IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 2, April, 2025, pp. 1029-1039
Journal homepage: http://iieta.org/journals/ts

Enhancing the Quality of Teacher–Student Interaction in Online English Instruction Through Facial Expression Recognition

Yun He

College of Foreign Languages, Quzhou University, Quzhou 324000, China

Corresponding Author Email: 33043@qzc.edu.cn

https://doi.org/10.18280/ts.420236

Received: 28 August 2024
Revised: 3 February 2025
Accepted: 17 March 2025
Available online: 30 April 2025

Keywords:
online English instruction, facial expression recognition, teacher–student interaction, multi-scale fused feature network, sentiment analysis

ABSTRACT

With the rapid advancement of information technology, online English instruction has emerged as a significant educational modality. However, insufficient teacher–student interaction remains a prevalent issue, negatively affecting instructional efficacy and student engagement. Facial expressions, as direct manifestations of emotional and cognitive responses, offer valuable insights into students' affective states during learning. In recent years, facial expression recognition (FER) technology has been increasingly integrated into educational contexts to enable real-time monitoring of emotional fluctuations and to provide feedback for teachers, thereby optimizing classroom interaction and improving teaching effectiveness. Despite prior investigations into the educational applications of FER, current approaches often suffer from limited accuracy and latency in complex online environments. Moreover, existing interaction strategies have largely remained unidirectional, lacking comprehensive interaction strategies based on FER results. How to improve the accuracy of FER and enhance the quality of teacher-student interaction in online English instruction through this technology remains a key issue that urgently needs to be addressed. Therefore, this study, which can be divided into two principal components, was conducted to address these challenges. First, an FER method based on a multi-scale fused feature network was proposed, with the objective of enhancing both accuracy and real-time performance in online English instruction. Second, based on the results of FER, strategies for improving the quality of teacher-student interaction in online English instruction were studied, and how to optimize the interaction process between teachers and students through facial expression data was explored. This study not only introduces technical innovations to support interaction in online English instruction but also expands the practical applications of FER technology within the broader field of education.

1. INTRODUCTION

With the rapid advancement of information technology, online education has increasingly emerged as a mainstream instructional modality [1-3], particularly within the field of English language instruction. By removing the spatial and temporal constraints of traditional classrooms, online English instruction has enabled flexible and diverse learning pathways [4–6]. However, the widespread adoption of online teaching has simultaneously highlighted persistent challenges, notably the insufficiency of teacher–student interaction [7, 8]. This deficiency has been shown to adversely affect both student motivation and instructional effectiveness. Facial expressions, as essential indicators of human emotional and cognitive responses [9], are capable of directly reflecting students' affective states and levels of attentiveness during the learning process. Consequently, the integration of FER technology into online English instruction has emerged as a critical area of inquiry, with the aim of improving the quality of teacher–student interaction.

In relevant research, FER has been recognized as an effective tool for enhancing the quality of educational interaction [10-13]. By enabling the real-time detection of students' facial expressions, FER allows teachers to better interpret learners' emotional fluctuations and engagement levels, thereby facilitating the dynamic adjustment of instructional strategies to enhance pedagogical outcomes. Furthermore, FER has been applied in sentiment analysis [14, 15], assisting educators in identifying emotional variations and enabling timely modifications to lesson pacing and instructional methods, thus promoting improved learning effectiveness. Accordingly, the investigation of FER-based interaction mechanisms in online education holds substantial theoretical significance and practical value.

Although the application of FER in educational contexts has been explored in prior studies, several limitations persist in existing methodologies. A considerable proportion of current FER techniques remain rooted in conventional image processing approaches [16, 17], which often overlook the importance of multi-scale feature fusion. This oversight has resulted in reduced accuracy and compromised real-time performance when applied in complex online instructional environments. Moreover, existing strategies for enhancing interaction have predominantly focused on unidirectional

feedback from teachers to students [18, 19], lacking the integration of FER-informed multidimensional interaction frameworks. Consequently, the effectiveness of such strategies in improving interaction quality has remained constrained. Therefore, how to improve the accuracy of FER technology and effectively apply it to online English instruction remains an urgent research problem to be solved.

To address these gaps, the present study was conducted, which comprises two principal components. First, an FER method based on a multi-scale fused feature network was proposed to enhance both the accuracy and real-time responsiveness of FER in online English instruction. Second, strategies for improving teacher–student interaction quality were investigated, utilizing FER results to optimize the interaction process between teachers and students. Through these two focal areas, this research seeks to offer novel technological support and theoretical grounding for advancing interaction quality in online English instruction, thereby contributing to the broader development of online education.

2. FER FOR ONLINE ENGLISH INSTRUCTION BASED ON A MULTI-SCALE FUSED FEATURE NETWORK

In online English instruction, facial expressions of both teachers and students serve as critical indicators influencing the quality of instructional interaction. Real-time observation of students' facial expressions enables teachers to assess emotional and attentional states, thereby allowing for timely adjustment of instructional strategies to enhance interactive effectiveness. However, FER in complex teaching scenarios is often challenged by various factors, including lighting conditions, camera angles, and inter-individual differences. High inter-class similarity among expressions further contributes to reduced recognition accuracy. To improve the recognition performance of facial expressions in online English instruction, it is essential to design a neural network model that is not only lightweight and accurate but also robust to complex variations in facial expressions. In this context, a blueprint-separable dilated convolution network module was designed in this study, which has significant application value. This module facilitates architectural optimization, parameter reduction, and enhanced generalization capability, thereby addressing the deficiencies of existing FER methods in terms of accuracy and computational efficiency. The network architecture designed for FER based on multi-scale fused features is illustrated in Figure 1.

Specifically, the integration of the Spatial Group-wise Enhance (SGE) attention mechanism plays a critical role in enhancing the precision of FER. In online English instruction scenarios, teachers are required to obtain students' emotional responses in real time to facilitate timely pedagogical adjustments. Given that changes in facial expressions are often subtle and transient, conventional convolutional neural networks (CNNs) frequently fail to capture key localized features effectively. By incorporating the SGE attention mechanism, the network is endowed with the capability to automatically analyze and amplify facial features that are contextually relevant to learning-related emotions, while simultaneously suppressing irrelevant or noisy features.

A blueprint-separable dilated convolution module was specifically designed for multi-scale feature fusion in teacher–student FER in this study, which plays a crucial role in improving recognition accuracy. In online English instruction, interaction typically occurs via video conferencing, where facial expression features may appear at varying scales due to differences in camera angles and environmental conditions. Conventional convolution operations often fail to effectively process such diversified features. In contrast, the adoption of the blueprint-separable dilated convolution module enables an expanded receptive field of the convolution kernel, thereby facilitating the capture of more comprehensive facial expression information.

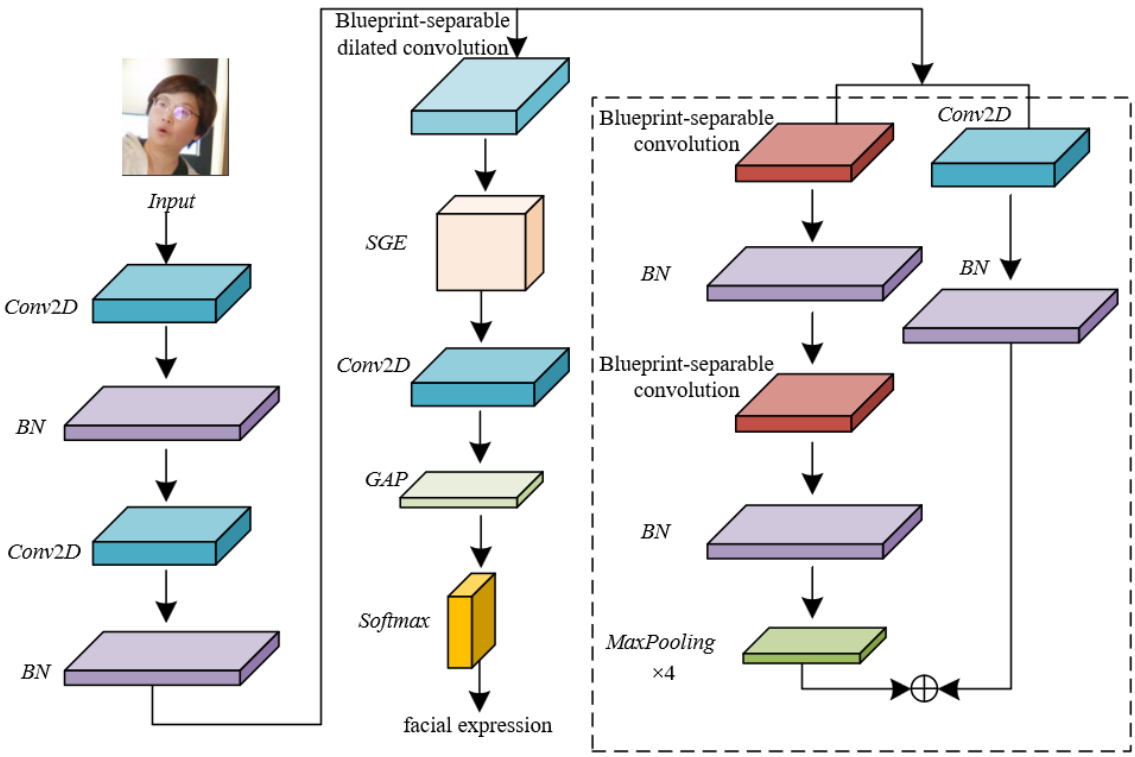


Figure 1. Network architecture for teacher–student FER

To address issues commonly associated with deep neural networks, such as overfitting, the fully connected layer was replaced with a global average pooling layer, and a cross-layer connection mechanism was incorporated. These modifications have significantly improved the model's generalization capability. In online English instructional settings, considerable variability exists in both instructional content and student emotional responses. Overfitting phenomenon may lead to unstable performance of the network when processing new data. The use of the global average pooling layer effectively reduces the number of model parameters, thereby lowering the risk of overfitting and enhancing the robustness of the network. Furthermore, the integration of the cross-layer connection mechanism facilitates more effective information flow between low-level and high-level features, fostering tighter coupling across feature hierarchies. This strategy can improve the adaptability and accuracy of the model across diverse students and instructional scenarios. As a result, FER in online English instruction can be performed with greater precision, thereby supporting improved teacher–student interaction quality.

2.1 Attention mechanism module

In the context of FER in online English instruction, the SGE attention mechanism was employed to optimize the network's responsiveness to students' expressions through weighted feature maps. In real instructional environments, students' facial expressions tend to be transient and subtle. Traditional CNNs often struggle to accurately capture these nuances, particularly in emotionally complex expressions such as tension, confusion, or fatigue, which frequently exhibit high inter-class similarity. The SGE mechanism addresses this challenge by applying Gaussian normalization to the variance feature maps, enabling the network to automatically identify regions sensitive to emotional variation while effectively suppressing background noise or irrelevant expression changes. This selective attention allows the network to focus on critical facial areas—such as the eyes and mouth—thereby improving the accuracy of emotion recognition. The implementation of this module follows a structured process:

(a) Initially, feature information within the image is divided into groups, followed by the application of global average pooling to extract global features. Each image region is then subjected to a pointwise multiplication operation to compute initial attention weights relative to the global context. In online English instructional settings, the recognition of students' facial expressions is often complicated by environmental factors such as lighting, posture, and background variability. Global average pooling serves to mitigate the influence of such variations, ensuring the preservation of holistic information. The subsequent element-wise multiplication enables the network to contrast global and local features to determine which regions of the face convey the most salient expressions. For instance, subtle micro-expressions such as confusion or engagement, which are often localized to specific facial regions, can be effectively identified through this operation.

(b) The attention-weighted masks computed by the SGE mechanism are subjected to normalization, and each feature group is independently scaled. This process ensures that no particular region receives disproportionate emphasis during the feature weighting stage, thus preserving the integrity of diverse expression information. In online English instruction, real-time interpretation of students' emotional changes is essential, particularly during extended sessions in which facial expressions may undergo subtle shifts, such as signs of fatigue or disengagement. If the network were to focus excessively on specific facial regions while neglecting others, the accuracy of FER would be compromised. Through normalization, the SGE mechanism enables equitable weighting of expression features across all regions. The subsequent scaling operation further adjusts the relative importance of each region's features, enhancing the network's adaptability and flexibility, and thereby improving recognition precision.

(c) The SGE attention mechanism also employs a sigmoid function to extract features that require more attention, followed by a scaling transformation for features of each region. The use of the sigmoid function enables finer modulation of the network's attention across regions, allowing context-sensitive calibration of feature importance. In online English instruction, students' facial expressions may vary dynamically in response to instructional content, interaction style, and classroom atmosphere. The sigmoid-based adjustment mechanism empowers the network to recalibrate attention weights in real time, focusing on regions indicative of emotional variation—such as a slight smile or furrowed brows. This context-aware adaptability enhances the network's sensitivity to subtle expressions, ensuring that teachers are provided with accurate and timely emotional feedback for precise pedagogical adjustments.

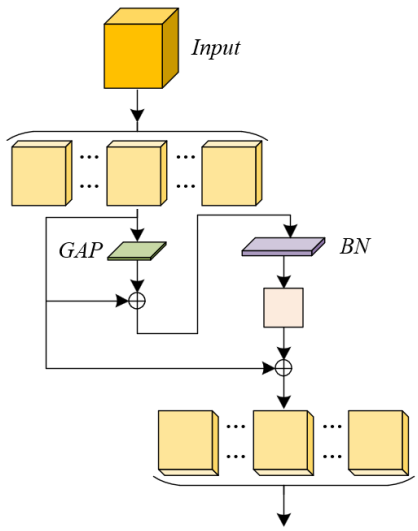


Figure 2. Network architecture of the SGE attention mechanism

Figure 2 illustrates the network architecture of the SGE attention mechanism. When implementing the SGE mechanism in FER tasks, particular attention must be paid to the partitioning of feature groups and the detailed processing of the network. The SGE mechanism processes the input feature map by dividing it into multiple feature groups, then computes the inter-feature similarity to generate attention coefficients. During this process, both the size and the quality of the feature map exert a substantial influence on the performance of the model. In online English instruction, facial expression details are frequently affected by the distance between students and teachers, the camera angle, and environmental variability. As such, the grouping of features must take into account the inherent diversity and complexity of these instructional settings. The granularity of feature group partitioning must be appropriately aligned with the

requirements of FER to ensure that critical facial regions—such as the eyes and mouth—are adequately emphasized. This targeted attention allows for the effective capture of micro-expressions such as anxiety or confusion, thereby enabling timely instructional adjustments by educators.

Furthermore, although the global average pooling operation in the SGE module enables the extraction of global feature information, reliance solely on global features may be insufficient for capturing the subtle variations in students' emotional states during FER. Accurate recognition of facial expressions often depends on fine-grained local information, such as minor muscle movements and region-specific expression changes. Accordingly, when the global average pooling function $D_h(\cdot)$ is applied, it must be ensured that the resulting global features, when integrated with local features, adequately preserve the subtle variations inherent in facial expressions. For example, slight changes around the eyes or the corners of the mouth are critical in identifying emotions. The attention-weighting capability of the SGE mechanism plays a pivotal role in amplifying such localized features, thereby enhancing recognition accuracy. Using global statistical features, the feature vector obtained after pooling is represented as:

$$h = D_h(\kappa) = \frac{1}{v}\sum_{u=1}^{v} a_u \tag{1}$$

The degree of similarity between the global feature a_u and a local feature h can be evaluated via an element-wise multiplication operation. The attention coefficients z_u for each feature can then be computed as follows:

$$z_u = h \cdot a_u \tag{2}$$

Normalization of the attention coefficients is another essential component of the SGE mechanism. In online English instructional contexts, students' expressions are subject to dynamic variation depending on content, situational context, and individual emotional states. Therefore, the normalization of the attention coefficients z_u is critical for preventing overfitting that may arise from expression disparities across samples. By normalizing the attention coefficients, the network is better equipped to accommodate a wide range of complex emotional states—such as anxiety or confusion encountered during learning difficulties—thus ensuring that the system can respond consistently and provide accurate feedback. Assuming the mean and standard deviation of the attention coefficients z_u are denoted by ω_z and δ_z, respectively, and the normalized attention coefficients are represented by $\hat{z}_u$. A small constant $\in$ was added for numerical stability. The following shows the specific normalization operation:

$$\omega_z = \frac{1}{v}\sum_{1}^{v} z_u \tag{3}$$

$$\delta_z^2 = \frac{1}{v}\sum_{k}^{v}(z_k - \omega_z)^2 \tag{4}$$

$$\hat{z}_u = \frac{z_u - \omega_z}{\delta_z + \in} \tag{5}$$

The design of the hyperparameters ε and α within the SGE attention mechanism plays a critical role in ensuring the model's flexibility and overall performance. In FER tasks, real-time tracking of students' emotional fluctuations is essential for effective instructional interaction. However, variations in teaching scenarios can lead to significant differences in facial expression features. For instance, during highly interactive speaking exercises, expressions may predominantly reflect tension and concentration, whereas listening tasks may elicit signs of confusion or fatigue. To accommodate such contextual variability, the hyperparameters ε and α in the SGE module were configured to match the number of feature groups, allowing for adaptive adjustment of feature weighting across different instructional contexts. Activation of the attention coefficients was performed using the sigmoid function, which refines the weighting and spatial scaling of each feature. This enables dynamic modulation of facial expression regions according to changes in situational demands, thereby enhancing both the robustness and accuracy of expression recognition. Such improvements support more accurate interpretation of student emotional states by teachers. Specifically, the transformation and scaling of the normalized attention coefficients $\hat{z}_u$ are governed by the following equation:

$$x_u = \varepsilon \hat{z}_u + \alpha \tag{6}$$

x_u was activated by the initial a_u via the sigmoid function $\delta(.)$. Then the activated x_u was element-wise multiplied with the original facial expression feature information. Scaling operation was performed in the spatial dimension using x_u.

$$\hat{a}_u = a_u \cdot \delta(x_u) \tag{7}$$

$$\hat{a} = \{\hat{a}_{1\ldots v}\}, \hat{a}_u \in \overline{E^{H}}, v = G \times Q \tag{8}$$

2.2 Improved dilated convolution module

Dilated convolution was utilized to expand the receptive field without increasing computational complexity by introducing a dilation rate into the convolution operation. This is particularly advantageous for FER, where subtle changes often occur in localized regions such as the corners of the mouth, eyes, or eyebrows. By applying dilated convolution, the network is enabled to capture broader contextual facial features without the need for additional network layers. This not only reduces the computational burden of convolutional operations but also improves the network's ability to interpret complex emotional fluctuations present in both teacher and student facial expressions. Accurate recognition of emotional states—such as confusion, concentration, or happiness—can thus be more effectively achieved. In the context of online English instruction, minimizing computational load is essential for improving real-time performance and responsiveness. Given that student facial expressions may shift rapidly during interactive sessions, a model capable of performing timely and accurate emotion recognition is required. To meet this demand, a lightweight blueprint-separable dilated convolution network was employed. This network structure was selected to ensure high recognition accuracy while simultaneously optimizing architectural efficiency, reducing resource consumption, and increasing processing speed.

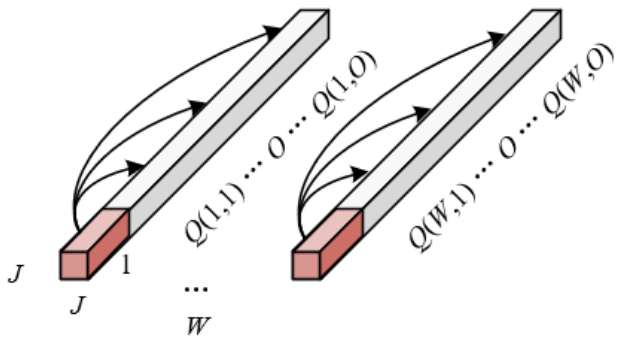


Figure 3. Schematic diagram of the blueprint-separable convolution

To support FER, a blueprint-separable convolution architecture was first designed to effectively extract critical features from subtle facial variations. FER tasks often require the extraction of multi-scale and multi-dimensional facial features. The blueprint-separable dilated convolution was adopted for this purpose, leveraging efficient pointwise and depthwise convolution operations. This allows the network to identify key expression-related features with reduced computational resource consumption, without compromising recognition accuracy. As a lightweight neural network module, the blueprint-separable dilated convolution is built upon the redundancy along the depth axis of the convolution kernel. By optimizing the convolutional computation process, both the number of parameters and computational complexity are significantly reduced, while maintaining or even enhancing the network's feature extraction capability. This module consists of two pointwise convolution layers and one depthwise convolution layer. Initially, a 1×1 pointwise convolution is applied to the input feature maps to transform them into a new set of weighted coefficients. This operation effectively captures the channel-level information from the input features. Subsequently, a second 1×1 convolution is used to further refine these coefficients, allowing for the extraction of multi-scale feature information. Throughout this process, the inherent redundancy of the convolution kernels is exploited to efficiently separate deep feature representations, leading to a considerable reduction in computational cost. A schematic illustration of the blueprint-separable convolution is provided in Figure 3.

The derivation of the constructed blueprint-separable convolution module is primarily oriented toward efficient facial feature extraction and computational optimization. A detailed breakdown is presented below.

To begin, consider the input feature map I with dimensions (Z,B,A), where Z denotes the number of channels, and B and A represent the spatial dimensions. The objective of the blueprint-separable convolution is to transform the input feature map I into an output feature map N with V output channels, maintaining the same spatial dimensions B and A. To achieve this transformation, V convolutional kernels were employed, denoted as $O^{(1)}$,$O^{(2)}$,$O^{(3)}$...$O^{(V)}$, where each kernel has dimensions (Z,C,C), with C representing the spatial size of the kernel. The central idea of this design lies in constructing each convolutional kernel through the combination of the blueprint and coefficients, which enables the efficient extraction of local facial expression features. Let the convolution operation be denoted by *. The computation step is as follows:

$$N_{v^*} = I * O^{(v)} \tag{9}$$

What distinguishes the blueprint-separable convolution from conventional approaches is the method by which the convolutional kernels are constructed. Each kernel $O^{(V)}$ is composed of a blueprint $O^{(v)}$, representing the structural pattern of the kernel, and a set of corresponding weight coefficients $\mu_{v,1}$, $\mu_{v,2}$, $\mu_{v,3}$, ..., $\mu_{v,Z}$. This decomposition facilitates a more flexible convolution operation. By decoupling the convolution kernel into a blueprint and components, both parameter count and computational complexity are significantly reduced, thereby improving the efficiency of FER. In the context of online English instruction, the rapid and accurate recognition of subtle student facial expressions is critical. The blueprint-separable convolution leverages this decomposition strategy to achieve efficient recognition while minimizing computational resource usage. Specifically, the blueprint-separable convolution kernel $O^{(v)}$ is expressed as:

$$O_{l^*}^{(v)} = \mu_{v,l} \cdot Y^{(v)} \tag{10}$$

Compared to conventional convolution, the blueprint-separable convolution significantly reduces the number of parameters. For a standard convolutional kernel of size (Z,C,C), the total number of parameters is Z×C×C. In contrast, the blueprint-separable convolution requires only C×C+Z parameters. By decomposing each convolutional kernel into a blueprint and a corresponding set of coefficients, the total parameter count is substantially reduced, which in turn lowers the computational burden. This characteristic is particularly important in FER tasks, especially within real-time online English instructional settings, where computational efficiency and responsiveness are critical. Through this parameter reduction, the blueprint-separable convolution network is capable of processing facial images more rapidly, thereby enhancing real-time performance without compromising the accuracy of expression recognition. Expanding the previous formulation yields:

$$N_{v^*} = \sum_{z=1}^{Z} \left(I_* * O_*^{(v)} \right) \tag{11}$$

Each element was computed from a specific channel z within the input convolutional kernel channel.

$$N_{v^*} = \sum_{z=1}^{Z} \left(I_* * \left(\mu_{v,z} \cdot Y^{(v)} \right) \right) \tag{12}$$

Given that $Y^{(v)}$ is independent of Z and that $\mu_{v,z}$ is a scalar, the above expression can be further expanded as:

$$N_{v^*} = \left(\sum_{z=1}^{Z} I_* \cdot \mu_{v,z} \right) * Y^{(v)} \tag{13}$$

$$N'_{v^*} = \sum_{z=1}^{Z} I_* \cdot \mu_{v,z} \tag{14}$$

By rearranging the weights $\mu_{v,1}$, $\mu_{v,2}$, $\mu_{v,3}$…$\mu_{v,z}$ into $\mu\tilde{}_v$ with Z×1×1, the resulting operation becomes:

$$N'_{v^*} = I * \tilde{\mu}_{v,z} \tag{15}$$

$$N_* = N'_* * Y^{(v)} \tag{16}$$

As indicated by the preceding formulations, the output N'_{v^*} of the blueprint-separable convolution—corresponding to the v-th channel—is obtained through a weighted summation across all input channels I. The weights $\mu_{v,1}$, $\mu_{v,2}$, $\mu_{v,3}$, …, $\mu_{v,z}$ correspond to each input channel and are computed using a 1×1 pointwise convolution. This operation is functionally equivalent to applying a pointwise convolution to perform a weighted summation of the input I, thereby integrating feature information across different channels. Simultaneously, the blueprint $Y^{(v)}$, with dimensions $1\times1\times C\times C$, operates analogously to a depthwise convolution and serves to extract feature information along the depth dimension. In FER tasks—particularly in dynamic and complex teacher–student interactions—accurate extraction of deep expression features is essential for identifying subtle variations. The structure of the blueprint-separable convolution is capable of capturing these nuanced changes efficiently, thereby supporting teachers in better understanding students' emotional states.

In the baseline lightweight neural network architecture, the replacement of conventional convolution with blueprint-separable dilated convolution can significantly reduce convolutional redundancy. Specifically, the original input feature map I possesses the dimensionality (Z, B, A). In the optimized blueprint-separable dilated convolution, the operation is divided into two stages: an initial weighted pointwise convolution, followed by a depthwise convolution. In both stages, traditional convolutional layers are substituted with dilated convolutions. By expanding the receptive field, dilated convolution enables the capture of subtle facial variations—such as minor movements around the eyes or mouth—using fewer layers. The introduction of dilated convolution also enhances the network's capacity to extract multi-scale facial features, thereby improving both the accuracy and robustness of expression recognition.

During the weighted pointwise convolution stage, the input feature map is processed using a 1×1 convolution kernel in a pointwise manner. In conventional CNNs, this operation is typically performed using standard convolution. However, in the blueprint-separable dilated convolution, this step is replaced with a dilated convolution, thereby expanding the effective receptive field. Because subtle facial expression variations often span broad spatial regions, the use of dilated convolution enables the extraction of significant facial expression features over larger areas—without increasing computational complexity. As a result, the weighted features derived through pointwise convolution more accurately reflect student facial feedback during online English instruction, including emotional states such as anxiety, confusion, or concentration, thereby enhancing affective recognition performance.

In the depthwise convolution stage, the input feature maps undergo deep convolutional processing. Dilated convolution is again employed to enlarge the receptive field, further facilitating the extraction of high-level facial expression features along the depth dimension. Depthwise convolution operates independently on each channel, allowing fine-grained facial details to be captured for each specific channel. This is especially valuable in recognizing subtle facial expression changes, where the model is required to focus on highly localized expression regions. In online English instruction, teachers often rely on the real-time interpretation of students' micro-expressions to adjust teaching content and strategies. Accordingly, the model must be capable of extracting key expression-related information with both speed and precision. The integration of dilated convolution into the depthwise convolution stage enables the extraction of meaningful deep features while maintaining a low computational cost, thereby improving the accuracy of FER. Figure 4 shows the detailed architecture of the blueprint-separable dilated convolution network.

By integrating the two previously described stages, the complete blueprint-separable dilated convolution network achieves efficient extraction of facial expression features through a combination of weighted pointwise convolution and depthwise convolution. The use of dilated convolution increases the receptive field, enabling the network to capture a wider range of facial expression features with fewer layers and reduced computational overhead. In the context of online English instruction, this highly efficient FER capability is particularly valuable, as it enables teachers to detect students' emotional changes in real time. Such changes may reflect comprehension of the course content, signs of confusion, or varying levels of attention and engagement.

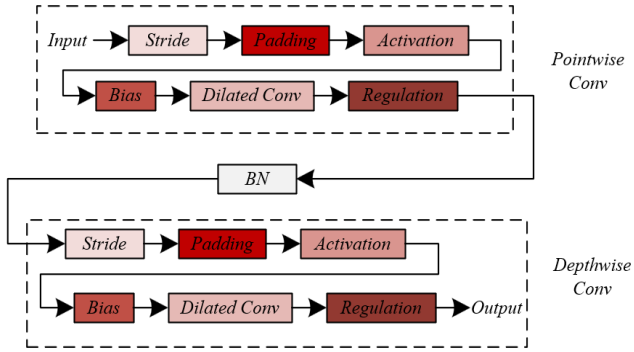


Figure 4. Detailed architecture of the blueprint-separable dilated convolution network

3. IMPROVING ONLINE ENGLISH INTERACTION WITH FER

Through the previously described methods, accurate recognition of student facial expressions was achieved using a blueprint-separable dilated convolution network module designed within a lightweight neural network framework. To further improve the quality of teacher–student interaction in online English instruction, several targeted strategies may be implemented as follows:

Strategy 1: The introduction of a real-time affective feedback mechanism constitutes a critical approach to enhancing interaction quality. By leveraging FER technology, the system is capable of continuously monitoring students' emotional states—such as confusion, anxiety, enjoyment, or focus—in real time. Based on detected emotional fluctuations, teachers can adjust their pedagogical strategies accordingly. For instance, when expressions of confusion or anxiety are detected, the system may prompt the teacher to consider modifying the pace or complexity of the instructional content. Additional clarification or increased interactivity may then be provided as needed. This real-time emotional feedback enables more precise adjustment of instructional pacing,

ensuring that students' affective needs are promptly addressed, thereby improving the overall quality of interaction.

Strategy 2: The optimization of personalized learning strategies also serves as an effective approach to enhancing interaction quality. FER not only provides real-time insight into students' emotional states but also enables the longitudinal analysis of emotional trends. Based on students' emotional patterns observed across different lessons or time intervals, tailored instructional strategies can be developed. For students exhibiting significant emotional fluctuations, individual support, adjustments to course content, or modifications to instructional methods may be employed to alleviate emotional burdens. Conversely, for students demonstrating emotional stability and sustained focus, the introduction of more challenging content may be used to stimulate interest and motivation. By implementing such personalized strategies, teachers can better adapt to the needs of each student, thereby enhancing the effectiveness and engagement of interaction.

Strategy 3: The enhancement of interactive teaching activity design represents another critical strategy for improving the quality of teacher–student interaction. In online English instruction, student facial expressions often convey authentic responses to instructional activities beyond verbal interaction. For example, expressions of enjoyment may indicate strong interest in a particular activity, prompting increased instructional emphasis on that component. Alternatively, facial expressions reflecting boredom or fatigue may suggest the need for additional interactive elements—such as group discussions or Question & Answer (Q&A) sessions—to boost engagement. By integrating feedback derived from FER, instructional content and interaction formats can be dynamically adjusted in real time. This approach fosters a more interactive and engaging classroom environment and further facilitates emotional communication and collaboration between teachers and students.

Strategy 4: The further development and integration of affective intelligence systems constitute a long-term measure for enhancing the quality of teacher–student interaction. Building upon the foundation of FER, the construction of more intelligent sentiment analysis systems can provide teachers with multidimensional emotional data. These data may include not only students' immediate affective responses but also their emotional preferences regarding various types of instructional content and their adaptability to different pacing patterns. Through such systems, a more comprehensive understanding of students' emotional states and learning needs can be obtained, enabling data-driven optimization of instructional design and interaction modalities. Moreover, the deep integration of affective intelligence systems with online English instruction platforms can facilitate real-time responsiveness to students' emotional changes. Instructional content and interaction formats can be automatically adjusted in accordance with these changes, thereby improving both teaching efficiency and interaction quality.

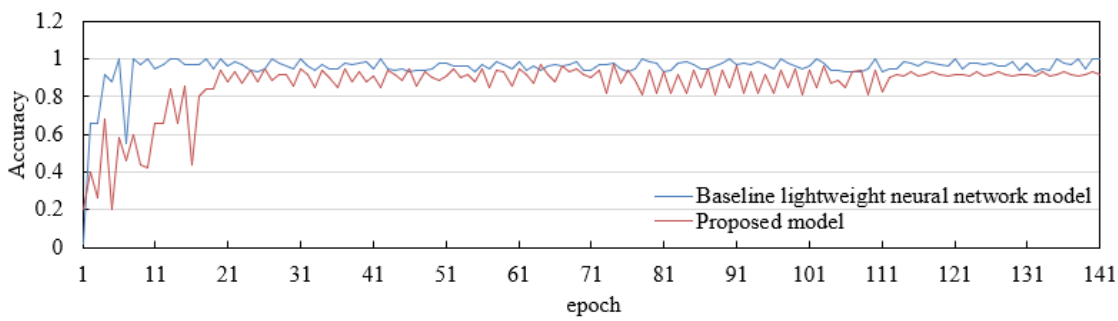
By incorporating real-time emotional feedback, personalized learning strategies, interactive instructional design, and integrated affective intelligence systems, teachers can more effectively understand and respond to students' emotional needs. Consequently, the quality of interaction in online English instruction can be significantly enhanced. As an effective tool for affective recognition, FER technology serves as a critical bridge in this process, facilitating the transition of online education from traditional knowledge transmission to a more personalized and interactive instructional paradigm.

4. EXPERIMENTAL RESULTS AND ANALYSIS

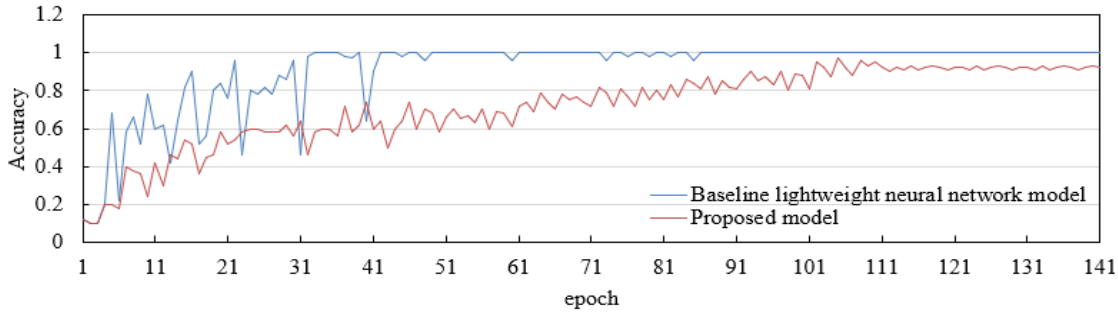
As shown in Table 1, the proposed model achieved significantly higher accuracy across all three datasets compared to the baseline lightweight neural network model. Specifically, for the student facial image dataset, the proposed model attained an accuracy of 98.61%, representing an improvement of approximately 5 percentage points over the baseline model's 93.56%. For the teacher facial image dataset, the proposed model achieved an accuracy of 98.51%, exceeding the baseline model's 95.47%. On the scene-associated image dataset, the proposed model reached an accuracy of 98.23%, while the baseline model recorded only 88.92%. These results demonstrate the superior performance of the proposed FER approach based on a multi-scale fused feature network. Enhanced recognition accuracy and stronger generalization capabilities were observed across diverse data sources, thereby validating the effectiveness of the proposed model in improving the precision of FER.

Table 1. Accuracy comparison between the baseline lightweight neural network model and the proposed model

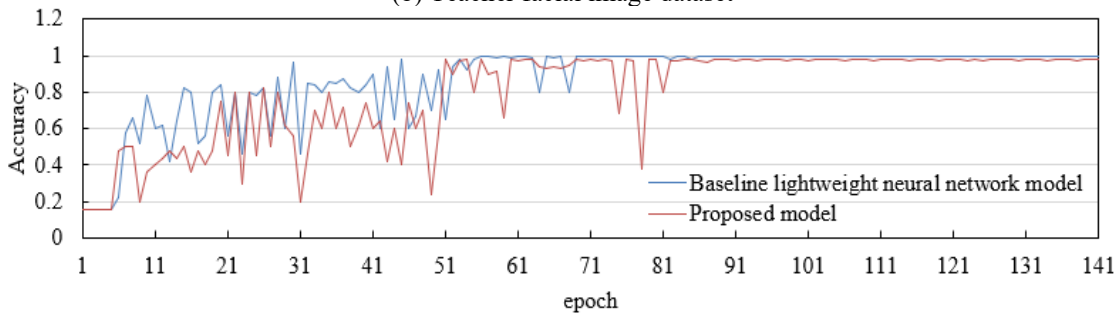
Model	Student Facial Image Dataset Accuracy (%)	Teacher Facial Image Dataset Accuracy (%)	Scene-Associated Image Dataset Accuracy (%)
Baseline lightweight neural network model	93.56	95.47	88.92
Proposed model	98.61	98.51	98.23



(a) Student facial image dataset



(b) Teacher facial image dataset



(c) Scene-associated image dataset

Figure 5. Accuracy curves of the baseline lightweight neural network model and the proposed model across three datasets

Table 2. Recognition accuracy comparison of multiple methods across three datasets

Dataset	Method	Recognition Accuracy (%)
Student facial image dataset	*StarGAN*	94.23
	MFNet	94.58
	Capsule-Emotion	96.32
	MTL-FER	97.85
	Proposed model	98.21
Teacher facial image dataset	*StarGAN*	92.65
	MFNet	95.36
	Capsule-Emotion	96.87
	MTL-FER	97.52
	Proposed model	98.26
Scene-associated image dataset	*StarGAN*	88.31
	MFNet	87.26
	Capsule-Emotion	87.56
	MTL-FER	98.31
	Proposed model	98.46

As illustrated in Figure 5, both the proposed model and the baseline lightweight neural network model exhibited a degree of fluctuation in the early training stages on the student facial image dataset, with both models eventually converging toward an accuracy of 1.00. However, the proposed model demonstrated greater overall stability. On the teacher facial image dataset, the baseline model experienced significant early fluctuations before stabilizing at a relatively high accuracy, while the proposed model showed a steady upward trend and similarly converged toward 1.00. In the scene-associated image dataset, the baseline model displayed notable instability, whereas the proposed model, despite some initial variation, also ultimately achieved near-perfect accuracy. These results visually depict the accuracy trends across training epochs for each model on the three types of datasets. From a broader perspective, the experimental outcomes across all three datasets indicate that the proposed model exhibited strong adaptability and effectiveness in diverse online English instruction facial image scenarios. Although some fluctuation was observed during early training phases, the final accuracy levels achieved by the proposed model were consistently high. Notably, the superior performance observed on both the student facial image dataset and the scene-associated dataset underscores the efficacy of the multi-scale fused feature network in enhancing FER accuracy. These findings provide a robust empirical foundation for subsequent strategies aimed at optimizing teacher–student interaction based on FER outcomes. Furthermore, they validate the practical value of the proposed approach to improving both the accuracy and real-time responsiveness of FER, while also reinforcing the reliability of leveraging recognition results to inform strategies for interaction quality enhancement.

Based on the data presented in Table 2, the proposed model achieved a markedly higher recognition rate than all comparative methods across the three evaluated datasets. On the student facial image dataset, the proposed model reached a recognition rate of 98.21%, outperforming Star Generative Adversarial Network (StarGAN) (94.23%), Multimodal Fusion Network (MFNet) (94.58%), Capsule-Emotion (96.32%), and Multi-Task Facial Expression Recognition (MTL-FER) (97.85%). On the teacher facial image dataset, an accuracy of 98.26% was attained, again surpassing StarGAN (92.65%), MFNet (95.36%), Capsule-Emotion (96.87%), and MTL-FER (97.52%). Notably, on the scene-associated image

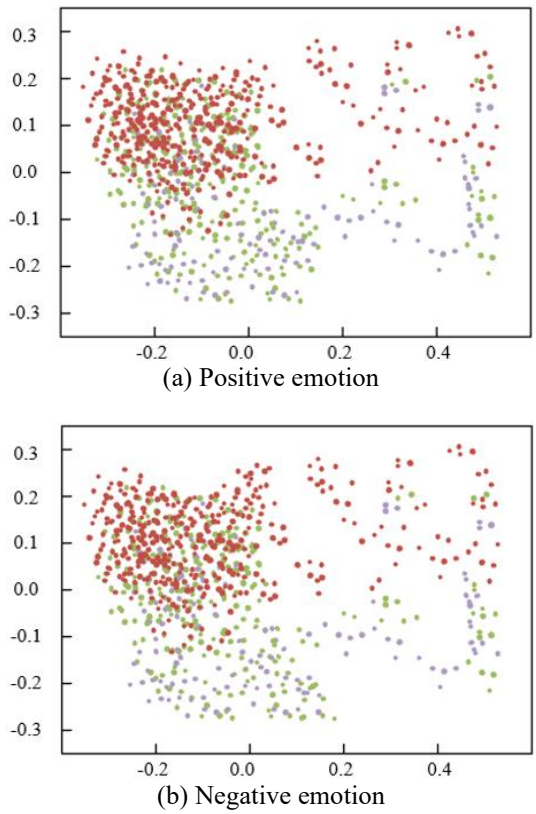
dataset, the proposed model achieved a recognition rate of 98.46%, exceeding MTL-FER (98.31%) as well as other baseline models, including StarGAN (88.31%), MFNet (87.26%), and Capsule-Emotion (87.56%). These experimental results confirm the proposed model's exceptional recognition accuracy across varied datasets. In particular, its performance remained robust even when applied to complex visual scenes involving diverse emotional expressions and background variability.

Table 3. Comparison of recognition accuracy and average mutual information for positive and negative features

Feature	Block Size	Accuracy (%)	Average Mutual Information
Positive emotion	1×1	54.23	1.115
	2×2	75.62	0.736
	5×5	91.23	0.415
	7×7	92.36	0.326
Negative emotion	1×1	61.25	1.526
	2×2	81.54	1.389
	5×5	88.96	0.856
	7×7	92.36	0.612

Based on the data presented in Table 3, significant differences were observed in recognition accuracy and average mutual information for positive and negative emotions across varying block sizes. For positive emotion recognition, accuracy was found to increase with larger block sizes. Specifically, an accuracy of 54.23% was achieved using a 1×1 block, which improved to 91.23% with a 5×5 block and reached 92.36% with a 7×7 block. However, this improvement in accuracy was accompanied by a progressive decrease in average mutual information, which declined from 1.115 at 1×1 to 0.326 at 7×7. This suggests that while larger block sizes enhanced recognition performance, a trade-off may have occurred in terms of information extraction efficiency. A similar trend was observed for negative emotion recognition. The accuracy increased from 61.25% with a 1×1 block to 81.54% with 2×2, 88.96% with 5×5, and stabilized at 92.36% with 7×7. Concurrently, accuracy for the negative emotion was stable with larger block sizes, and average mutual information decreased from 1.526 at 1×1 to 0.612 at 7×7. The results demonstrate that the proposed FER method maintained high recognition accuracy for both positive and negative emotions, particularly when larger block sizes were employed. These larger blocks proved effective in capturing subtle emotional variations within facial expressions, thus enhancing classification accuracy. However, this improvement came at the cost of reduced average mutual information, potentially indicating a compromise in fine-grained feature representation. In practical applications, a balance must be achieved between recognition accuracy and information extraction efficiency. For FER systems designed to support interaction in online English instruction, such a balance is essential to ensure that emotional data are utilized effectively, thereby optimizing teacher–student feedback mechanisms and instructional strategies.

Figure 6 presents the scatter plots of multi-scale fused facial expression features for positive and negative emotional states. In the plot corresponding to positive emotion, the distribution of points—distinguished by color—exhibits both aggregation and dispersion, indicating a rich and diverse representation of features under positive emotional conditions. In contrast, the scatter plot for negative emotion displays a more complex distribution pattern, with notable differences in density and spatial regions compared to positive emotion, thereby reflecting distinct characteristics in multi-scale feature expression between emotional categories. A degree of separability between the feature distributions of positive and negative emotions can be observed, suggesting that the multi-scale fused feature network was capable of effectively capturing emotion-specific patterns. Although partial overlap between feature points exists, the overall spatial structure reveals that emotion-discriminative multi-scale features were successfully extracted, supporting the FER process. This visual evidence aligns with the stated objective of enhancing recognition accuracy and substantiates the model's effectiveness in distinguishing between teacher and student emotional states. Moreover, it provides a reliable foundation for the subsequent optimization of interaction processes based on recognized emotional cues. These findings underscore the practical value of the proposed method in advancing the quality of teacher–student interaction within the context of online English instruction.



(a) Positive emotion

(b) Negative emotion

Figure 6. Scatter plot of multi-scale fused facial expression features

5. CONCLUSION

This study centers on the development of an FER method based on a multi-scale fused feature network, with the objective of enhancing both accuracy and real-time performance to improve the quality of teacher–student interaction in online English instruction. Through the introduction of the multi-scale fused feature network, a novel approach to FER was proposed. Experimental results obtained across multiple datasets demonstrated that the proposed model significantly outperformed existing methods in terms of recognition accuracy. Superior performance was achieved on

student, teacher, and scene-associated facial image datasets, with particularly notable adaptability and robustness observed in scenarios involving complex backgrounds. Furthermore, based on the outcomes of FER, strategies for optimizing the interactive processes between teachers and students to improve teaching effectiveness were explored, facilitating more personalized and responsive instructional delivery in online learning environments.

The contributions of this study are reflected in two primary dimensions. First, from a technological perspective, the proposed multi-scale fused feature network substantially improved the accuracy and real-time capabilities of FER, offering an efficient technical solution for online English instruction and related domains. Second, from an application standpoint, the integration of facial expression analysis into pedagogical practice enabled the refinement of teacher–student interaction, promoting personalization and real-time feedback, and thereby enhancing the overall learning experience. Nonetheless, several limitations remain. The application scope of the proposed model was primarily confined to FER. Although its relevance to instructional contexts was validated, its adaptability to broader affective recognition tasks warrants further investigation. In addition, the sensitivity of FER to environmental factors such as lighting conditions, occlusion, and facial variation may affect stability and robustness in real-world instructional settings. Moreover, individual differences in emotional expression among students were not fully accounted for, which may introduce bias or errors in emotion recognition under certain conditions.

Future research could be advanced in several key directions. First, efforts could be directed toward enhancing the robustness of the model, particularly in terms of maintaining stability under complex conditions such as variable lighting and facial occlusion or deformation. Second, the scope of application could be expanded through the integration of multimodal information—such as voice and contextual data—with facial expression features. Such multimodal fusion is expected to further improve the accuracy of emotion classification and enhance the granularity of emotional interpretation. Third, the model could be aligned with more refined personalized instructional strategies. By leveraging diverse emotion recognition outputs, new pathways may be explored to optimize interactive patterns between teachers and students, thereby enabling more intelligent online educational solutions. In conclusion, this study provides strong technical support and innovative insights for the field of online education. Nonetheless, further development and refinement are required to ensure that the proposed approach meets the evolving demands of increasingly complex and diverse educational environments.

REFERENCES

[1] Khalili, H. (2020). Online interprofessional education during and post the COVID-19 pandemic: A commentary. Journal of Interprofessional Care, 34(5): 687-690. https://doi.org/10.1080/13561820.2020.1792424

[2] Ortagus, J.C., Derreth, R.T. (2020). "Like having a tiger by the tail": A qualitative analysis of the provision of online education in higher education. Teachers College Record, 122(2): 1-32. https://doi.org/10.1177/016146812012200207

[3] Mao, S., Guo, L., Li, P., Shen, K., Jiang, M., Liu, Y. (2023). New era of medical education: Asynchronous and synchronous online teaching during and after COVID-19. Advances in physiology education, 47(2): 272-281. https://doi.org/10.1152/advan.00144.2021

[4] Panaligan, J.H., Curran, N.M. (2022). "We are cheaper, so they hire us": Discounted nativeness in online English teaching. Journal of sociolinguistics, 26(2): 246-264. https://doi.org/10.1111/josl.12543

[5] Madanat, H., Ab Rashid, R., Hashmi, U.M., Alqaryouti, M.H., Mohamed, M., Al Smadi, O.A. (2024). Jordanian English language educators' perceived readiness for virtual learning environment. Heliyon, 10(4): e25766. https://doi.org/10.1016/j.heliyon.2024.e25766

[6] Ma, Y.Y., Lin, C.L., Lin, H.L. (2023). Ranking of service quality index and solutions for online English teaching in the post-COVID-19 crisis. Mathematics, 11(18): 4001. https://doi.org/10.3390/math11184001

[7] Gherghel, C., Yasuda, S., Kita, Y. (2023). Interaction during online classes fosters engagement with learning and self-directed study both in the first and second years of the COVID-19 pandemic. Computers & Education, 200: 104795. https://doi.org/10.1016/j.compedu.2023.104795

[8] Attardi, S.M., Barbeau, M.L., Rogers, K.A. (2018). Improving online interactions: Lessons from an online anatomy course with a laboratory for undergraduate students. Anatomical Sciences Education, 11(6): 592-604. https://doi.org/10.1002/ase.1776

[9] Olderbak, S., Hildebrandt, A., Pinkpank, T., Sommer, W., Wilhelm, O. (2014). Psychometric challenges and proposed solutions when scoring facial emotion expression codes. Behavior Research Methods, 46: 992-1006. https://doi.org/10.3758/s13428-013-0421-3

[10] Wang, S., Chen, M., Ratnavelu, K., Shibghatullah, A. S. B., Keoy, K.H. (2024). Online classroom student engagement analysis based on facial expression recognition using enhanced YOLOv5 for mitigating cyberbullying. Measurement Science and Technology, 36(1): 015419. https://doi.org/10.1088/1361-6501/ad8a80

[11] Bhatti, Y.K., Jamil, A., Nida, N., Yousaf, M.H., Viriri, S., Velastin, S.A. (2021). Facial expression recognition of instructor using deep features and extreme learning machine. Computational Intelligence and Neuroscience, 2021(1): 5570870. https://doi.org/10.1155/2021/5570870

[12] Gu, M., Feng, J., Chu, Y. (2024). A novel multi-scale facial expression recognition algorithm based on improved Res2Net for classroom scenes. Multimedia Tools and Applications, 83(6): 16525-16542. https://doi.org/10.1007/s11042-023-16115-0

[13] Liu, T., Wang, J., Yang, B., Wang, X. (2021). Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom. Infrared Physics & Technology, 112: 103594. https://doi.org/10.1016/j.infrared.2020.103594

[14] Kayser, D., Egermann, H., Barraclough, N.E. (2022). Audience facial expressions detected by automated face analysis software reflect emotions in music. Behavior Research Methods, 54(3): 1493-1507. https://doi.org/10.3758/s13428-021-01678-3

[15] Black, D. (2023). Facial analysis: automated surveillance

and the attempt to quantify emotion. Information, Communication & Society, 26(7): 1438-1451. https://doi.org/10.1080/1369118X.2021.2011948

[16] Khan, R.A., Meyer, A., Konik, H., Bouakaz, S. (2013). Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recognition Letters, 34(10): 1159-1168. https://doi.org/10.1016/j.patrec.2013.03.022

[17] Kim, T.H., Yu, C., Lee, S.W. (2018). Facial expression recognition using feature additive pooling and progressive fine-tuning of CNN. Electronics Letters, 54(23): 1326-1328. https://doi.org/10.1049/el.2018.6932

[18] Doyle, N.B., Downer, J.T., Brown, J.L., Lowenstein, A.E. (2022). Understanding high quality teacher-student interactions in high needs elementary schools: An exploration of teacher, student, and relational contributors. School Mental Health, 14(4): 997-1010. https://doi.org/10.1007/s12310-022-09519-0

[19] Koenen, A.K., Vervoort, E., Verschueren, K., Spilt, J.L. (2019). Teacher–student relationships in special education: The value of the teacher relationship interview. Journal of Psychoeducational Assessment, 37(7): 874-886. https://doi.org/10.1177/0734282918803033