, it is not suitable to collect physiological signals by wearing professional equipment for e-learning, because professional equipment is extremely expensive, and it cannot widen the scope of application. More importantly, it may affect students' learning state. To broaden the application scenario and reduce the interference, videos collected during the learning process can be used to analyze student's learning state, i.e., facial expressions and micro-expression [7], eye states [8], gaze points [9], head posture [10], and non-contact physiological signals [11]. These states can be extracted from videos, that can further be used as cues to monitor students' learning states. Compared with other methods, videos can senselessly record learning state, which can provide a true, natural, and objective reflection of the student's emotional state.

2. The development of trusted emotion recognition in the field of education

2.1. The development of trusted emotion recognition

In 1997, Professor Rosalind Picard of the Massachusetts Institute of Technology (MIT) introduced the concept of affective computing, which emphasizes the connection between computing and emotions [12]. In the 1990s, Japan pioneered the research of Kansei engineering by combining aesthetics with engineering. Kansei engineering aims to enhance user satisfaction by incorporating human emotional needs into the process of product design and manufacturing. During the same period, countries and companies in Europe also started research work on facial expression recognition, measurement of emotional information and wearable computing, such as the Emotion Research laboratory led by Klaus Soberer at the University of Geneva and the Emotion Robotics Research Group led by Canamero at the Free University of Brussels. In terms of market applications, a multi-model shopping assistant based on the EMBASSI system was proposed in 2001 with the

support of the German Ministry of Education and Research as well as the participation of more than 20 universities and companies. This shopping assistant will take into account the psychological and environmental needs of consumers.

2.2. The development of trusted emotion recognition in education

The research focus on affective computing is to capture signals generated by physiological indicators or behavioral characteristics using various types of sensors. After signals are obtained, models can be established. The signals include facial expressions, micro-expressions, speech, body gestures, hand gestures, electrocardiograms, electroencephalograms, etc. Affective computing has been applied in many fields. In the field of driving, affective computing can reduce accidents by alerting drivers when they are not concentrating or are fatigued [13]. In the field of e-commerce, affective computing automatically senses the user's purchase intention and makes accurate recommendations [14]. In the medical field, it detects possible psychological and mental anomalies by analyzing the changes in patients' emotional and psychological states to provide assistance to doctors' diagnosis [15]. Emotional fluctuation can reflect changes in learners' psychology as well as state. Emotional fluctuation refers to the changes or fluctuations in students' emotional states over a period of time. In the educational context, students' emotions can be influenced by various factors such as the difficulty of the learning content, the effectiveness of teaching methods, and individual factors. Integrating artificial intelligence with education enables the analysis of students' emotional fluctuations during lessons, facilitating teachers in promptly adapting teaching methods, stragtegies and content. This, in turn, enhances students' learning effectiveness and outcomes. Teaching adjustment involves adapting and modifying teaching strategies, content, and support based on students' emotional fluctuation and learning needs. Understanding students' emotional fluctuation is crucial for teachers because emotions can impact students' learning experiences and outcomes. By observing and analyzing students' emotional fluctuation, teachers can make appropriate teaching adjustments to better meet students' needs and facilitate effective learning.

In the field of education, interviews and questionnaires are two commonly used methods to detect student's emotions. Interviews are used to ask students face-to-face questions about a set of relevant issues and the results are analyzed to identify students' emotions. Interviews are the simplest form of emotion recognition. Although this method is relatively simple, it has the following problems. First, students may try to hide their true emotions. Second, emotion is characterized by its temporary nature. As students need to evaluate after the lesson, disengagement from the teaching situation can lead to emotional forgetting, and it may distort evaluation results. Questionnaires are another subjective method of emotion recognition. Participants are asked to complete a questionnaire designed to collect the emotional state of the participants and emotion recognition is achieved by analyzing the results of the questions [16]. The Academic Emotions Scale for University Students (AEQ) is widely used for questionnaire-based emotion recognition, which combines activation and validity dimensions to determine the intensity of continuous emotions. The questionnaire-based emotion recognition method has the following shortcomings. Initially, there is a concern regarding the reliability of the survey outcomes. Additionally, the quantity of questionnaire items influences the analysis findings. When there are fewer questions, it does not always reflect the results of emotion recognition. When there are more questions, students get bored, which will affect the final results. Third, the recovery rate of the questionnaire is not guaranteed. Therefore, the traditional method of

emotion recognition based on interviews and questionnaires cannot reflect the trusted emotions.

As compared to the traditional methods, physiological recognition can truly reflect emotional fluctuations, which is widely used in the field of emotion recognition [17−20]. However, professional equipment should be used to collect physiological signals, which may not only affect the learner's state but also increase the cost. Thus, it can only be used in laboratory environments. Except for physiological information, speech [21], textual [22], and body [23] information has also been used for emotion recognition tasks. However, it is difficult to obtain much speech and body movement data during the e-learning.

3. Trusted emotion recognition based on video

Established multimodal emotion recognition methods need to incorporate multiple types of state data, including image, audio, text, and physiological state. In order to establish a multimodal emotion recognition model, different devices need to be utilized to obtain raw data, which may increase the cost. In addition, there is a difference in sampling frequency between different collection devices. Therefore, complex preprocessing operations are required for the alignment of the raw data. The accuracy of data alignment affects the results of emotion recognition.

Unlike emotion recognition based on multi-source state data, emotion recognition based on video data can be achieved at a low cost. Second, only video is used, which does not require complex data alignment operations. Finally, video can record student's state during the learning process.

Image, speech, and non-contact physiological signals can be obtained from video, which can reflect emotional changes. Furthermore, different state data can be aligned under the same equipment and the complementarity between different state data can be used to improve the accuracy of emotion recognition. The state data mainly includes a one-dimensional time series and a two-dimensional image series. The one-dimensional time series includes speech, head posture, non-contact physiological signals, eye states, and so on. The one-dimensional time series can be defined as:

$$y = f_1(t) \tag{1}$$

where t represents time, y represents the signal amplitude.

The two-dimensional image series includes facial expression and micro-expression, which can be defined as:

$$z = f_2(x, y) \tag{2}$$

where x, y represent the row and column of the image, and z represents the pixel value at (x, y).

In the following section, we will summarize the relevant state data that can be extracted from videos and analyze the advantages and limitations of performing authentic emotion recognition based on different types of state data.

3.1. Trusted emotion recognition based on facial expressions

3.1.1. Facial expression recognition database

Facial expression recognition database can be divided into three types, i.e., images collected from laboratory conditions, images collected from the internet, and images collected from specific

environments. The commonly used facial expression recognition databases are listed in Table 1. Images collected from the laboratory have good quality due to the control of external factors such as lighting conditions and posture. Images collected from the internet have rich types comprising different poses, occlusions, and lighting conditions. In addition, there are databases for specific application environments. For example, the NIR KMU-FED database is used for drivers' emotion recognition.

Table 1. Commonly used facial expression recognition databases.

Database	Category	Resolution	Type	Quantity	Collection environment
NIR KMU-FED [24]	6	1600 × 1200	1	1106	Driving environment
MMEW [25]	7	1920 × 1080	1	300	Laboratory
FaceWarehouse [26]	20	640 × 480	2	3000	Laboratory
Radboud Faces [27]	8	1024 × 681	3	59616	Laboratory
JAFFE [28]	7	256 × 256	3	213	Laboratory
CK+ [29]	7	48 × 48	3	593	Laboratory
OULU-CASIA [30]	6	100 × 100	3	15339	Laboratory
FER2013 [31]	7	320 × 240	3	35887	Internet
AFFECTNET [32]	8	224 × 224	3	450000	Internet
RAF-DB [33]	7	48 × 48	3	35887	Internet
EmotioNet [34]	22	400 × 300	3	1000000	Internet

Note: 1 represents image sequence, 2 represents depth image, 3 represents image.

3.1.2. Facial expression recognition methods

Facial expressions are commonly used for emotion recognition tasks. In 1976, a facial action coding system (FACS) was proposed by Ekman. FACS divides the facial region into several independent and interrelated action units (AUs). By analyzing the movement characteristics of these AUs, the controlled facial regions, and the corresponding facial expressions, standard facial actions can be derived [35]. A trained human can recognize different emotions by analyzing the movements of AUs. However, manually calibrated facial units require a lot of manpower and time resources. With the development of artificial intelligence and computer vision, automated facial expression recognition has received widespread attention. Facial expression recognition can be divided into traditional methods and deep learning methods.

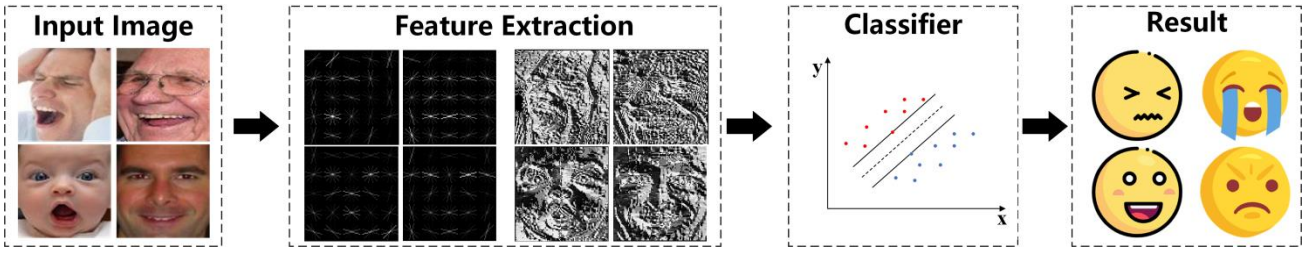


Figure 1. Facial expression recognition based on traditional features.

The flowchart of facial expression recognition methods based on traditional features is shown in

Figure 1. It is necessary to extract features from the image. Then, a classifier is selected to classify the extracted features. Traditional features include LBP [36], SIFT [37], HOG [38], and color features [39]. LBP features demonstrate resilience to variations in lighting and changes in pose Therefore, it is widely used in facial expression recognition tasks. LBP features can be expressed as:

$$LBP(x_c, y_c) = \sum_{p=0}^{P-1} 2^p s(i_p - i_c) \tag{3}$$

with x_c, y_c as central pixel with intensity i_c; and i_n being the intensity of the neighbor pixel. *S* is the sign function defined as:

$$s(x) = \begin{cases} 1 \; if \; x \geq 0 \\ 0 \qquad else \end{cases} \tag{4}$$

However, the primitive LBP features have high dimensionality, which require significant memory resources and have slow computational speed. In order to overcome the shortcomings of primitive LBP features, researchers have successively proposed double LBP [40] and Riu-LBP [36]. Double LBP and Riu-LBP are defined as (5) and (8), respectively.

$$LBP_{P,R}^{Double}(x) = \{LBP_{P,R}^{+}(x), LBP_{P,R}^{-}(x)\} \tag{5}$$

$LBP_{P,R}^{+}(x)$ and $LBP_{P,R}^{-}(x)$ can be expressed as:

$$LBP_{P,R}^{+}(x) = \sum_{p=0}^{P-1} s(g_p - g_x - n)2^p \, , s(u) = \begin{cases} 1, u > 0 \\ 0, u \leq 0 \end{cases} \tag{6}$$

$$LBP_{P,R}^{-}(x) = \sum_{p=0}^{P-1} s(g_p - g_x - n)2^p, s(u) = \begin{cases} 1, u < 0 \\ 0, u \geq 0 \end{cases} \tag{7}$$

$$LBP_{P,R}^{Riu} = \begin{cases} \sum_{p=0}^{P} s(g(p) - g(c)) \\ P + 1 \end{cases} \tag{8}$$

where *c* is the center pixel, *g()* denotes gray level of pixel and *s()* is the sign function. However, external factors such as illumination, pose change, occlusion, and image quality. can affect the results of facial expression recognition based on traditional features.

Unlike traditional features, deep learning-based methods can automatically extract features. The flowchart of facial expression recognition based on deep learning methods is shown in Figure 2. First, it is necessary to collect a large number of diverse facial expression images as inputs. To increase the robustness of the model, data augmentation strategies are applied to enrich the input images. Then, deep learning models are designed for feature extraction. Finally, the model is trained by minimizing the loss function to achieve facial expression recognition. Compared with traditional methods, deep learning-based methods can effectively overcome the influence of external factors on recognition results. In [41], sparse representation and extreme learning are combined to overcome the impact of head and lighting variations on facial expression recognition accuracy. Reference [42] proposed the IERN model and eliminated the effect of background and dataset bias on recognition accuracy from the perspective of causal inference. In [43], EASE was proposed to overcome the effect of image quality and inaccurate training data labeling on model accuracy and robustness. The model can accurately recognize emotions with ambiguity in noisy data. Reference [44] proposed AffMen model, which integrates the common features of emotions and the personality features learned in real time when performing continuous emotion recognition to reduce the impact on the accuracy of emotion recognition due to cultural background, gender and personality differences. Facial expression recognition algorithms in mask-obscured scenes is relatively low due to the problem of missing

facial information caused by face masks. To solve the problem mentioned above, TKNN was proposed in [45], which integrates eyebrow and eye state information in mask-obscured scenes. The proposed algorithm utilizes facial feature points in the eyebrow and eye regions to calculate various relative distances and angles, capturing the state information of eyebrows and eyes.

In 2020, researchers at the University of California, Berkeley, analyzed the extent to which 16 types of facial expressions occurred in thousands of situations in 6 million videos from 144 countries and territories. They discovered that 16 facial expressions appeared in similar contexts around the world. The researchers also found that each specific facial expression was uniquely associated with a set of contexts that were more similar. These backgrounds were maintained 70 percent across 12 regions. Their findings reveal a fine-grained model of human facial expressions preserved throughout the modern world [46]. Most research work is based on the study of basic discrete emotions proposed by Ekman. Facial expressions are complex, as the same expression under different intensities can reflect different emotional states, such as smile and laughter, sadness and grief. Therefore, references [47,48] investigated the problem of facial expression recognition under fine granularity and classified each basic expression into multiple categories according to its intensity. Moreover, they implemented facial expression recognition under fine granularity based on convolutional neural networks and graph neural networks, respectively. Compared to convolutional neural networks, graph neural networks can achieve better performance.

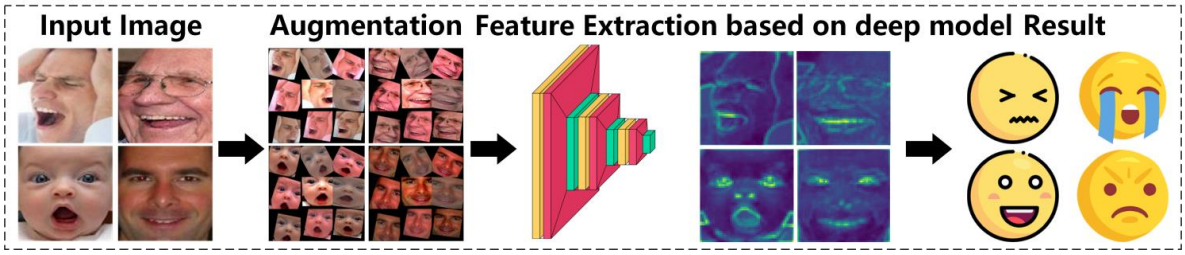


Figure 2. Facial expression recognition based on deep learning.

Although facial expressions are commonly used for emotion recognition, they cannot truly reflect a student's emotional changes, for facial expressions can be deceptive. As students attempt to hide their true emotions, relying solely on facial expressions may not accurately reflect their emotions. In addition, facial expressions are diverse, they cannot capture the difference between different cultures and backgrounds. Furthermore, facial expressions can be ambiguous, as the same expression can convey different emotions. For example, a smile can express joy and friendliness, but it can also express concealment or insincerity. Finally, existing facial expression recognition algorithms primarily focus on basic facial expressions, which may not comprehensively reflect student's emotions. Therefore, emotion recognition based on facial expression has inherent limitations.

3.2. Trusted emotion recognition based on micro-expressions

3.2.1. Micro-expression recognition database

Micro-expression can express the true emotions that people are trying to suppress and hide.

However, it is more difficult to recognize if it is untrained due to its short duration and low intensity. Therefore, most of the databases are captured using high-speed cameras in laboratory environments. The commonly used micro-expression recognition databases are listed in Table 2.

Table 2. Commonly used micro-expression recognition databases.

Database	Category	Resolution	Type	Quantity	Collection environment
SMIC [34]	3	640 × 480	100	164	Laboratory
CASME [49]	7	640 × 480 720 × 1280	60	195	Laboratory
CASME II [50]	5	640 × 480	200	247	Laboratory
CASME III [51]	7	1280 × 720	30	5902	Laboratory
SAMM [25]	8	2040 × 1088	200	159	Laboratory
SAMM Long [52]	8	2040 × 1088	200	147	Laboratory
MEVIEW [53]	7	1280 × 720	25	40	Real scenario
York DDT [54]	2	320 × 240	25	20	Laboratory
USF-HD [55]	6	720 × 1280	29.7	100	Laboratory
POLIKOVSKY [56]	6	640 × 480	200	42	Laboratory
MMEW [57]	7	1920 × 1080	90	300	Laboratory

CASME, CASMEII, and CASME III databases have been created successively by the Institute of Psychology, Chinese Academy of Sciences. CASME III database introduced depth information and physiological information for the first time to provide multi-dimensional data support for comprehensive and accurate analysis of micro-expression changes.

3.2.2. Micro-expression recognition methods

Compared to facial expressions, micro-expressions can reveal the true emotions that people are trying to hide. Micro-expression recognition methods can be divided into traditional methods and deep learning methods.

Micro-expression recognition based on traditional methods cannot obtain better performance due to the short duration and low intensity. The reason is that there are no significant differences between the features of different micro-expressions (results can be seen in Figure 3). Therefore, it is not possible to accurately identify different micro-expressions based on traditional features.

Compared with traditional algorithms, deep learning methods can effectively improve classification performance. Convolution neural networks are widely used in micro-expression recognition tasks. In [58], DTSCNN was proposed for micro-expression recognition and an accuracy of 66.67% was achieved on CASME micro-expression dataset. Reference [59] proposed a deep neural network for micro-expression intensity recognition based on temporal-spatial features. This was used to generate temporal-spatial features with distinguishability and it achieved 60.98% classification accuracy on CASME II dataset. Reference [60] proposed the LEARNet model which can effectively detect the small changes in facial muscles. Reference [61] proposed the TSCNN model, which consists of dynamic-temporal, static-spatial and local-spatial components. These components are used to extract micro-expression time-varying features, appearance and contour features, and local features, respectively. Based on the fused features, an accuracy of 80.97% was

achieved on the CASME II dataset. In [62], deep recurrent convolution neural networks were used to extract micro-expression time-varying features in terms of both appearance and shape, respectively, and an accuracy of 80.3% was achieved for the CASME II dataset. There are fewer publicly available micro-expression recognition datasets that can be used to train the deep neural networks. To solve the shortcoming mentioned above. Transfer learning was used by [63], and the results showed that the accuracy can be improved by 8% based on transfer learning.

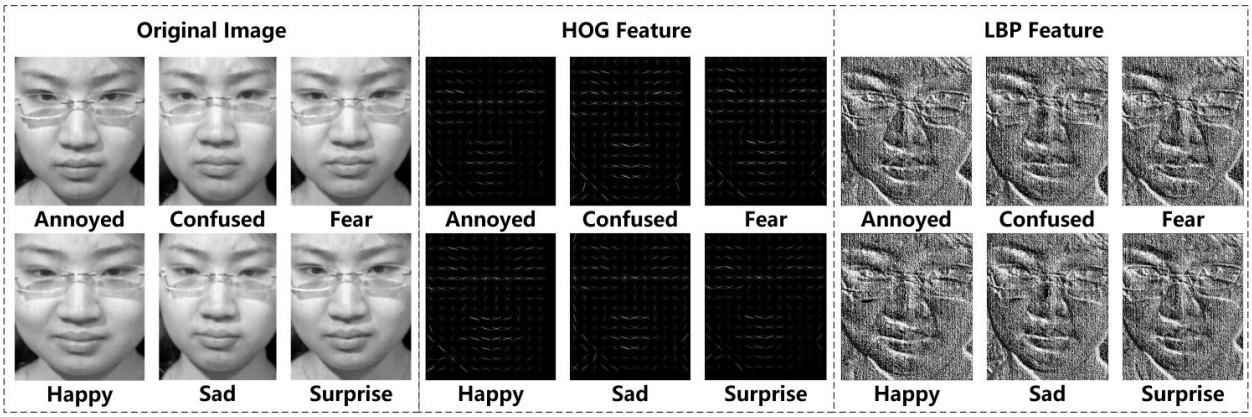


Figure 3. Feature extraction based on traditional methods for micro-expressions.

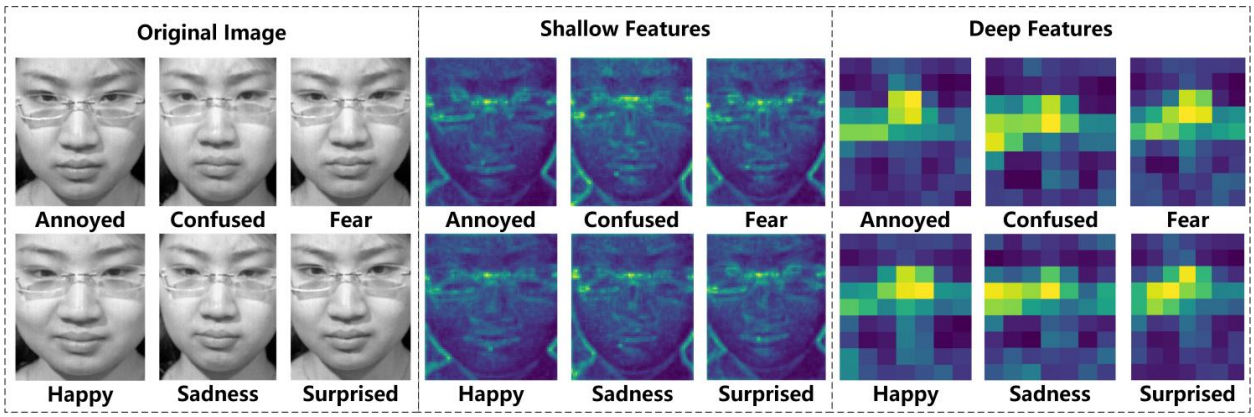


Figure 4. Feature extraction based on CNNs for micro-expressions.

The reason why CNNs are commonly used for micro-expression recognition is that time and spatial domain features can be extracted which can reflect the micro-expression changes. Feature extraction based on deep learning is shown in Figure 4. Results show that there are some differences in depth features between different micro-expressions compared to traditional features. Detailed features such as texture and contour, can be extracted at shallow layers. As the number of model layers increases, the extracted features are more discriminative. Although the performance of micro-expression recognition can be improved by CNNs, it cannot build the relationship between different features. Furthermore, a large amount of training data is required to train the model. Therefore, micro-expression recognition based on CNNs cannot obtain better performance.

Facial images reflect relatively stable local structures and local texture patterns in the images. Therefore, facial images can be effectively represented by graphs, which prossess strong processing

capabilities for non-Euclidean structures. The graph can be expressed by $G = (V, E)$, where V is the set of vertexes and E is the set of edges. An edge connects v_i and $v_j \in V$ is denoted as e_{ij}. If there is an edge connecting v_i and v_j, v_i is the neighbor of v_j, and vice versa. Assume that all neighbors of v_i are $N(v_i)$; thus, $N(v_i)$ can be given by $N(v_i) = \{v_j | \exists e_{ij} \in E \ or \ e_{ji} \in E\}$. The number of edges with v_i as the endpoint is called the degree of v_i, denoted as deg (v_i)=$|N(v_i)|$. In the graph structure, the facial feature points are represented as nodes and the distances between feature points are represented as edges. The movement of facial muscles can be represented as the movement of feature points. Different expressions cause feature points in different regions of the face to move with different tendencies. Therefore, micro-expression recognition can be achieved by modeling the motion of feature points. Micro-expression recognition based on a graph neural network is shown in Figure 5. First, a data augmentation strategy is used. Second, the graph structure is created based on landmarks. Third, feature extraction is performed based on the graph structure. Finally, micro-expression recognition is performed based on the extracted features. Micro-expression is characterized by short duration and low amplitude, as well as it requires the use of high-speed video cameras for video capture. When students use portable devices such as mobile phones, tablets, and computers for e-learning, these devices are equipped only with ordinary video cameras, which makes it difficult to accurately detect micro-expressions on videos captured by ordinary cameras.

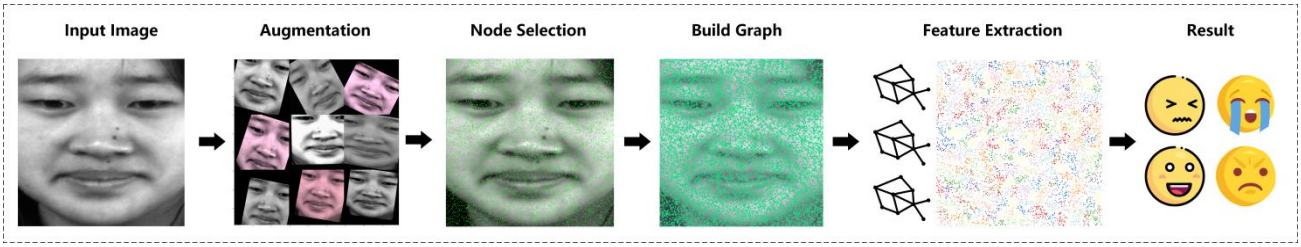


Figure 5. Feature extraction based on graph neural network for micro-expressions.

3.3. Trusted emotion recognition based on eye states

3.3.1. Eye region structure database

Eye region structure database can be divided into three types according to the auxiliary light source, that is, images collected under infrared light source, ordinary light source, and natural light source, respectively. Commonly used eye region structure databases are listed in Table 3.

The eye consists of pupil, iris and sclera. In the facial image, the eye region is small, and pupil and iris have similar colors. The variability of eye structures can be enhanced by using infrared light sources. Therefore, most of the eye structure datasets are acquired using specialized equipment with the assistance of infrared light sources. In order to improve the accuracy of eye structure segmentation, researchers have successively proposed eye structure databases based on ordinary light sources and without light source assistance.

Table 3. Commonly used eye region structure databases.

Database	Category	Resolution	Equipment	Light	Distance
CASIA.v1 [64]	320 × 280	IRI	Near-infrared camera	1	Close range
CASIA.v2 [65]	640 × 480	IRI	Near-infrared camera	1	Close range
CASIA.v3 [66]	Multi-type	IRI	Near-infrared camera	1	Close range
CASIA.v4 [67]	Multi-type	IRI	LMBS	1	3m
ICE 2005 [68]	480 × 640	IRI	LG EOU 2200	1	Close range
WVU [69]	480 × 640	IRI	OKI Iris Pass-H	1	Close range
LPW [70]	640 × 480	IRI	Head-camera	1	Close range
UPOL [71]	576 × 768	RGB	Visible light camera	2	Close range
UBIRISv1 [72]	Multi-type	RGB	Nikon E5700v1.0	2	< 50 cm
UBIRISv2 [73]	400 × 300	RGB	Ordinary camera	2	4-8m
TEyeD [74]	Multi-type	IRI	Head-camera	1	Close range

Note: 1, 2, and 3 stand for infrared light source, ordinary light source and natural light, respectively. IRI and RBG stand for infrared image and colour image, respectively.

3.3.2. Eye states recognition methods

Eye states can truly reflect emotion changes. Eye tracking is a method that can record users' real, natural, and objective behavior. By tracking the user's point of gaze and eye movements, eye tracking technology analyses the factors that attract the user's attention and emotional changes. Professional eye tracking devices are now widely used in education. The eye tracking methods can be classified as invasive and non-invasive methods.

Invasive methods were commonly used in the early stage, including direct observation method, mechanical recording method, electromagnetic induction method, and optical recording. Invasive methods cannot obtain high accuracy and it may cause some damage to the eye. Non-invasive methods can be divided into wearable and non-wearable methods. Wearable eye tracking systems require the user to wear special equipment such as helmets or glasses equipped with optical cameras. The weight of the helmet affects the ease of use. However, wearable eye tracking systems offer high accuracy. The head can be moved over a wide range. The non-wearable eye tracking system acquires the user's facial image through the camera. It analyses and extracts features from the face as well as the human eye to get the feature parameters that can reflect the changes in the line of sight. Then, the feature parameters of the human eye are transformed into three-dimensional of the line of sight through the feature parameter and mapping model, so as to estimate the direction of the line of sight and the position of the landing point. Non-wearable eye tracking systems have the advantages of low interference, easy operation, and wide applicability. Current commercial eye tracking devices include Tobii, Polhemus, ASL, SMI, and Eyelink. Commercial devices are mainly based on the pupillary corneal reflection method, which first uses an infrared light source to produce a Pulchin spot on the cornea. Then, the direction vector between the centre of the pupil and the Pulchin spot is calculated. Finally, the direction of gaze is estimated based on this direction vector. This method requires the addition of an auxiliary light source. It has high requirements for image collection equipment and the equipment is expensive. It can only be used in experimental settings and is not easily generalized. For application environments such as e-learning, how to use ordinary cameras to acquire the region of interest during students' learning process, has received the attention of relevant researchers.

Depending on the number of cameras and auxiliary light sources, gaze estimation detection methods based on ordinary cameras can be classified as single camera single light source, single camera multiple light sources, and multi-camera multiple light source methods.

The single camera single light source system needs to obtain the human eye invariant parameters in the acquired images. The hardware configuration of this system is simple. However, it requires a complex calibration process. In addition, head movement may affect the accuracy of gaze estimation. The single camera multiple light source system with multiple light sources will produce multiple Pulchin spots in the eye region. The light sources are in fixed positions. Therefore, gaze can be determined by the relative positional relationship between the Pulchin spots. The single camera multiple light source system reduces the sensitivity to head movements. It allows the head to move within a certain range while maintaining the accuracy of the line of sight. For the multiple camera multiple light source system, parameters such as the centre of curvature of the cornea and the centre of the pupil, which are relevant for the detection of the visual field, are obtained according to the principle of multi camera stereo vision. The multi camera multiple light sources improves the effects of head movement and light variations on the accuracy of gaze tracking. However, it requires a complex calibration process, which includes calibration of the light source, user position, display position, and camera. Moreover, auxiliary light can affect student's learning state and increase usage costs. Therefore, how to accurately detect gaze estimation using ordinary cameras without disturbing the user's state and increasing the usage cost has become an urgent problem [75].

The region of interest may change as learning content changes. Although it cannot be directly used for emotion recognition, it can provide supportive clues for emotion recognition, as shown in Figure 6.

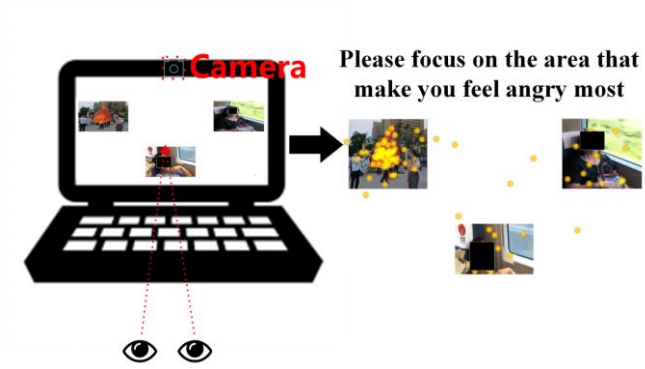


Figure 6. Emotion recognition based on gaze estimation.

Unlike gaze points, gaze duration, gaze counts, pupil size, eye closure time, and blink frequency can also be used to reflect emotion change.. However, when using ordinary cameras for eye state analysis, exogenous factors such as lighting, posture, occlusion, and image quality can affect the accuracy of eye state analysis because the eye occupies only a small part of the face region and the pupil and iris have similar colors. The current detection of the eye state requires the use of auxiliary light to improve the distinguishability of the eye region structure, and the use of professional equipment such as head-mounted cameras, infrared cameras, or depth cameras to capture high-resolution eye images to achieve the detection of the eye state. It is difficult to analyze the eye state based on the image sequence captured by ordinary cameras, as shown in Figure 7.

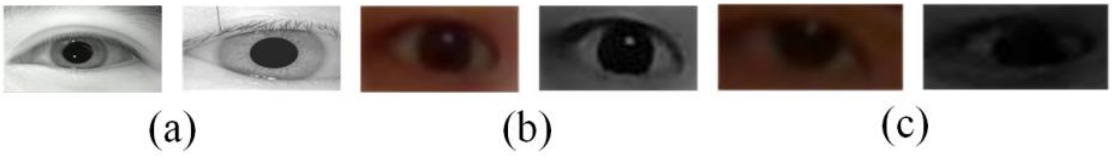


Figure 7. Eye region image captured from (a) near-infrared camera, (b) visible light camera (daylight), (c) visible light camera (night).

Moreover, when employing deep learning techniques for eye region structure segmentation, the training data's proximity to the real-world application scenario directly impacts the model's training efficacy and the accuracy of the segmentation results. However, most of the datasets (as shown in Table 3) are collected in laboratories or under controlled conditions based on professional equipment with the assistance of light sources. Therefore, the insufficiency of relevant datasets, restricts the development of deep learning-based methods for eye region structure segmentation of images captured by ordinary cameras under natural light.

3.4. Trusted emotion recognition based on non-contact physiological signals

Table 4. Commonly used non-contact physiological signals databases.

Database	Subjects	Samples	Length	Equipment (Video/BVP)
COHFACE [76]	40	160	1 min	Logitech HD C525/SA9311M
PFF [77]	13	85	3 min	---/MIO Alpha II
PURE [78]	10	60	1 min	ECO274CVGE/CMS50E
UBFC-rPPG [79]	42	42	2 min	Logitech C920HD/CMS50E
VIPL-HR [80]	107	3130	30 s	Logitech C310/CMS60C
VIPL-HR-V2 [81]	500	2500	10 s	RealSense F200/CMS60C
MMSE-HR [82]	40	102	---	FLIP A655sc/Biopac MP150
MR-NIRP-Car [83]	19	190	2−5 min	GS3U341C6NIRC/CMS50D
MR-NIRP-Indoor [84]	12	---	30 s	BFLYU323S6CC/CMS50D+
OBF [85]	106	200	5 min	SN9C201&202/NeXus-10MKII

Note: "---" means not introduced in the text.

Physiological signals such as heart rate, oxygen saturation, respiratory rate, heart rate variability, and blood pressure can reflect emotional changes. Unlike the Collection of physiological signals using specialized equipment, non-contact physiological signals can be extracted from video based on remote photoplethysmography.

3.4.1. Non-contact physiological signals database

The database of non-contact physiological signals is acquired in a laboratory environment. Heart rate information is recorded using sensors while the camera records images of the subject's face. Commonly used non-contact physiological signals databases are shown in Table 4.

3.4.2. Trusted emotion recognition method based on non-contact physiological signals

Heart rate, heart rate variability, and pulse oximeter values can be extracted from video based on rPPG, the reflection from skin recorded by the camera can be defined as a time-varying function of the color channel, which can be expressed as:

$$C_k(t) = I(t) \times (V_s(t) + V_d(t)) + V_n(t) \tag{9}$$

where $C_k(t)$ represents the kth pixel value at time t; $I(t)$ represents the illumination intensity at time t, which can be affected by the change of illumination and the distance between the illumination, the skin and the camera; $V_s(t)$, $V_d(t)$, and $V_n(t)$ represent the specular reflection, the diffuse reflection, and the noise at time t, respectively.

As light reaches the skin, a large amount of light is reflected off the skin surface and only a small amount of light enters the tissue, therefore the specular reflection does not contain pulse information, the specular reflection can be expressed as:

$$V_s(t)=u_s \times (s_0 + s(t)) \tag{10}$$

Different from the specular reflection, the concentration of hemoglobin changes periodically with the pulse of the heart, so the light intensity of menstrual hemoglobin also changes periodically, which can be used to reflect the pulse. The diffuse reflection can be expressed as:

$$V_d(t) = u_d \times d_0 + u_p \times p(t) \tag{11}$$

where $p(t)$ represents the intensity of heart rate.

Taking into account the specular reflection and the diffuse reflection, $C_k(t)$ can be expressed as

$$C_k(t)= I(t) \times (u_s \times s_0 + u_s \times s(t) + u_d \times d_0 + u_p \times p(t)) + V_n(t) \tag{12}$$

Define $u_c \times c_0 = u_s \times s_0 + u_s \times s(t)$, $I(t)= I_0 + I_0 \times i(t)$, where u_c represent the unit vector of the skin's intrinsic reflections, c_0 represent the light intensity, I_0 represent the fixed light intensity, $I_0 \times i(t)$ represents the varying light intensity, the skin reflection model can be expressed as:

$$C_k(t) = (I_0 + I_0 \times i(t)) \times (u_c \times c_0 + u_s \times s(t) + u_p \times p(t)) + V_n(t) \tag{13}$$

According to the skin reflection model, we can observe the image sequence $C_k(t)$ collected by the camera contains the pulse signal $p(t)$, and decomposition algorithm can be used to extract the pulse signal. The flowchart of non-contact physiological signals can be seen in Figure 8. First, the facial part is captured using camera. After face is detected, the region of interest is extracted from facial parts. Second, in order to obtain better performance, signal processing is conducted, including detrending, filtering, and signal decomposition, etc. Then, BVP signals are extracted from the preprocessed signal. Finally, FFT is used to transfer BVP signal from the time domain to the frequency domain and the heart rate can be calculated based on the signal frequency results.

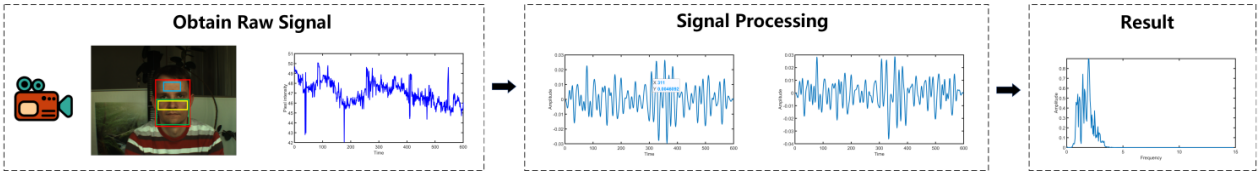


Figure 8. The flowchart of non-contact physiological detection.

Non-contact physiological signals have been applied to emotion recognition, medical care, state monitoring, biometric recognition and other fields [86,87]. In [88], non-contact physiological signals were used to recognize micro-expression. First, heart rate variability is extracted from video. Then, time-domain, frequency-domain and statistical features are extracted separately. Finally, the extracted features are fused and micro-expression recognition is performed based on the fused features. The accuracy is improved by 17.05% compared to micro-expression recognition based on facial images. In [89], pulse rate variability was extracted from videos for micro-expression recognition. CASME II micro-expression dataset was used to test the performance. Results show that an accuracy of 60% is reached. Compared to micro-expression recognition based on heart rate variability, the accuracy is improved by 0.21%. In [90], heart rate variability was extracted from the video to monitor the player's emotions. Most research work is based on forward faces for physiological signal detection. Because forward faces can stably be extracted to face-related regions. References [91] and [92] addressed heart rate detection when face information is missing in educational scenarios. The proposed method can obtain stable and accurate heart rate measurements when faces are missing.

Compared with measurement methods that require wearing sensors, such as ECG and EMG, rPPG-based measurement of physiological data does not affect the user's state. In addition, it expands the application scenarios by getting rid of the limitation of professional equipment. However, the accuracy of measurement results can be affected by various aspects such as light, posture, image resolution, region of interest, signal decomposition algorithm, length of preprocessed video, and initial value of heart rate. In addition, physiological signals are private [93]. How to accurately measure relevant physiological signals in a safe and reliable environment is the prerequisite and guarantee for emotion recognition. Although researchers have proposed more robust and more accurate algorithms [94,95] to achieve the measurement of physiological parameters under different exogenous situations. However, the accuracy needs to be further improved compared to contact methods.

3.5. Trusted emotion recognition based on speech

3.5.1. Speech emotion recognition database

The dataset of speech emotion recognition can be divided into discrete emotion recognition databases and continuous emotion recognition databases. Commonly used speech emotion recognition databases are shown in Table 5. Discrete emotion recognition databases are used to judge the category of emotion by the annotator's vote. Continuous emotion recognition databases are used to quantify the values of different dimensions with the help of MAAT or SAM system.

Table 5. Commonly used speech emotion recognition databases.

Database	Language	Samples	Participants	Emotional category
CASIA [96]	Chinese	9600	4	5 categories of discrete emotions
RAVDESS [97]	English	1440	24	8 categories of discrete emotions
RML [98]	Multiple	720	8	6 categories of discrete emotions
BAUM-1s [99]	Turkish	1222	31	8 categories of discrete emotions
IEMOCAP [100]	English	1150	10	V, A, D continuous emotions
CreativeIT [101]	English	8 h	16	V, A, D continuous emotions
VAM [102]	German	1018	47	V, A, D continuous emotions
SEMAINE [103]	German	80 h	150	V, A, D, E, I continuous emotions
RECOLA [104]	French	9.5 h	46	V, A continuous emotions

Note: V, A, D, E and I stand for Valence, Arousal, Dominance, Expectancy and Intensity respectively.

3.5.2. Speech emotion recognition methods

Speech signals are also widely used in the field of emotion recognition. The flowchart of emotion recognition based on speech signals is shown in Figure 9.

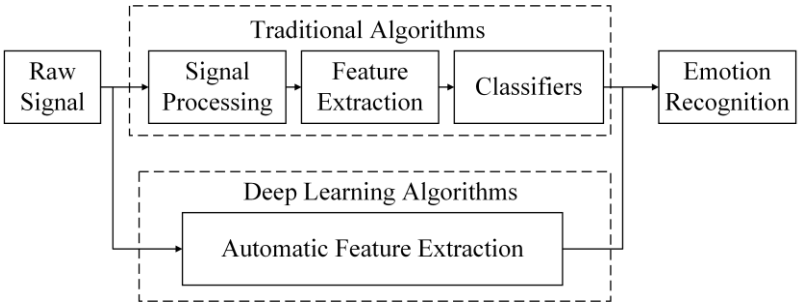


Figure 9. The flowchart of speech emotion recognition.

Speech emotion recognition algorithms can be divided into traditional algorithms and deep learning algorithms. For traditional algorithms, acoustic features should be extracted from the original signals first. Then, classifiers need to be selected to classify the extracted features. The acoustic features include rhythm, power spectrum, acoustic, non-linearity, etc. Among them MFCC features are widely used in speech recognition tasks. Defined $x(n)$ is the input speech signal. First, pre-emphasis is applied to the input signal, which enhances the higher-frequency components relative to the lower-frequency components. This can be achieved using the following formula:

$$y(n) = x(n) - \alpha x(n-1) \tag{14}$$

where *y(n)* is the pre-emphasized signal, *x(n)* is the original speech signal, and α is the pre-emphasis coefficient. Then, the pre-emphasized signal *y(n)* is divided into short frames of length *N*. Overlapping frames can be obtained using a frame shift of *M* with a desired overlap ratio R. Each frame can be represented as:

$$x_i(n) = y(n + i \times M), 0 \le i \le L - 1 \tag{15}$$

where L is the number of frames, and n is the starting position of the frame. After the process of frame segmentation, a window function $w(n)$ is applied to each frame signal. Commonly used window functions include the Hamming window. The windowed frame can be calculated as:

$$x_i(n) = x_i(n) \times w(n), 0 \le n \le N-1 \tag{16}$$

The fourth step is applying the DFT (discrete fourier transform) to each windowed frame $x_i(n)$ to obtain the frequency spectrum $x_i(k)$, and then a set of Mel filters are applied to the frequency spectrum. These filters are uniformly spaced in the Mel frequency scale. The output of each Mel filter can be computed as:

$$H_m(k) = \begin{cases} 0, k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, f(m-1) \le k < f(m) \\ 1, f(m) \le k \le f(m+1) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, f(m) < k \le f(m+1) \\ 0, k > f(m+1) \end{cases} \tag{17}$$

where m is the index of the Mel filter, k is the frequency bin index, and $f(m)$ is the center frequency of the m-th Mel filter. The center frequencies can be calculated using the inverse Mel frequency scale function:

$$f(m) = f_{mel}^{-1}(m) \tag{18}$$

The sixth step is to compress the dynamic range by taking the logarithm of the output energy from each Mel filter, it can be computed as:

$$s_i(m) = log(\sum_{k=0}^{K-1} |X_i(k)|^2 \cdot |H_m(k)|^2) \tag{19}$$

where $s_i(m)$ is the output energy of the m-th Mel filter for the i-th frame, K is the length of the spectrum. Then, DCT (discrete cosine transform) is applied to the log-compressed energy spectrum $s_i(m)$ to obtain the cepstral coefficients. The DCT can be computed as:

$$C_i(l) = \sum_{m=0}^{M-1} \left(S_i(m) \cdot cos\left[\frac{\pi}{M}(l+0.5)m \right] \right) \tag{20}$$

where $C_i(l)$ is the value of the I-th cepstral coefficient for the i-th frame, M is the desired number of cepstral coefficients to be retained. Finally, only the first few cepstral coefficients are retained as the final MFCC features.

Reference [105] extracted MFCC features in speech signals and HMM was used as a classifier for emotion recognition. Reference [106] proposed hierarchical speech recognition framework. They obtained the probability of emotion recognition in the case of SVM and GMM as classifiers based on MFCC features in the first level respectively. After that, the obtained probability values are used as inputs to the SVM classifier in the second level to get the final recognition results. Testing is performed on Emo-DB dataset. The accuracy is improved by 8% and 9% compared to single SVM and GMM classifiers, respectively. Due to the environment containing noise that affects speech recognition performance, the functional paralanguage contained in speech, although it carries a lot of emotional information, such as sighs, questioning, laughter, etc., also affects the accuracy of speech emotion recognition due to the interference of feature bursts. Reference [107] proposed a speech

emotion recognition method that fuses functional paralanguage. The functional paralanguage of the utterance to be recognized is detected by a functional paralanguage automatic detection algorithm based on fixed-length segmentation and the SPS-DM model, as well as the pure speech signal and the functional paralanguage signal, are obtained. The two types of signals are then fused using a confidence-based adaptive weight fusion method to improve the accuracy and robustness of recognition. A recognition accuracy of 67.41% is obtained for the PFSED dataset. Reference [108] performed feature extraction based on deep networks for emotional and functional paralanguage, respectively. Then, they discussed the effect of different feature fusion algorithms on the recognition results. For feature-level fusion, affective features and functional paralinguistic features are fused using splicing. The fused features are used as inputs to the Bi-LSTM model for emotion recognition. For model-level fusion, cell-coupled LSTM is used instead of Bi-LSTM. The cell-coupled LSTM contains two parts. One-part deals with affective features and the other part deals with functional paralinguistic features. For decision-level fusion, two different classifiers are trained using sentiment features and functional paralinguistic features, respectively, and the results are fused using linear weighting. Tested on the NNIME dataset, the decision-level fusion achieved the highest recognition accuracy of 61.92%.

Deep learning-based speech emotion recognition algorithm using end-to-end approach for emotion recognition can effectively extract emotion features with high recognition accuracy and robustness. Since convolution neural network and recurrent neural network can effectively perform feature extraction and capture contextual information in speech, it is widely used in deep learning based speech emotion recognition. Reference [109] used convolution neural network and time-domain pyramid matching for feature extraction of speech signals and recognized the fused features based on SVM. First, the Mel spectrogram of the speech signal is extracted and the Mel spectrogram is segmented according to the speech segments. Then, the high-level features of the Mel spectrogram of each segment are extracted using the AlexNet model. The learned high-level features are fused using time domain pyramid for feature fusion. Finally, the fused features are classified using SVM. EMO-DB, RML, eNTERFACE05, and BAUM-1s datasets were used to test the performance. Results show that all of these yielded high classification accuracy. Reference [110] performed feature extraction and feature fusion on speech spectrograms based on convolution neural network and multiple kernel learning, respectively, and then the support vector machines algorithm was used for emotion recognition. It was tested on EMO-DB and CASIA datasets and achieved 86% and 88% recognition accuracy. Reference [111] extracted MFCC and Mel spectrograms for speech signals and performed speech emotion recognition based on the bidirectional LSTM model. The bidirectional LSTM model consists of two parts for emotion recognition based on MFCC features and two Mel spectrograms with different temporal frequency resolutions. After obtaining the output of the two-part structure, feature fusion is performed at the decision level. Reference [112] performed emotion recognition based on 1D-CNN-LSTM and using speech and Mel spectrograms, respectively. The results show that 2D-CNN-LSTM is able to learn local features, global features and temporal dependencies of emotions compared to 1D-CNN-LSTM. It was tested on Emo-DB and IEMOCAP datasets and obtained 92.9% and 89.16% recognition accuracies, respectively. Reference [113] extracted spectrograms, MFCC, cochleograms and fractal dimensions for speech signals, transformed time sequences into image sequences and used an end-to-end approach combined with an attention mechanism based on a 3D CNN-LSTM model to achieve motion recognition. It was tested on RAVDESS, RML, and SAVEE datasets and achieved 96.18%, 93.2% and 87.5%

recognition accuracy, respectively.

Feature extraction of speech spectrogram based on convolution neural network will be missing spatial information, which is significantly related to low-level features such as resonance peaks and treble. Capsule network can extract spatial information from the speech spectrogram and transfer the information by dynamic routing. Reference [114] introduced capsule networks into the speech emotion recognition task. To improve the model's ability to capture contextual features, they introduced a recurrent neural network to capture the time domain information in speech and achieved 72.73% accuracy in the EMOCAP dataset. It has an internal recurrent routing protocol algorithm. Therefore, capsule networks are slower. Especially as the size of the dataset increases, the number of training epochs for the model needs to be increased. Furthermore, the model compression algorithm applied to the convolution neural network cannot be directly applied to the capsule network. Therefore, capsule networks have higher complexity. Reference [115] proposed the DC-LSTM COMP-CapsNet capsule network model which can reduce the model's computational complexity while ensuring recognition accuracy. It was tested on four public data and achieved 89.3% recognition accuracy.

Compared to convolution neural network and recurrent neural network, the transformer has superior performance in feature extraction and long sequence modeling. Therefore, the transformer was introduced into the field of speech emotion recognition. Reference [116] proposed unsupervised domain adaptation approach based on Transformer and mutual information for cross-corpus speech emotion recognition. The transformer is used to extract dynamic features of speech from Mel's spectrogram. Then, common and individual features of different precursor corpora are learned from the extracted features based on maximum and minimum mutual information strategies to eliminate the differences between corpora. Finally, an interactive multi-head attention fusion strategy is proposed to learn the complementarities between different features, and speech emotion recognition is performed based on the fused features. Reference [117] proposed a transformer-like model for speech emotion recognition in order to reduce the problem of excessive time and memory overhead during the training process for the transformer model. The method achieves similar recognition accuracy as the original Transformer model while reducing the time and memory overheads during model training. Transformer-based approaches usually require frame-by-frame computation of attention coefficients, which cannot capture local emotion information and are susceptible to noise interference. Reference [118] proposed the BAT model. By calculating the self-attention of the spectrogram after chunking, they can mitigate the effect of local noise generated by high-frequency energy while capturing real emotions. In addition, they proposed a cross-block attention mechanism to facilitate the information interaction between blocks. Moreover, they integrated the FCCE module used to reduce attentional bias. IEMOCAP and Emo-DB datasets were used to test the performance, an accuracy of 75.2% and 89% were obtained. Transformer based speech emotion recognition model requires a large amount of data to train the model adequately. In order to remedify the problem of insufficient training data, [119] proposed the ADAN model for generating training samples. They combined the generated data and tested it on Emo-DB and IEMOCAP datasets, which achieved recognition accuracies of 84.49% and 66.92%, respectively.

3.6. Trusted emotion recognition based on other signals

Head posture, hand posture, and facial feature points can also be used for emotion recognition.

head posture i.e., head motion is characterized by calculating the pitch, yaw and roll angles of the head in space. The flowchart of emotion recognition based on head posture is shown in Figure 10. First, feature point detection is performed in the facial image and the detected facial feature points are matched with the 3D face model. Second, according to the relationship between the world coordinate system, camera coordinate system and imaging plane in order to determine the affine transformation matrix from the 3D face model to the 2D face feature points. The euler angles consist of yaw, roll and pitch, which can be defined as α, β , and γ and the quaternions can be defined as follows.

$$R_{wb} = (R_x(\alpha)R_y(-\beta)R_z(r))^T \tag{21}$$

where $R_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\alpha) & sin(\alpha) \\ 0 & -sin(\alpha) & cos(\alpha) \end{bmatrix}$, $R_y(-\beta) = \begin{bmatrix} cos(\beta) & 0 & sin(\beta) \\ 0 & 1 & 0 \\ -sin(\beta) & 0 & cos(\beta) \end{bmatrix}$, $R_z(\gamma) = \begin{bmatrix} cos(\gamma) & sin(\gamma) & 0 \\ -sin(\gamma) & cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Thus, yaw, roll, and pitch can be calculated as follows:

$$\alpha = arctan2(R_{wb}(3,2), R_{wb}(3,3)) \tag{22}$$

$$\beta = arcsin\big(R_{wb}(3,1)\big) \tag{23}$$

$$\gamma = arctan2(R_{wb}(2,1), R_{wb}(1,1)) \tag{24}$$

Third, the time series of head posture is obtained based on the Euler angles and quaternions. Finally, feature extraction of head posture is performed for emotion recognition.

Facial feature points can also be extracted as features for emotion recognition. Reference [120] judged the user's blinking based on the positional relationship between the eye feature points. Moreover, they extracted the duration of blinking, frequency of blinking, average aspect ratio, and eye distance when eyes are open as eye movement features based on the obtained time series to judge the learning engagement of students. Reference [121] recognized four discrete emotions through the relative positional relationships between facial feature points and the angle relationships between feature points in different parts of the face. When the model detects negative emotions in the learning process of students with autism, it automatically regulates the learning content and attention by means of animation in order to enhance the online learning ability of students with autism.

The online learning environment, although it is not possible to perform affective analysis by analyzing the learner’s complete body posture, affective changes can be analyzed by analyzing the Hand-over-Face (HOF), a movement that often occurs during the learning process. It has been shown that different hand movements and hand shapes imply different affective states, which in turn affects behavioral engagement [122,123], as shown in Figure 11.

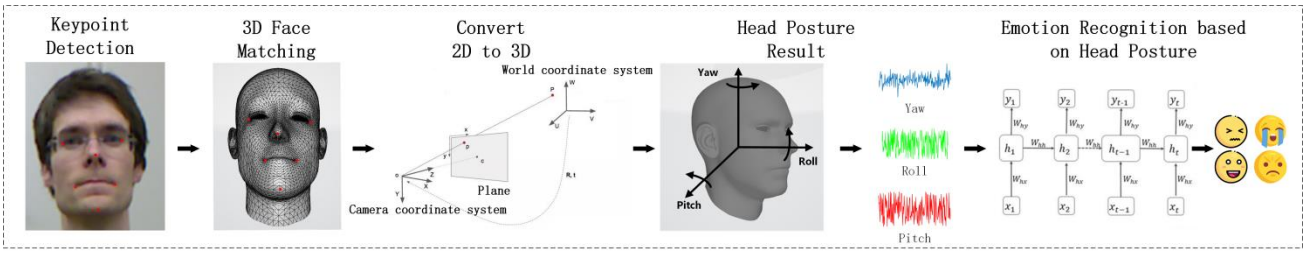
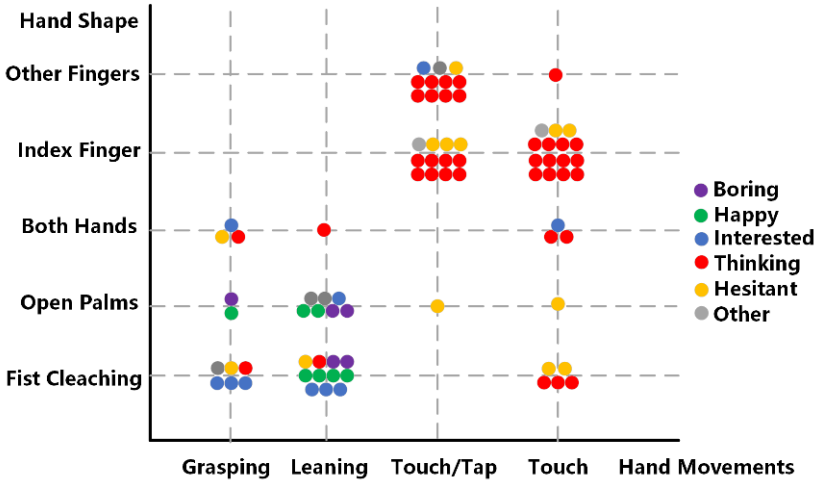


Figure 10. The flowchart of emotion recognition based on head posture.

(a) Examples of HOF images



(b) Hand states and hand shapes in different emotions

Figure 11. The relationship between HOF and emotions.

3.7. Trusted emotion recognition based on multiple signals

Emotion recognition based on a single signal cannot obtain better performance, which limits the scope of application. The advantages and disadvantages can be seen in Table 6.

Table 6. Summary of the emotion recognition based on signal feature.

Methods	Advantage	Disadvantage
3.1	Facial images are less difficult to acquire	Does not always truly reflect changes in emotion
3.2	Can truly reflect changes in emotions	Short duration, low amplitude and high difficulty.
3.3	Can truly reflect changes in emotions	Low accuracy of segmentation based on common equipment
3.4	Can truly reflect changes in emotions	Easy to be disturbed by external factors, low accuracy
3.5	May provide clues for emotion recognition	Low recognition accuracy under noise interference
3.6	May provide clues for emotion recognition	Low accuracy in single state recognition

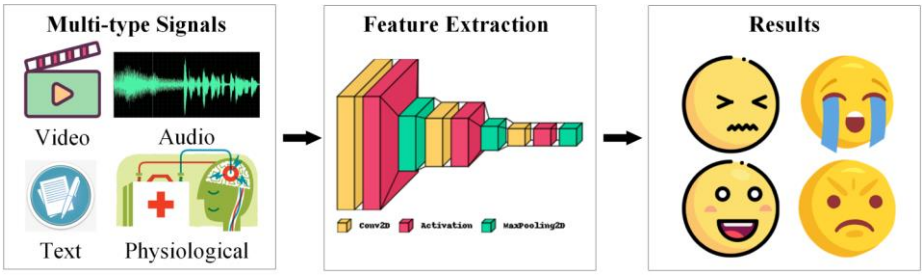


Figure 12. The flowchart of emotion recognition based on multiple signals.

The accuracy of emotion recognition can be improved based on multiple types of signals compared to single signal by exploiting the complementarity between signals. Established research methods need to combine video with audio, textual content and physiological to achieve emotion recognition. The multimodal emotion recognition process is shown in Figure 12.

There are large differences in dimensions and types between different signals. Therefore, data fusion is needed to take advantage of the complementarity between multimodal signals for emotion recognition. Feature fusion methods can be classified into data-level fusion, feature-level fusion, decision-level fusion and model-level fusion. Data-level fusion is also known as sensor-level fusion, which is a direct combination of the original data collected by the sensors and generates a new set of data. Data level fusion preserves the original data information and maintains the integrity of the information. However, data-level fusion is complicated and cumbersome. Feature-level fusion is to extract features from different original data and then compose a new feature set from each feature. Feature-level fusion can retain the information of the original data to the maximum extent. It has been shown that feature-level fusion can achieve the best recognition performance [124]. Decision-level fusion requires a joint decision on the credibility of each model. It is easier to implement than feature-level fusion. The main strategy adopted for model-level fusion is to learn more complex changes by building deep network models. It can fit more complex features to increase the nonlinear fitting capability. Signals can be extracted from video as shown in Figure 13, including one-dimensional signals and two-dimensional signals. Among them, the one-dimensional signals include head posture, non-contact physiological signal, pupil size and eye state. The two-dimensional signals include facial expression, micro-expression and gaze estimation. Data fusion can be performed at the feature-level (e.g., Figure 14(a)) and decision-level (e.g., Figure 14(b)) for emotion detection.

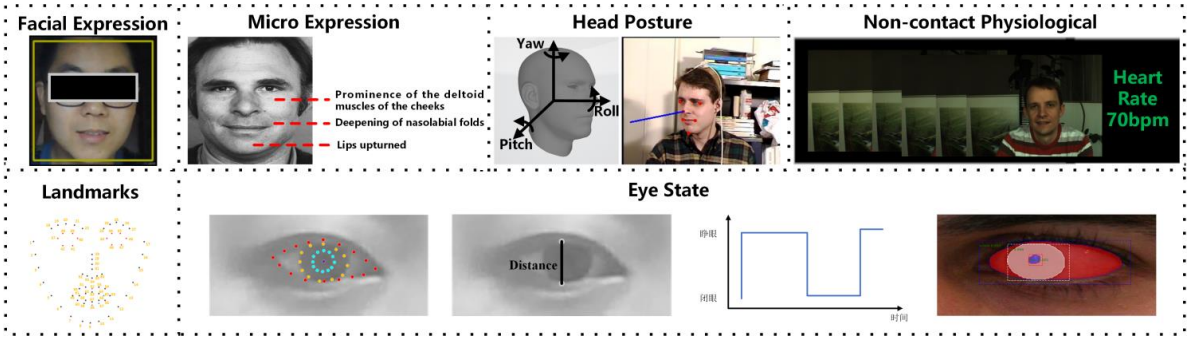
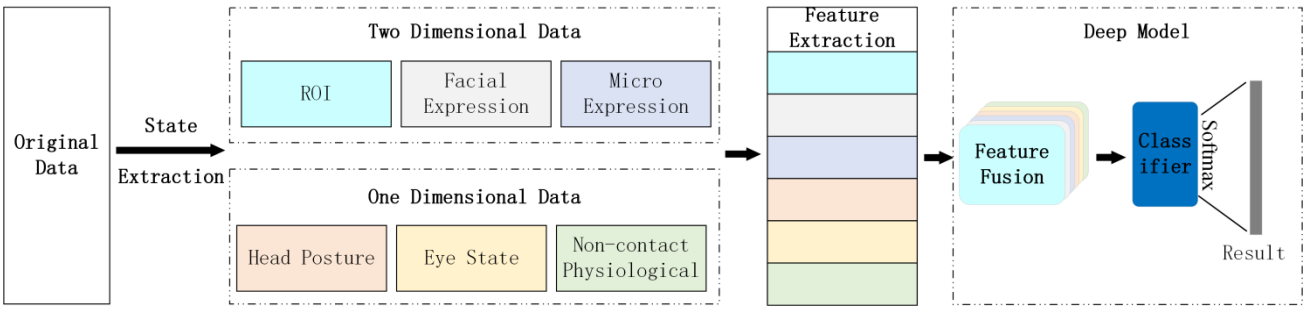
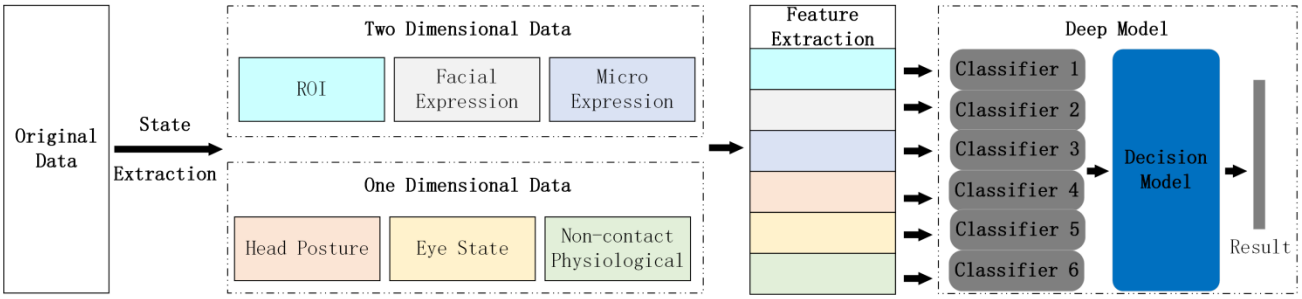


Figure 13. Multiple signals captured from video.



(a) Feature-level fusion



(b) Decision-level fusion

Figure 14. Fusion method based on different dimensions and types.

3.7.1. Multimodal emotion recognition databases

The multimodal emotion recognition database contains visual, speech, textual, and physiological data, of which visual includes facial expression, body posture, eye state, etc. Speech contains speech content that can reflect the subject's emotion, and physiological includes EEG, EMG, heart rate, and so on. According to the emotion label category, the database can be divided into discrete emotion recognition database and continuous emotion recognition database. The multimodal emotion recognition database is shown in Table 7.

Table 7. Commonly used multimodal emotion recognition databases

Database	Samples	Type	Category
DEAP [125]	1280	Images + Physiological Signals	Continuous emotion
CMU-MOSI [126]	2199	Image + Eye	Continuous emotion
CH-SIMS [127]	2281	Audio + Text	Continuous emotion
Multi-ZOL [128]	28,469	Image + Eye	Continuous emotion
LIRIS-ACCEDE [129]	9800	Audio + Text	Continuous emotion
MAHNOB-HCI [130]	532	Image + Text	10 categories of discrete emotions
eNTERFACE'05 [131]	1166	Image + Audio	5 categories of discrete emotions
CMU-MOSEI [126]	23,453	Image + Eye	6 categories of discrete emotions
IEMOCAP [100]	10,039	Physiological signals	9 categories of discrete emotions

3.7.2. Multimodal emotion recognition methods

Video and audio are commonly combined for emotion recognition tasks. Reference[132] proposed a bilinear pooled attention network model based on adaptive and multilayer decomposition for emotion recognition. For audio signals, an audio encoder was used to encode the spectrogram of audio signals and designed a convolutional network with attention mechanism and global response normalisation to extract audio features. For video signals, a video encoder was used to encode the extracted faces in the video and weighted the encoding results using a one-dimensional attention mechanism. Then, they proposed the FBP method based on the self-attention mechanism to achieve the fusion of audio and video features. Finally, emotion recognition was performed based on the fused features. Results show that an accuracy of 61.1% was reached combining video and audio data, compare with single data, an improvement of 9.03% and 26.11% was reached, respectively. Reference [133] proposed a convolution neural network model based on a two-stage fuzzy fusion strategy for emotion recognition based on the facial image and speech extracted from the video. Experiments were conducted on three multimodal datasets with facial expression and speech extraction. The results show that compared with single facial expression or speech, the accuracy can be improved up to 17.55%, 26.36%, and 37.25% on Entereface'05, AFEW and SAVEE dataset, respectively. Similar to [133,134] detected not only facial image but also facial feature points in the video. They extract HOG-TOP features for facial images and geometric features for facial feature points. Furthermore, they used HOG-TOP features and geometric features to represent visual features. They extracted the relevant acoustic feature for the speech signal. Finally, SVM classifier was used to recognize emotion based on fused features. Results show that an accuracy of 45.2% on AFEW 4.0 database can be obtained based on dynamic, acoustic and geometric features. When a user speaks, facial changes due to speech can be confused with facial changes due to emotional changes. Therefore, [135] first divided the face into upper and lower parts. Second, the pitch and content of the speech were combined with the upper and lower parts of the facial image, respectively. The emotion at that particular time was estimated separately. Third, emotion recognition was performed on the speech signal based on SVM classifier. Finally, the recognition results based on upper and lower facial image and speech signals were fused at the decision level for emotion recognition. IEMOCAP and SAVEE datasets were used to evaluate the algorithm's performance, an accuracy of 67.22% and 86.01%was achieved respectively. Further analysis finds that the upper face region, particularly the eyebrow region, is highly associated with speech emphasis. Reference [136] combined facial expression, intonation, and speech content. Initially, facial features were extracted

using FEA2. Then, features were extracted for intonation and speech content separately. Finally, the fused features were subjected to emotion recognition based on SVM. Visual, audio and text were combined to achieve emotion recognition, an accuracy of 95.6% was reached.C with each single signal, the accuracy can be improved up to 51.2%, 9.6%, and 21.7%, respectively.

In order to improve the performance of emotion recognition, text is also combined with video and audio for emotion recognition. Due to the differences in dimensional among different modal data, most of the existing research methods need to design different fusion networks. These fusion network designs were based on the differences in dimensions when performing multimodal data fusion. In [137], a modality-independent fusion network based on transformer was proposed which can fuse video, audio and textual features. The experiments carried out on the Hume-Reaction datasets and the performance were reported in terms of Pearson correlation. The best performance can be obtained by combining image, text, audio, and FAU, which is statistically significant with other configurations. Reference [138] investigated the fusion of different modal features. They first used transformer for feature extraction of three modal data. Then, the fusion was performed based on the intermediate layer features. The experimental results show that better recognition accuracy can be obtained based on intermediate layer features compared to deep features. Multimodal data fusion can take advantage of the complementarity between different data. While variations in features are expected among different modal data, they may also introduce redundant information. This leads to incomplete learned multimodal features. The experimental results on IEMOCAP and CMU-MOSEI show that compared with all baseline models, multimodal data fusion obtains the best results, which demonstrates the effectiveness of sufficiently leveraging the multimodal interactions among the intermediate representations. Reference [139] proposed FDMER, a multimodal emotion recognition method based on decoupled representation learning. FDMER maps each modality to modality-invariant subspace and modality-specific subspace respectively. The modality-invariant subspace and modality-specific subspace learn common and individual features among different models, respectively. The learned features are cascaded and adaptive weights are learned using the cross-modal attention fusion module to obtain effective multimodal features. Furthermore, most multimodal models fail to capture the association between each modality. Since fine-grained information is ignored, discriminative features cannot be extracted and it limits the generality. CMU-MOSI, CMU-MOSEI, and UR_FUNNY dataset were used to test the performance. Compared with single text, audio and visual pattern, the best performance can be obtained based on multimodal data fusion and the accuracy is 84.2%, 83.9%, and 70.55%, respectively. [140] proposed a feature fusion mechanism to encode temporal-spatial features to learn a more complete feature representation. Then, a late fusion strategy is used to capture fine-grained relationships between multimodalities, which is combined with an integration strategy to improve model performance. Results show that multimodal fusion is an efficient approach for further performance improvement by integrating complementary emotion-related information from each modality and the ensemble method can effectively combine the advantages of the single model.

Besides audio and text, EEG are combined with facial images for emotion recognition. Reference [141] combined facial expression with haptics. They collected pressure values under different emotions through pressure sensors and combined the pressure values with facial expressions for continuous emotion detection. Analysis shows that the participants usually integrated both visual and touch information to evaluate the emotional valence and adding this type of tactile stimulation to multimodal emotional communications platforms might enhance the quality of the

mediated emotional interactions. In [142], DCN was applied to fuse EEG signals and facial expressions in feature level to classify positive, neutral and negative emotion for deaf subjects. First, the EEG signals and facial expressions were simultaneously collected when they were watching emotional movie clips. Then, the EEG signals were preprocessed, and the DE feature in different bands were extracted to obtain 310 dimensions features. The first frame of one second were extracted from facial video records to select a portion of the facial expression, and resize the image to 30 × 30. Finally, two modal features were fused as the input of DCN to classify three emotions. The classification accuracy is 98.35%, which is 0.14% higher than facial expression emotion recognition and 0.96% lower than EEG emotion recognition. Unlike [142] who extracted facial expression and EEG features separately, [143] first transformed EEG data from time series to brain map form. Then, they used the VGG-16 depth model for feature extraction. For facial expressions, 30-dimensional facial features were formed by detecting the distance between feature points based on the facial feature points. The extracted EEG features and facial features were used as inputs to the LSTM model after PCA dimensionality reduction in order to achieve continuous emotion recognition based on EEG and facial expression. The classification accuracy is 54.22%, which is 10.7% and 5.39% higher than EEG and facial expression emotion recognition, respectively.

In [144], speech and textual were combined for emotion recognition. For speech information, CNN, RNN and attention mechanism were used for speech coding. For textual information, the Bert model was used for textual content coding. After obtaining the speech and textual content features, fusion was performed at the feature level. Results show that corporate with variable text benefit from the combination, which is 9.1% better than speech only and 1.7% better than text only. Reference [145] proposed a regularized deep fusion model for emotion recognition based on multiple types of physiological signals. First, feature extraction was performed from different types of physiological signals, with feature embedding using kernel matrix. Then, a specific representation of each physiological signal was learned from the feature embedding based on the depth model. Finally, a global fusion layer with regularization terms was designed for feature fusion. The experimental results show than the best performance was obtained when fusing all of the modalities both on arousal and valence, irrespective of the dataset. Reference [146] also performed emotion recognition based on multiple types of physiological signals. The time domain and frequency domain features of physiological signals have different activation levelswhere the activation level in the time domain can reflect the brain activity. High and low activation levels correspond to positive and negative emotions, respectively. In the frequency domain, negative emotions correspond to high activation. In addition, there is heterogeneity and correlation in multimodal physiological data. Heterogeneity refers to the differences between various signal properties collected from different organs. Correlation refers to the relationship between channels of the same modality or different modalities. Existing methods struggle to utilize the complementarity, heterogeneity, and correlation of features in the time-space-frequency domain of multimodal time series for emotion detection.

To address this, researchers proposed HetEmotionNet, a heterogeneous graphical neural network that aims to achieve the simultaneous modeling of the complementarity, heterogeneity, and correlation of features of multimodal data under a unified framework. The model uses GTN and GCN to model the heterogeneity and correlation of multimodal physiological data, respectively, and GRU to extract the dependencies between time-frequency and frequency domains of multimodal physiological data. The result indicates that the effects of only using EEG signals is better than only using PPS, because EEG

signals are the main physiological signals used in emotion recognition. Besides, fusing data of different modalities further improves the performance and reaches the best results. References [147] and [148] combined EEG and eye states for emotion recognition. Reference [147] used eye-tracker X120 to capture changes in pupil size and gaze point as eye movement features. Moreover, they fused the eye movement features with EEG features at the feature level and decision level, respectively, in order to verify the effect of different fusion strategies on the accuracy of emotion recognition. The best classification accuracies of 68.5% for three labels of valence and 76.4% for three labels of arousal were obtained using a modality fusion strategy and a support vector machine. Unlike [147], who used telemetric eye-tracker X120 to acquire eye states, [148] used a head-mounted eye-tracker to acquire eye states. The experimental results show the advantage of EEG features for the recognition of happy emotions. The recognition result of eye state for fear emotion is superior to EEG. This shows that different types of data have complementary advantages. Emotion recognition based on multiple types of data can improve the accuracy of emotion recognition. Reference [149] combined EEG and speech for emotion recognition. For EEG signals, 160-dimensional, differential entropy features were extracted in different frequency bands. For speech signals, MFCC coefficients, differential and acceleration coefficients were extracted to form 36-dimensional speech features. After feature extraction, feature fusion was performed at the feature level and decision level, respectively, to achieve emotion recognition. The experimental results show that the accuracy of the multimodal based approach is improved by 1.67% and 25.92% respectively compared to EEG or speech only for emotion recognition. This proves the effectiveness of the multimodality approach.

Although the accuracy of motion recognition can be improved based on multimodal data, the multimodal data requires the use of different devices for raw data acquisition, which increases the economic cost. Furthermore, due to the difference in data acquisition frequency between different devices, it is necessary to align the relevant state data acquired, as well as data alignment and annotation will increase the manpower cost. Emotions are complex and can be simultaneously reflected in changes between different state data. In the case that the data cannot be accurately aligned, it will affect the accuracy of emotion recognition. In addition, further summary and analysis of existing research work can be found that existing research work focuses more on the recognition of basic emotions. The basic emotions cannot comprehensively reflect the changes in students' emotions during the learning process. In addition, learning emotions such as confusion, thinking and other types of learning emotions that occur during the learning process are not included in the basic emotions. Therefore, the existing methods have some limitations. In the next study, the basic emotion categories can be subdivided by combining the multi-type state data extracted from the video to achieve fine-grained emotion recognition on the one hand. On the other hand, the basic emotion categories are extended to achieve the recognition of learned emotions.

4. The application of emotion recognition technology in intelligent education

4.1. The application of contact physiological signals detection in the field of education

Contact physiological Signals are detected with a high degree of accuracy, which can provide data support for analyzing students' learning state, improving students' academic performance, detecting changes in students' emotional state during the learning process, and improving students' participation in the classroom. Heart rate variability can be used to recognize academic performance,

cognitive load, learning anxiety and physical activity participation level. Reference [150] investigated the relationship between adolescents' personality and heart rate variability and academic performance by analyzing the changes in heart rate variability of 91 seventh-grade students before, during, and after watching stress-inducing videos. Furthermore, students' ability to adapt and recover from stress can be judged by the changes in HRV. The experimental results showed that heart rate variability decreased during the viewing of the evoked video. Heart rate variability increased during the recovery period after the end of the video. Extroverted students and introverted students with greater increases in heart rate variability during the recovery period had better academic performance. Reference [151] investigated the relationship between changes in heart rate variability and academic performance through a game. Analyzing the trend of heart rate variability collected during the game revealed that being positively motivated by the game led to an increase in heart rate variability. This indicates that students' concentration is improved.Heart rate variability is also associated with mental load. Heart rate variability decreases as mental load increases and high-achieving students have lower mental load compared with underperforming students. The performance of learning outcomes may be influenced by cognitive load and prior knowledge together. When cognitive load significantly exceeds the prior knowledge, it can have a negative impact on student's learning outcomes. In [152], heart rate was used to detect students' and teachers' cognitive load during the chemistry course. Results showed that students' heart rate variability increased as cognitive load increased. Differently, although there is an increase in cognitive load, there is no significant difference between the changes in teachers' heart rate and cognitive load.

In [153], ATS emotional aid system was proposed to detect students' emotions through heart rate variability, which can improve student's learning outcomes. In [154], heart rate was collected to investigate whether it was possible to relieve anxiety by performing positive thinking exercises via mobile phone. Results show that anxiety can be alleviated based on mindfulness practice.

Physiological signals can also be used to analyze a student's motion state. Reference [155] aims to investigate whether physical education classes can increase the amount of physical activity among adolescent students, which can lead to a healthy lifestyle. ActiGraph was used during the experiment to record students' heart rate and step count. Results show that students who participated in physical education classes had more total daily exercise, more exercise intensity, and were able to meet the required exercise standards in greater numbers than students who did not participate in physical education classes. In addition, exercise duration can be effectively increased through human intervention. It is a challenge to increase girls' participation and perceptual ability in physical education. Compared with boys, girls have lower levels of participation in team sports, such as football and basketball. Reference [156] investigated whether the use of single-gender grouping strategy could enhance girls' participation and perceptual abilities in physical education. During the experiment, heart rate was monitored as a cue to judge activity intensity. The results showed a significant improvement in girls' perceptual abilities based on single-gender grouping strategy.

4.2. The application of eye states recognition in the field of education

Gaze points can truly reflect the region of interest during the class. The application of eye states can be divided into the following two categories. First, eye state can be used to analyze the impact of different presentation methods of the same knowledge content on students' learning effects, which can optimize the teaching methods.

The presentation mode of learning content may affect student's learning outcomes. Teachers can develop personalized learning styles and content presentations based on students' different learning preferences. Reference [157] investigated the effects of graphical and flux-based presentations on the understanding of the physical concept of divergence by analyzing the students' line of sight. The results of the experiment showed that better performance can be obtained by combining graphical and flux-based explanations. In addition, the effect of the learning style chosen by the students was better than the one assigned by the teacher. Reference [158] investigated the effect of different layouts of text and images in instructional materials on student's learning outcomes. The experimental results showed that the closer the text and image position, the higher the learning efficiency. Moreover, the use of text to supplement the content of images helps to improve the retention of the teaching content. Reference [159] investigated the effect of viewing visual learning process data on students' learning outcomes. They randomly divided students into two groups. The results of the pre-test showed no difference in prior knowledge and comprehension between the two groups. One group of students was able to view eye-tracking data from their learning process. Then, both groups of students reread the new material. The results of the post-test showed that the students who viewed the learning process data had stronger text processing and comprehension abilities regarding the learning content. In [160], the impact of virtual teacher and gaze guidance was combined to investigate the learning outcomes in e-learning. Results show that learning performance was not affected by virtual teacher and students were able to effectively allocate their attention between virtual teacher and the learning content. In [161], the impact of cognitive level on students' knowledge acceptance was studied. Sixty-two students participated in the experiment. They were asked to read the pre-revised and post-revised mathematics textbooks and then gaze points were analyzed. Results show that students with prior knowledge exhibited superior cognitive processing abilities for the revised materials compared to the original materials, while students lacking prior knowledge showed no significant differences.

Second, eye states can be used to analyze the differences in eye states between high-achieving students and lagging students, so as to optimize learning habits. Academic performance can be affected by learning habits. By analyzing the gaze points of high-achieving and low-achieving students during the learning process, it is possible to provide learning guidance for low-achieving students and optimize their learning habits. Research in [162] was studying the reading habits of elementary school students. Results show that high-achieving students paid more attention to graphical information and they had greater learning efficiency and comprehension. In [163], the change in pupil diameter was used to analyze the relationship between the difficulty of problems and cognitive level. Results show that pupil diameter will change as the difficulty of the problem increases. Reference [164] investigated the common and individual characteristics of eye movement patterns of students with different reading habits. Analyzing the gaze points, it was found that students with high reading ability had smooth visual sweeps and high accuracy in locating key information.

4.3. The application of facial expression recognition in the field of education

Emotions not only affect students' attention, memory and decision making, but also influence learning motivation and interest. The main purpose of facial emotion recognition in the field of education is to detect student's learning engagement in the learning process, which can help teachers

timely adjust the teaching content and strategy. Engagement includes emotion engagement, behavior engagement and learning engagement. Facial expressions are often used to determine students' emotional engagement in the learning process. Unlike the six basic emotions of anger, disgust, fear, happiness, sadness, and surprise proposed by Ekman, academic emotion refers to the facial expressions that students show during the learning process. It includes confusion, boredom, sleepiness, concentration, and anxiety. Existing research work has been done more on designing different depth models to detect the six basic emotions proposed by Ekman. The detection of basic emotions can provide some support for teachers to judge students' emotional engagement. However, basic emotions cannot fully reflect academic emotions and it cannot judge students' emotional engagement. Although basic emotion datasets are available for model training, there is a lack of learning-based emotion recognition datasets, which limits the development of academic emotion recognition. Micro-expression can accurately reflect students' emotions, but micro-expression is short with low amplitude, which makes it difficult to detect. Therefore, less research work has been done to achieve the analysis of emotional engagement based on micro-expression in the field of education. Reference [165] proposed a micro-expression recognition model based on a hybrid neural network with bimodal spatial-temporal feature representation for micro-expression recognition in e-learning. The model consists of two modalities. The first modality models the changing geometry and dynamics of the face. First, feature points are detected for facial images. Then, the facial feature points are combined into vectors. Finally, the dynamic geometric features of the face are learned using DNN. The second modality is to model the facial appearance and dynamic changes. First, CNN was used to extract the appearance features. Then, the extracted appearance features are used as inputs to LSTM to learn the dynamic features of micro-expression changes. Finally, the learned features from both modalities are fused at the feature level for micro-expression recognition.

Head posture and eye state are combined with facial expression to recognize student's behavioral engagement. Gaze tracking is a method of recording real, natural, and objective user behavior, which can analyze students' behavioral engagement by analyzing head posture, gaze points, etc. In [166], blink frequency, body posture, and facial expression were extracted from videos during online exam to recognize student's behavioral engagement. Different from [166], in [167], facial expressions, hand gestures, and body postures were extracted to analyze both individual and overall behavioral engagement.

A single engagement cannot fully reflect the learning state. In [168], emotional and behavioral engagement were combined to recognize learning engagement. First, facial landmarks were extracted from video and head posture was calculated based on the extracted landmarks. Second, action units were detected from the images. Third, on the one hand head posture and gaze points were combined to recognize behavioral engagement, on the other hand action units were used to recognize emotional engagement. Finally, behavioral engagement and emotional engagement were combined to recognize learning engagement.

4.4. The real-world applications of emotion recognition in the field of education

The application of emotion recognition systems in the field of education carries significant importance. First, emotion recognition systems assist educators in understanding students' emotional states and needs, enabling them to provide personalized learning support. By monitoring students' emotional responses in real-time, the system can adjust learning resources, feedback, and teaching

strategies to cater for individual learning needs, thereby enhancing learning outcomes and engagement. Second, emotion recognition systems help identify students' emotional issues and mental well-being. Educators can detect emotional distress early and provide appropriate support and guidance. This positively impacts students' emotion management, mental health, and overall well-being. Third, emotion recognition systems aid educators in refining teaching methods and assessment approaches. By understanding students' emotional responses during the learning process, educators can adapt their teaching strategies, provide more accurate assessment and feedback, cater to students' emotional needs, and improve assessment accuracy. Fourth, emotion recognition systems can be utilized to create emotionally intelligent learning environments. By recognizing students' emotional states, the system can automatically adjust factors such as music, colors, and ambiance within the learning environment to foster a positive emotional atmosphere, thereby enhancing students' emotional engagement and learning outcomes. Finally, emotion recognition systems contribute to increased student engagement and motivation. By promptly understanding students' emotional states, educators can provide relevant incentives and support based on their emotional needs, encouraging active participation in learning activities and enhancing students' interest and motivation.

There are many emotion recognition systems used in the field of education. EmotionTracker is an application that utilizes emotion recognition technology in the education field. It is a system designed to monitor and track students' emotional states in real-time. Through various sensors, such as facial expression analysis, voice tone analysis, or physiological measurements, EmotionTracker captures data related to students' emotions during learning activities. The purpose of EmotionTracker is to provide educators with valuable insights into students' emotional experiences and needs. By analyzing the collected data, the system can identify patterns and trends, allowing educators to better understand how students are feeling during different learning tasks, lessons, or assessments. EmoLearn is a specific application that leverages emotion recognition technology in the education domain. It is a platform designed to deliver personalized learning experiences based on students' emotional states and needs. The primary goal of EmoLearn is to create a learning environment that takes into account students' emotions and tailors' instructional content and strategies. By utilizing emotion recognition algorithms, the system identifies and analyzes students' emotions in real-time during their learning activities. The system can dynamically adjust the learning materials, pacing, and difficulty level to match each student's emotional state. Affectiva is a company that specializes in emotion recognition technology and provides solutions for emotion AI. They have developed advanced algorithms and software tools to detect and analyze human emotions through facial expressions, voice tone, and physiological signals. Affectiva's emotion recognition technology can also assist in assessing students' emotional well-being and mental health. By detecting signs of stress, frustration, or other negative emotions, educators and counselors can intervene and provide appropriate support to students in need.

5. Discussion and conclusions

In this paper, the detailed related datasets and technologies are used for emotion recognition. The performance of trusted emotion recognition can be improved by combining multiple types of signals. However, existing emotion recognition methods that combine physiological information, speech, micro-expressions, and posture have some limitations for e-learning. In the context of online education, video can senselessly record learners' behaviors during the learning process. Extracting features that

authentically capture changes in emotion holds significant practical importance for the future implementation of e-learning. However, how to integrate different types of signals and leverage the complementary information between different data is a challenge that should be addressed.

Apart from technical aspects, it is crucial to consider different student populations, including students with disabilities or from diverse cultural backgrounds, addressing the accessibility and inclusivity issues of trusted emotion recognition technologies in education. The application of emotion recognition technologies needs to consider the needs of students with disabilities to ensure their equal participation in the learning process. For instance, adaptive interfaces or assistive tools can be provided for students with visual or hearing impairments to interact and express emotions with the technology. Additionally, collaborating with representatives and experts from the disability community to understand their needs and perspectives is essential in ensuring the accessibility and effectiveness of the technology. The accuracy and applicability of emotion recognition technologies may vary across different cultural backgrounds. Interpretation of facial expressions and body language can differ in various cultures. Therefore, caution should be exercised in the use of emotion recognition technologies in education to avoid imposing specific cultural norms of emotional expression as the standard. Developing and training emotion recognition algorithms should account for diversity and cultural differences to ensure the applicability and accuracy of the technology across different cultural backgrounds. Collecting and utilizing diverse datasets is crucial to enhance the inclusivity of emotion recognition technologies. The datasets should include samples from diverse age groups, gender, races, cultures, and disability populations. This helps to reduce data biases and improve the universality of the technology. Additionally, protecting data privacy and ensuring the representativeness of the data are important considerations. Relevant training and education are essential to improve educators' and students' understanding and acceptance of emotion recognition technologies. Educators should learn about the limitations and potential biases of the technology and how to correctly interpret and use the results of emotion recognition. Moreover, educators should educate students about the diversity of emotional expression and foster a learning environment that is inclusive and respectful of different ways of expressing emotions. In the field of education, the accessibility and inclusivity of emotion recognition technologies are key to ensuring that every student gets benefits. Considering the needs of students with disabilities and those from diverse cultural backgrounds, along with incorporating diverse datasets, is crucial for addressing accessibility and inclusivity challenges in emotion recognition technologies. Additionally, offering adaptive tools and training can further enhance our ability to tackle these issues effectively in education. This ensures that all students can equally avail themselves of the benefits offered by these technologies.

With advancements in technology, video-based emotion recognition devices have made significant progress. These devices analyze facial expressions, speech, and body language to infer an individual's emotional state. Emotion recognition technology holds great potential for various fields, including education, healthcare, marketing, and human-computer interaction. Firstly, attention should be given to the applicability of emotion recognition equipment. While emotion recognition technology has shown promising accuracy in laboratory settings, it faces challenges in real-world applications. Factors such as lighting conditions, noise, and individual differences can impact the accuracy of emotion recognition. Therefore, further improvements and optimization of algorithms are necessary to enhance the applicability and accuracy of the equipment in practical scenarios. Second, the economic significance of emotion recognition equipment should be considered. As the demand for emotion recognition technology grows, the economic implications of these devices

become more apparent. In the education sector, for instance, emotion recognition equipment can help teachers gain a better understanding of students' emotional states, enabling personalized adjustments to teaching strategies and enhancing learning outcomes. Nevertheless, with the proliferation of devices and their widening applications, it becomes imperative to consider cost-effectiveness factors such as equipment costs, maintenance expenses, and data privacy and security concerns. In conclusion, it is important to focus on the applicability and economic significance of video-based emotion recognition equipment. By continuously improving technology, addressing practical challenges, and weighing cost-effectiveness factors, we can harness the potential of emotion recognition equipment and achieve broader applications across various domains.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Humanities and Social Science Fund of Ministry of Education (22YJA880091), CN.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Q. Hu, L. Liu, N. Ding, The dilemma and solution of online education in the perspective of educational equity, *China Educ. Technol.*, **8** (2020), 14−21. https://doi.org/10.3969/j.issn.1006-9860.2020.08.003
2. M. Balaam, G. Fitzpatrick, J. Good, R. Luckin, Exploring affective technologies for the classroom with the subtle stone, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (2010), 1623−1632. https://doi.org/10.1145/1753326.1753568
3. A. Hutanu, P. E. Bertea, A review of eye tracking in elearning, in *Proceedings of the 15th International Scientific Conference eLearning and Software for Education*, (2019), 281−287. https://doi.org/10.12753/2066-026X-21-038
4. Y. Wang, Q. Wu, S. Wang, X. Q. Fang, Q. Ruan, MI-EEG: Generalized model based on mutual information for EEG emotion recognition without adversarial training, *Expert Syst. Appl.*, **244** (2024), 122777. https://doi.org/10.1016/j.eswa.2023.122777
5. T. Fan, S. Qiu, Z. Wang, H. Zhao, J. Jiang, Y. Wang, et al., A new deep convolutional neural network incorporating attentional mechanisms for ECG emotion recognition, *Comput. Biol. Med.*, **159** (2023), 106938. https://doi.org/10.1016/j.compbiomed.2023.106938
6. Q. Xu, W. Sommer, G. Recio, Control over emotional facial expressions: Evidence from facial EMG and ERPs in a Stroop-like task, *Biol. Psychol.*, **181** (2023), 108611. https://doi.org/10.1016/j.biopsycho.2023.108611

7. J. J. Zhang, G. M. Sun, K. Zheng, S. Mazhar, X. H. Fu, Y. Li, et al., SSGNN: A macro and microfacial expression recognition graph neural network combining spatial and spectral domain features, *IEEE Trans. Human-Mach. Syst.*, **52** (2022), 747−760. https://doi.org/10.1109/THMS.2022.3163211
8. J. Zhang, K. Zheng, S. Mazhar, X. Fu, J. Kong, Trusted emotion recognition based on multiple signals captured from video, *Expert Syst. Appl.*, **233** (2023), 120948. https://doi.org/10.1016/j.eswa.2023.120948
9. J. Zhang, G. Sun, K. Zheng, Review of gaze tracking and its application in intelligent education, *J. Comput. Appl.*, **40** (2020), 3346. https://doi.org/10.11772/j.issn.1001-9081.2020040443
10. P. Van Cappellen, M. E. Edwards, M. N. Shiota, Shades of expansiveness: Postural expression of dominance, high-arousal positive affect, and warmth, *Emotion*, **23** (2023), 973−985. https://doi.org/10.1037/emo0001146
11. Z. Yu, X. Li, G. Zhao, Facial-video-based physiological signal measurement: Recent advances and affective applications, *IEEE Signal Process. Mag.*, **38** (2021), 50−58. https://doi.org/10.1109/MSP.2021.3106285
12. R. W. Picard, *Affective Computing*, MIT Press, (2000), https://doi.org/10.7551/mitpress/1140.001.0001
13. J. J. Wang, Y. H. Gong, Recognition of multiple drivers' emotional state, in *Proceedings of the 19th International Conference on Pattern Recognition*, (2008), 1−4. https://doi.org/10.1109/icpr.2008.4761904
14. F. Ungureanu, R. G. Lupu, A. Cadar, A. Prodan, Neuromarketing and visual attention study using eye tracking techniques, in *Proceedings of the 21st International Conference on System Theory, Control and Computing*, (2017), 553−557. https://doi.org/10.1109/icstcc.2017.8107093
15. M. Uljarevic, A. Hamilton, Recognition of emotions in autism: A formal meta-analysis, *Journal of Autism and Developmental Disorders*, **43** (2013), 1517−1526. https://doi.org/10.1007/s10803-012-1695-5
16. I. Lopatovska, Searching for good mood: examining relationships between search task and mood, *ASIS&T*, **46** (2009), 1−13. https://doi.org/10.1002/meet.2009.1450460222
17. P. Sarkar, A. Etemad, Self-supervised ECG representation learning for emotion recognition, *IEEE Trans. Affect. Comput.*, **13** (2022), 1541−1554. https://doi.org/10.1109/taffc.2020.3014842
18. P. Pandey, K. R. Seeja, Subject independent emotion recognition from EEG using VMD and deep learning, *J. King Saud. Univ. Comput. Inf. Sci.*, **34** (2022), 1730−1738. https://doi.org/10.1016/j.jksuci.2019.11.003
19. G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, M. Tsiknakis, Review on psychological stress detection using biosignals, *IEEE Trans. Affective Comput.*, **13** (2019), 440−460. https://doi.org/10.1109/taffc.2019.2927337
20. D. J. Diaz-Romero, A. M. R. Rincon, A. Miguel-Cruz, N. Yee, E. Stroulia, Recognizing emotional states with wearables while playing a serious game, *IEEE Trans. Instrum. Meas.*, **70** (2021), 1−12. https://doi.org/10.1109/tim.2021.3059467
21. S. Zhang, X. Zhao, Q. Tian, Spontaneous speech emotion recognition using multiscale deep convolutional LSTM, *IEEE Trans. Affective Comput.*, **13** (2019), 680−688. https://doi.org/10.1109/taffc.2019.2947464

22. S. Peng, R. Zeng, H. Liu, L. Cao, G. Wang, J. Xie, Deep broad learning for emotion classification in textual conversations, *Tsinghua Sci. Technol.*, **29** (2024), 481−491. https://doi.org/10.26599/tst.2023.9010021
23. A. Kleinsmith, N. Bianchi-Berthouze, Affective body expression perception and recognition: A survey, *IEEE Trans. Affective Comput.*, **4** (2013), 15−33. https://doi.org/10.1109/t-affc.2012.16
24. M. Jeong, B. C. Ko, Driver's facial expression recognition in real-time for safe driving, *Sensors (Basel)*, **18** (2018), 4270. https://doi.org/10.3390/s18124270
25. A. K. Davison, C. Lansley, N. Costen, K. Tan, M. H. Yap, SAMM: A spontaneous micro-facial movement dataset, *IEEE Trans. Affective Comput.*, **9** (2018), 116−129. https://doi.org/10.1109/taffc.2016.2573832
26. C. Cao, Y. Weng, S. Zhou, Y. Tong, K. Zhou, FaceWarehouse: A 3D facial expression database for visual computing, *IEEE Trans. Visual. Comput. Graph.*, **20** (2014), 413−425. https://doi.org/10.1109/tvcg.2013.249
27. O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, A. D. Van Knippenberg, Presentation and validation of the radboud faces database, *Cognit. Emotion*, **24** (2010), 1377-1388. https://doi.org/10.1080/02699930903485076
28. M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, (1998), 200−205. https://doi.org/10.1109/afgr.1998.670949
29. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, (2010), 94−101. https://doi.org/10.1109/cvprw.2010.5543262
30. G. Zhao, X. Huang, M. Taini, S. Z. Li, M. Pietikalnen, Facial expression recognition from near-infrared videos, *Image Vision Comput.*, **29** (2011), 607−619. https://doi.org/10.1016/j.imavis.2011.07.002
31. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, et al., Challenges in representation learning: A report on three machine learning contests, *Neural Networks*, **65** (2015), 59−63. https://doi.org/10.1016/j.neunet.2014.09.005
32. A. Mollahosseini, B. Hasani, M. H. Mahoor, Affectnet: A database for facial expression, valence, and arousal computing in the wild, *IEEE Trans. Affect. Comput.*, **10** (2017), 18−31. https://doi.org/10.1109/taffc.2017.2740923
33. S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, *IEEE Trans. Image Process.*, **28** (2018), 356−370. https://doi.org/10.1109/tip.2018.2868382
34. C. F. Benitez-Quiroz, R. Srinivasan, A. M. Martinez, EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild, in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, (2016), 5562−5570. https://doi.org/10.1109/cvpr.2016.600
35. P. Ekman, W. V. Friesen, Measuring facial movement, *J. Nonverbal. Behav.*, **1** (1976), 56−75. https://doi.org/10.1007/BF01115465
36. Y. Fang, J. Luo, C. Lou, Fusion of multi-directional rotation invariant uniform LBP features for face recognition, in *2009 Third International Symposium on Intelligent Information Technology Application*, (2009), 332−335. https://doi.org/10.1109/iita.2009.206

37. T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, K. Yan, A deep neural network-driven feature learning method for multi-view facial expression recognition, *IEEE Trans. Multimedia*, **18** (2016), 2528−2536. https://doi.org/10.1109/TMM.2016.2598092
38. P. Kumar, S. L. Happy, A. Routray, A real-time robust facial expression recognition system using HOG features, in *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, (2016), 289−293. https://doi.org/10.1109/CAST.2016.7914982
39. N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, A. M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, *Neurocomputing*, **273** (2018), 643−649. https://doi.org/10.1016/j.neucom.2017.08.043
40. X. Jian, D. X. Qing, W. S. Jin, W. Y. Shou, Background subtraction based on a combination of texture, color and intensity, in *Proceedings of the 9th International Conference on Signal Processing*, (2008), 1400−1405. https://doi.org/10.3724/sp.j.1004.2009.01145
41. S. Shojaeilangari, W. Y. Yau, K. Nandakumar, J. Li, E. K. Teoh, Robust representation and recognition of facial emotions using extreme sparse learning, *IEEE Trans. Image Process*, **24** (2015), 2140−2152. https://doi.org/10.1109/TIP.2015.2416634
42. Y. D. Chen, X. Yang, T. J. Cham, J. F. Cai, Towards unbiased visual emotion recognition via causal intervention, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 60−69. https://doi.org/10.1145/3503161.3547936
43. L. Wang, G. Jia, N. Jiang, H. Wu, J. Yang, EASE: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 218−227. https://doi.org/10.1145/3503161.3548005
44. P. Barros, E. Barakova, S. Wermter, Adapting the interplay between personalized and generalized affect recognition based on an unsupervised neural framework, *IEEE Trans. Affect. Comput.*, **13** (2022), 1349−1365. https://doi.org/10.1109/TAFFC.2020.3002657
45. K. Zheng, L. Tian, Z. Li, H. Li, J. Zhang, Incorporating eyebrow and eye state information for facial expression recognition in mask-obscured scenes, *Electron. Res. Arch.*, **32** (2024), 2745−2771. https://doi.org/10.3934/era.2024124
46. A. S. Cowen, D. Keltner, F. Schroff, B. Jou, H. Adam, G. Prasad, Sixteen facial expressions occur in similar contexts worldwide, *Nature*, **589** (2021), 251−257. https://doi.org/10.1038/s41586-020-3037-7
47. K. Zheng, D. Yang, J. Liu, Recognition of teachers' facial expression intensity based on convolutional neural network and attention mechanism, *IEEE Access*, **8** (2020), 226437−226444. https://doi.org/10.1109/access.2020.3046225
48. J. J. Zhang, G. M. Sun, K. Zheng, S. Mazhar, X. H. Fu, D. Yang, Emotion recognition based on graph neural networks, in *Proceedings of the International Conference on Cognitive Systems and Signal Processing ICCSIP 2020: Cognitive Systems and Signal Processing*, (2021), 472−480. https://doi.org/10.1007/978-981-16-2336-3_45
49. W. J. Yan, Q. Wu, Y. J. Liu, S. J. Wang, X. Fu, CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces, in *Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, (2013), 1−7. https://doi.org/10.1109/fg.2013.6553799

50. W. J. Yan, X. Li, S. J. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, X. Fu, CASME II: An improved spontaneous micro-expression database and the baseline evaluation, *PLoS One*, **9** (2014), e86041. https://doi.org/10.1371/journal.pone.0086041
51. J. Li, Z. Dong, S. Lu, S. J. Wang, W. J. Yan, Y. Ma, et al., CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2023), 2782−2800. https://doi.org/10.1109/tpami.2022.3174895
52. C. H. Yap, C. Kendrick, M. H. Yap, SAMM Long Videos: A spontaneous facial micro- and macro-expressions dataset, in *Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition*, (2020), 771−776. https://doi.org/10.1109/fg47880.2020.00029
53. P. Husak, J. Cech, J. Matas, Spotting facial micro-expressions in the wild, in *Proceedings of the 22nd Computer Vision Winter Workshop*, (2017). https://api.semanticscholar.org/CorpusID:21669949
54. G. Warren, E. Schertler, P. Bull, Detecting deception from emotional and unemotional cues, *J. Nonverbal Behav.*, **33** (2009), 59−69. https://doi.org/10.1007/s10919-008-0057-7
55. M. Shreve, S. Godavarthy, D. Goldgof, S. Sarkar, Macro-and micro-expression spotting in long videos using spatio-temporal strain, in *Proceedings of the 2011 IEEE International Conference on Automatic Face and Gesture Recognition*, (2011), 51−56. https://doi.org/10.1109/fg.2011.5771451
56. S. Polikovsky, Y. Kameda, Y. Ohta, Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor, in *Proceedings of the 3rd International Conference on Image for Crime Detection and Prevention*, (2009), 16−21. https://doi.org/10.1049/ic.2009.0244
57. X. Ben, Y. Ren, J. Zhang, S. J. Wang, K. Kpalma, W. Meng, et al., Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, *IEEE Trans. Pattern Anal.*, **44** (2022), 5826−5846. https://doi.org/10.1109/tpami.2021.3067464
58. M. Peng, C. Wang, T. Chen, G. Liu, X. Fu, Dual temporal scale convolutional neural network for micro-expression recognition, *Front. Psychol.*, **8** (2017), 1745. https://doi.org/10.3389/fpsyg.2017.01745
59. D. H. Kim, W. J. Baddar, J. Jang, Y. M. Ro, Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition, *IEEE Trans. Affect. Comput.*, **10** (2017), 223−236. https://doi.org/10.1109/taffc.2017.2695999
60. M. Verma, S. K. Vipparthi, G. Singh, S. Murala, LEARNet: Dynamic imaging network for micro expression recognition, *IEEE Trans. Image Process.*, **29** (2019), 1618−1627. https://doi.org/10.1109/tip.2019.2912358
61. B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, et al., Recognizing spontaneous micro-expression using a three-stream convolutional neural network, *IEEE Access*, **7** (2019), 184537−184551. https://doi.org/10.1109/access.2019.2960629
62. Z. Xia, X. Hong, X. Gao, X. Feng, G. Zhao, Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions, *IEEE Trans. Multimedia*, **22** (2019), 626−640. https://doi.org/10.1109/tmm.2019.2931351

63. M. Peng, Z. Wu, Z. Zhang, T. Chen, From macro to micro expression recognition: Deep learning on small datasets using transfer learning, in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, (2018), 657−661. https://doi.org/10.1109/fg.2018.00103
64. L. Ma, T. Tan, Y. Wang, D. Zhang, Efficient iris recognition by characterizing key local variations, *IEEE Trans. Image Process.*, **13** (2004), 739−750. https://doi.org/10.1109/tip.2004.827237
65. Z. N. Sun, T. N. Tan, Ordinal measures for iris recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, **31** (2009), 2211−2226. https://doi.org/10.1109/tpami.2008.240
66. Z. F. He, T. N. Tan, Z. N. Sun, X. Qiu, Towards accurate and fast iris segmentation for iris biometrics, *IEEE Trans. Pattern Anal. Mach. Intell.*, ***31*** (2009), 1670−1684. https://doi.org/10.1109/tpami.2008.183
67. T. N. Tan, Z. F. He, Z. N. Sun, Efficient and robust segmentation of noisy iris images for non-cooperative iris recognition, *Image Vision Comput.*, **28** (2010), 223−230. https://doi.org/10.1016/j.imavis.2009.05.008
68. P. J. Phillips, K. W. Bowyer, P. J. Flynn, X. Liu, W. T. Scruggs, The iris challenge evaluation 2005, in *Proceedings of the 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, (2008), 1−8. https://doi.org/10.1109/btas.2008.4699333
69. S. Shah, A. Ross, Generating synthetic irises by feature agglomeration, in *Proceedings of the IEEE International Conference on Image Processing*, (2006), 317−320. https://doi.org/10.1109/icip.2006.313157
70. M. Tonsen, X. C. Zhang, Y. Sugano, A. Bulling, Labelled pupils in the wild: A dataset for studying pupil detection in unconstrained environments, in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications*, (2016), 139−142. https://doi.org/10.1145/2857491.2857520
71. M. Dobes, J. Martinek, D. Skoupil, Z. Dobesova, J. Pospisil, Human eye localization using the modified Hough transform, *Optik*, **117** (2006), 468−473. https://doi.org/10.1016/j.ijleo.2005.11.008
72. H. Proenca, L. A. Alexandre, UBIRIS: A noisy iris image database, in *Proceedings of the 13 International Conference on Image Analysis and Processing*, (2005), 970−977. https://doi.org/10.1007/11553595_119
73. H. Proenca, S. Filipe, R. Santos, J. Oliveira, L. A. Alexandre, The UBIRIS.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance, *Trans. Pattern Anal. Mach. Intell.*, **32** (2009), 1529−1535. https://doi.org/10.1109/tpami.2009.66
74. W. Fuhl, G. Kasneci, E. Kasneci, TEyeD: Over 20 million real-world eye image with pupil, Eyelid, and Iris 2D and 3D segmentations, 2D and 3D landmarks, 3D eyeball, gaze vector, and eye movement types, in *Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality*, (2021), 367−375. https://doi.org/10.1109/ismar52148.2021.00053
75. G. Sun, J. Zhang, K. Zheng, X. Fu, Eye tracking and roi detection within a computer screen using a monocular camera, *J. Web Eng.*, (2020), 1117−1146. https://doi.org/10.13052/jwe1540-9589.19789
76. G. Heusch, A. Anjos, S. Marcel, A reproducible study on remote heart rate measurement, preprint, arXiv: 1709.00962.

77. G. G. Hsu, A. Ambikapathi, M. S. Chen, Deep learning with time-frequency representation for pulse estimation from facial videos, in *Proceedings of the 2017 IEEE International Joint Conference on Biometrics*, (2017), 383−389. https://doi.org/10.1109/btas.2017.8272721
78. R. Stricker, S. Muller, H. M. Gross, Non-contact video-based pulse rate measurement on a mobile service robot, in *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, (2014), 1056−1062. https://doi.org/10.1109/roman.2014.6926392
79. S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, *Pattern Recogn. Lett.*, **124** (2019), 82−90. https://doi.org/10.1016/j.patrec.2017.10.017
80. X. Niu, H. Han, S. Shan, X. Chen, VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video, in *Proceedings of the Asian Conference on Computer Vision*, (2018), 562−576. https://doi.org/10.1007/978-3-030-20873-8_36
81. X. Li, H. Han, H. Lu, X. Niu, Z. Yu, A. Dantcheva, et al., The 1st challenge on remote physiological signal sensing, in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2020), 1274−1281. https://doi.org/10.1109/cvprw50498.2020.00165
82. Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, et al., Multimodal spontaneous emotion corpus for human behavior analysis, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 3438−3446. https://doi.org/10.1109/CVPR.2016.374
83. E. M. Nowara, T. K. Marks, H. Mansour, A. Veeraraghavan, Near-infrared imaging photoplethysmography during driving, *IEEE Trans. Intell. Trans. Syst.*, **23** (2022), 3589−3600. https://doi.org/10.1109/tits.2020.3038317
84. E. M. Nowara, T. K. Marks, H. Mansour, SparsePPG: Towards driver monitoring using camera-based vital signs estimation in near-infrared, in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (2018), 1353−1362. https://doi.org/10.1109/cvprw.2018.00174
85. X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, et al., The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection, in *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, (2018), 242−249. https://doi.org/10.1109/fg.2018.00043
86. Y. C. Chou, B. Y. Ye, H. R. Chen, Y. H. Lin, A real-time and non-contact pulse rate measurement system on fitness equipment, *IEEE Trans. Instrum. Meas.*, **71** (2021), 1−11. https://doi.org/10.1109/TIM.2021.3136173
87. Q. V. Tran, S. F. Su, W. Sun, M. Q. Tran, Adaptive pulsatile plane for robust noncontact heart rate monitoring, *IEEE Trans. Syst. Man Cybern.*, **51** (2021), 5587−5599. https://doi.org/10.1109/TSMC.2019.2957159
88. R. Belaiche, R. M. Sabour, C. Migniot, Y. Benezeth, D. Ginhac, K. Nakamura, et al., Emotional state recognition with micro-expressions and pulse rate variability, in *Proceedings of the 20th International Conference on Image Analysis and Processing*, (2019), 26−35. https://doi.org/10.1007/978-3-030-30642-7_3

89. R. M. Sabour, Y. Benezeth, F. Marzani, K. Nakamura, R. Gomez, F. Yang, Emotional state classification using pulse rate variability, in *Proceedings of the 4th International Conference on Signal and Image Processing*, (2019), 86−90. https://doi.org/10.1109/siprocess.2019.8868781
90. F. Bevilacqua, H. Engstrom, P. Backlund, Game-calibrated and user-tailored remote detection of stress and boredom in games, *Sensors-Basel*, **19** (2019), 2877. https://doi.org/10.3390/s19132877
91. K. Zheng, K. Ci, H. Li, L. Shao, G. Sun, J. Liu, et al., Heart rate prediction from facial video with masks using eye location and corrected by convolutional neural networks, *Biomed. Signal Process.*, **75** (2022), 103609. https://doi.org/10.1016/j.bspc.2022.103609
92. K. Zheng, K. Ci, J. Cui, J. Hong, J. Zhou, Non-contact heart rate detection when face information is missing during online learning, *Sensors-Basel*, **20** (2020), 7021. https://doi.org/10.3390/s20247021
93. K. Zheng, J. J. Shen, G. M. Sun, H. Li, Y. Li, Shielding facial physiological information in video, *Math. Biosci. Eng.*, **19** (2022), 5153−5168. https://doi.org/10.3934/mbe.2022241
94. S. K. A. Prakash, C. S. Tucker, Bounded Kalman filter method for motion-robust, non-contact heart rate estimation, *Biomed. Opt. Express*, **9** (2018), 873−897. https://doi.org/10.1364/boe.9.000873
95. Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, A. El Saddik, EVM-CNN: Real-time contactless heart rate estimation from facial video, *IEEE Trans. Multimedia*, **21** (2018), 1778−1787. https://doi.org/10.1109/tmm.2018.2883866
96. W. J. Han, H. F. Li, H. B. Ruan, L. Ma, Review on speech emotion recognition, *J. Software*, **25** (2014), 37−50. https://doi.org/10.13328/j.cnki.jos.004497
97. S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English, *PLoS One*, **13** (2018), e0196391, https://doi.org/10.1371/journal.pone.0196391
98. Y. Wang, L. Guan, Recognizing human emotional state from audiovisual signals, *IEEE Trans. Multimedia*, **10** (2008), 659−668. https://doi.org/10.1109/tmm.2008.927665
99. S. Zhalehpour, O. Onder, Z. Akhtar, C. E. Erdem, BAUM-1: A spontaneous audio-visual face database of affective and mental states, *IEEE Trans. Affect. Comput.*, **8** (2017), 300−313. https://doi.org/10.1109/taffc.2016.2553038
100. C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, et al., IEMOCAP: Interactive emotional dyadic motion capture database, *Lang. Resour. Eval.*, **42** (2008), 335−359. https://doi.org/10.1007/s10579-008-9076-6
101. A. Metallinou, Z. Yang, C. C. Lee, C. Busso, S. Carnicke, S. Narayanan, The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations, *Lang. Resour. Eval.*, **50** (2016), 497−521. https://doi.org/10.1007/s10579-015-9300-0
102. M. Grimm, K. Kroscher, S. Narayanan, The Vera am Mittag German audio-visual emotional speech database, in *Proceedings of 2008 IEEE International Conference on Multimedia and Expo*, (2008), 865−868. https://doi.org/10.1109/icme.2008.4607572
103. G. Mckown, M. Valstar, R. Cowie, M. Pantic, M. Schroder, The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent, *IEEE Trans. Affect. Comput.*, **3** (2012), 5−17. https://doi.org/10.1109/t-affc.2011.20

104. F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, in *Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, (2013), 1−8. https://doi.org/10.1109/fg.2013.6553805
105. V. V. Nanavare, S. K. Jagtap, Recognition of human emotions from speech processing, *Procedia Comput. Sci.*, **49** (2015), 24−32. https://doi.org/10.1016/j.procs.2015.04.223
106. P. Vasuki, C. Aravindan, Improving emotion recognition from speech using sensor fusion techniques, in *Proceedings of TENCON 2012 IEEE Region 10 Conference*, (2012), 1−6. https://doi.org/10.1109/tencon.2012.6412330
107. X. L. Zhao, Q. R. Mao, Y. Z. Zhan, New method of speech emotion recognition fusing functional paralanguages, *J. Front. Comput. Sci. Technol.*, **8** (2014), 186−199. https://doi.org/10.3778/j.issn.1673-9418.1309002
108. J. H. Hsu, M. H. Su, C. H. Wu, Y. H. Chen, Speech emotion recognition considering nonverbal vocalization in affective conversations, *IEEE-ACM Trans. Audio Speech Lang. Process.*, **29** (2021), 1675−1686. https://doi.org/10.1109/taslp.2021.3076364
109. S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, *IEEE Trans. Multimedia*, **20** (2017), 1576−1590. https://doi.org/10.1109/tmm.2017.2766843
110. Z. M. Wang, G. Liu, H. Song, Speech emotion recognition method based on multiple kernel learning feature fusion, *Comput. Eng.*, **45** (2019), 248−254. https://doi.org/10.19678/j.issn.1000-3428.0053232
111. J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, V. Tarokh, Speech emotion recognition with dual-sequence LSTM architecture, in *IEEE International Conference on Acoustics, Speech and Signal Processing* , (2020), 6474−6478. https://doi.org/10.1109/icassp40776.2020.9054629
112. J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, *Biomed. Signal Process.*, **47** (2019), 312−323. https://doi.org/10.1016/j.bspc.2018.08.035
113. O. Atila, A. Sengur, Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition, *Appl. Acoust.*, **182** (2021), 108260. https://doi.org/10.1016/j.apacoust.2021.108260
114. X. Wu, Y. Cao, H. Lu, S. Liu, D. Wang, Z. Wu, et al., Speech emotion recognition using sequential capsule networks, *IEEE-ACM Trans. Audio Speech Lang. Process.*, **29** (2021), 3280−3291. https://doi.org/10.1109/taslp.2021.3120586
115. I. Shahin, N. Hindawi, A. B. Nassif, A. Alhudhaif, K. Polat, Novel dual-channel long short-term memory compressed capsule networks for emotion recognition, *Expert Syst. Appl.*, **188** (2022), 116080. https://doi.org/10.1016/j.eswa.2021.116080
116. S. Zhang, R. Liu, Y. Yang, X. Zhao, J. Yu, Unsupervised domain adaptation integrating transformer and mutual information for cross-corpus speech emotion recognition, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 120−129. https://doi.org/10.1145/3503161.3548328
117. D. Jing, T. Manting, Z. Li, Transformer-like model with linear attention for speech emotion recognition, *J. Southeast Univ. (Engl. Ed.)*, **37** (2021), 164−170. https://doi.org/10.3969/j.issn.1003-7985.2021.02.005
118. J. Lei, X. Zhu, Y. Wang, BAT: Block and token self-attention for speech emotion recognition, *Neural Networks*, **156** (2022), 67−80. https://doi.org/10.1016/j.neunet.2022.09.022

119. L. Yi, M. W. Mak, Improving speech emotion recognition with adversarial data augmentation network, *IEEE Trans. Neur. Net. Learn. Syst.*, **33** (2020), 172−184. https://doi.org/10.1109/tnnls.2020.3027600
120. Z. Yucel, S. Koyama, A. Monden, M. Sasakura, Estimating level of engagement from ocular landmarks, *Int. J. Hum. Comput. Int.*, **36** (2020), 1527−1539. https://doi.org/10.1080/10447318.2020.1768666
121. Z. Pi, M. Chen, F. Zhu, J. Yang, W. Hu, Modulation of instructor's eye gaze by facial expression in video lectures, *Innov. Educ. Teach. Int.*, **59** (2022), 15−23. https://doi.org/10.1080/14703297.2020.1788410
122. M. Mahmoud, P. Robinson, Interpreting hand-over-face gestures, in *International Conference on Affective Computing and Intelligent Interaction*, (2011), 248−255. https://doi.org/10.1007/978-3-642-24571-8_27
123. K. Zheng, J. Kong, L. Tian, B. Li, H. Li, J. Zhou, Hand-over-face occlusion and distance adaptive heart rate detection based on imaging photoplethysmography and pixel distance in online learning, *Biomed. Signal Process.*, **85** (2023), 104898, https://doi.org/10.1016/j.bspc.2023.104898
124. M. Haghighat, M. Abdel-Mottaleb, W. Alhalabi, Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition, *IEEE Trans. Inf. Forensics Secur.*, **11** (2016), 1984−1996. https://doi.org/10.1109/tifs.2016.2569061
125. S. Koelstra, C. Muehl, M. Soleymani, A. Yazdani, T. Ebrahimi, T. Pun, et al., DEAP: A database for emotion analysis using physiological signals, *IEEE Trans. Affect. Comput.*, **3** (2012), 18−31. https://doi.org/10.1109/t-affc.2011.15
126. A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, L. P. Morency, Multi-attention recurrent network for human communication comprehension, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2018), 5642−5649. https://doi.org/10.1609/aaai.v32i1.12024
127. W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, K. Yang, CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020), 3718−3727. https://doi.org/10.18653/v1/2020.acl-main.343
128. N. Xu, W. Mao, G. Chen, Multi-interactive memory network for aspect based multimodal sentiment analysis, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2019), 371−378. https://doi.org/10.1609/aaai.v33i01.3301371
129. Y. Baveye, E. Dellandrea, C. Chamaret, LIRIS-ACCEDE: A video database for affective content analysis, *IEEE Trans. Affect. Comput.*, **6** (2015), 43−55. https://doi.org/10.1109/taffc.2015.2396531
130. M. Soleymani, J. Lichtenauer, T. Pun, A multimodal database for affect recognition and implicit tagging, *IEEE Trans. Affect. Comput.*, **3** (2012), 42−55. https://doi.org/10.1109/t-affc.2011.25
131. O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE'05 audio-visual emotion database, in *Proceedings of the 22nd International Conference on Data Engineering Workshops*, (2006). https://doi.org/10.1109/icdew.2006.145
132. H. Zhou, J. Du, Y. Zhang, Q. Wang, Q. F. Liu, C. H. Lee, Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition, *IEEE-ACM Trans. Audio Speech Lang. Process.*, **29** (2021), 2617−2629. https://doi.org/10.1109/taslp.2021.3096037

133. M. Wu, W. Su, L. Chen, W. Pedrycz, K. Hirota, Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition, *IEEE Trans. Affect. Comput.*, **13** (2020), 805−817. https://doi.org/10.1109/taffc.2020.2966440
134. J. Chen, Z. Chen, Z. Chi, H. Fu, Facial expression recognition in video with multiple feature fusion, *IEEE Trans. Affect. Comput.*, **9** (2018), 38−50. https://doi.org/10.1109/taffc.2016.2593719
135. Y. Kim, E. M. Provost, ISLA: Temporal segmentation and labeling for audio-visual emotion recognition, *IEEE Trans. Affect. Comput.*, **10** (2017), 196−208. https://doi.org/10.1109/taffc.2017.2702653
136. P. Bhattacharya, R. K. Gupta, Y. P. Yang, Exploring the contextual factors affecting multimodal emotion recognition in videos, *IEEE Trans. Affect. Comput.*, **14** (2023), 1547−1557. https://doi.org/10.1109/taffc.2021.3071503
137. L. Vaiani, M. L. Quatra, L. Cagliero, P. Garza, ViPER: Video-based perceiver for emotion recognition, in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, (2022), 67−73. https://doi.org/10.1145/3551876.3554806
138. Y. Wu, Z. Y. Zhang, P. Peng, Y. Y. Zhao, B. Qin, Leveraging multi-modal interactions among the intermediate representations of deep transformers for emotion recognition, in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, (2022), 101−109. https://doi.org/10.1145/3551876.3554813
139. D. K. Yang, S. Huang, H. P. Kuang, Disentangled representation learning for multimodal emotion recognition, in *Proceedings of the 30th ACM International Conference on Multimedia*, (2022), 1642−1651. https://doi.org/10.1145/3503161.3547754
140. Y. P. Liu, W. Sun, X. Zhang, Y. B. Qin, Improving dimensional emotion recognition via feature-wise fusion, in *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, (2022), 55−60. https://doi.org/10.1145/3551876.3554804
141. M. Y. Tsalamlal, M. A. Amorim, J. C. Martin, M. Ammi, Combining facial expression and touch for perceiving emotional valence, *IEEE Trans. Affect. Comput.*, **9** (2018), 437−449. https://doi.org/10.1109/taffc.2016.2631469
142. Y. Yang, Q. Gao, Y. Song, X. L. Song, Z. M. Mao, J. J. Liu, Investigating of deaf emotion cognition pattern by EEG and facial expression combination, *IEEE J. Biomed. Health*, **26** (2022), 589−599. https://doi.org/10.1109/jbhi.2021.3092412
143. Siddharth, T. P. Jung, T. J. Sejnowski, Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing, *IEEE Trans. Affect. Comput.*, **13** (2022), 96−107. https://doi.org/10.1109/taffc.2019.2916015
144. N. Braunschweiler, R. Doddipatla, S. Keizer, S. Stoyanchev, Factors in emotion recognition with deep learning models using speech and text on multiple corpora, *IEEE Signal Proc. Lett.*, **29** (2022), 722−726. https://doi.org/10.1109/lsp.2022.3151551
145. X. Zhang, J. Liu, J. Shen, S. Li, K. Hou, B. Hu, et al., Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine, *IEEE Trans. Cybern.*, **51** (2021), 4386−4399. https://doi.org/10.1109/tcyb.2020.2987575
146. Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, C. Chen, HetEmotionNet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition, in *Proceedings of the 29th ACM International Conference on Multimedia*, (2021), 1047−1056. https://doi.org/10.1145/3474085.3475583

147. M. Soleymani, M. Pantic, T. Pun, Multimodal emotion recognition in response to videos, *IEEE Trans. Affect. Comput.*, **3** (2011), 211−223. https://doi.org/10.1109/t-affc.2011.37
148. W. L. Zheng, W. Liu, Y. Lu, B. L. Lu, A. Cichocki, Emotionmeter: A multimodal framework for recognizing human emotions, *IEEE Trans. Cybern.*, **49** (2018), 1110−1122. https://doi.org/10.1109/tcyb.2018.2797176
149. Q. Wang, M. Wang, Y. Yang, X. Zhang, Multi-modal emotion recognition using EEG and speech signals, *Comput. Biol. Med.*, **149** (2022), 105907. https://doi.org/10.1016/j.compbiomed.2022.105907
150. S. Scrimin, U. Moscardino, L. Finos, L. Mason, Effects of psychophysiological reactivity to a school-related stressor and temperament on early adolescents' academic performance, *J. Early Adolesc.*, **39** (2019), 904−931. https://doi.org/10.1177/0272431618797008
151. B. Cowley, N. Ravaja, T. Heikura, Cardiovascular physiology predicts learning effects in a serious game activity, *Comput. Educ.*, **60** (2013), 299−309. https://doi.org/10.1016/j.compedu.2012.07.014
152. K. N. Cranford, J. M. Tiettmeyer, B. C. Chuprinko, S. Jordan, N. P. Grove, Measuring load on working memory: The use of heart rate as a means of measuring chemistry students' cognitive load, *J. Chem. Educ.*, **91** (2014), 641−647. https://doi.org/10.1021/ed400576n
153. N. Thompson, T. J. McGill, Genetics with Jean: The design, development and evaluation of an affective tutoring system, *Educ. Technol. Res.*, **65** (2017), 279−299. https://doi.org/10.1007/s11423-016-9470-5
154. A. Versluis, B. Verkuil, P. Spinhoven, J. F. Brosschot, Feasibility and effectiveness of a worry-reduction training using the smartphone: A pilot randomised controlled trial, *Br. J. Guid. Couns.*, **48** (2020), 227−239. https://doi.org/10.1080/03069885.2017.1421310
155. K. Fromel, Z. Svozil, F. Chmelik, L. Jakubec, D. Groffik, The role of physical education lessons and recesses in school lifestyle of adolescents, *J. School Health*, **86** (2016), 143−151. https://doi.org/10.1111/josh.12362
156. M. Slingerland, L. Haerens, G. Cardon, L. Borghouts, Differences in perceived competence and physical activity levels during single-gender modified basketball game play in middle school physical education, *Eur. Phys. Educ. Rev.*, **20** (2014), 20−35. https://doi.org/10.1177/1356336x13496000
157. P. Klein, J. Viiri, S. Mozaffari, A. Dengel, J. Kuhn, Instruction-based clinical eye-tracking study on the visual interpretation of divergence: How do students look at vector field plots?, *Phys. Rev. Phys. Educ. Res.*, **14** (2018), 010116. https://doi.org/10.1103/physrevphyseducres.14.010116
158. A. I. Molina, O. Navarro, M. Ortega, M. Lacruz, Evaluating multimedia learning materials in primary education using eye tracking, *Comput. Stand. Int.*, **59** (2018), 45−60. https://doi.org/10.1016/j.csi.2018.02.004
159. L. Mason, P. Pluchino, M. C. Tornatora, Using eye-tracking technology as an indirect instruction tool to improve text and picture processing and learning, *Br. J. Educ. Technol.*, **47** (2016), 1083−1095. https://doi.org/10.1111/bjet.12271
160. M. Van Wermeskerken, T. Van Gog, Seeing the instructor's face and gaze in demonstration video examples affects attention allocation but not learning, *Comput. Educ.*, **113** (2017), 98−107. https://doi.org/10.1016/j.compedu.2017.05.013

161. V. Clinton, J. L. Cooper, J. E. Michaelis, M. W. Alibali, M. J. Nathan, How revisions to mathematical visuals affect cognition: Evidence from eye tracking, in *Eye-Tracking Technology Applications in Educational Research*, (2017), 195−218. https://doi.org/10.4018/978-1-5225-1005-5.ch010
162. Y. C. Jian, Eye-movement patterns and reader characteristics of students with good and poor performance when reading scientific text with diagrams, *Reading. Writing.*, **30** (2017), 1447−1472. https://doi.org/10.1007/s11145-017-9732-6
163. J. M. Karch, J. C. Garcia Valles, H. Sevian, Looking into the black box: Using gaze and pupillometric data to probe how cognitive load changes with mental tasks, *J. Chem. Educ.*, **96** (2019), 830−840. https://doi.org/10.1021/acs.jchemed.9b00014
164. K. Krstic, A. Soskic, V. Kovic, K. Holmqvist, All good readers are the same, but every low-skilled reader is different: an eye-tracking study using PISA data, *Eur. J. Psychol. Educ.*, **33** (2018), 521−541. https://doi.org/10.1007/s10212-018-0382-0
165. X. Zhu, Z. Chen, Dual-modality spatiotemporal feature learning for spontaneous facial expression recognition in e-learning using hybrid deep neural network, *Vis. Comput.*, **36** (2020), 743−755. https://doi.org/10.1007/s00371-019-01660-3
166. B. T. Shobana, G. A. Kumar, I-Quiz: An intelligent assessment tool for non-verbal behaviour detection, *Comput. Syst. Sci. Eng.*, **40** (2022), 1007−1021. https://doi.org/10.32604/csse.2022.019523
167. T. S. Ashwin, R. M. R. Guddeti, Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework, *User Model. User-Adap. Int.*, **30** (2020), 759−801. https://doi.org/10.1007/s11257-019-09254-3
168. I. Alkabbany, A. Ali, A. Farag, I. Bennett, M. Ghanoum, A. Farag, Measuring student engagement level using facial information, in *2019 IEEE International Conference on Image Processing (ICIP)*, (2019), 3337−3341. https://doi.org/10.1109/icip.2019.8803590

AIMS Press