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

Traitement du Signal
Vol. 41, No. 6, December, 2024, pp. 3285-3293
Journal homepage: http://iieta.org/journals/ts

A Multimodal Image Recognition System for Student Behavior Analysis in Smart Classrooms in Universities

Yuan Zhou*, Juan Wang, Jianhua Zhang

College of Science, Chang'an University, Xi'an 710061, China

Corresponding Author Email: cz1979@chd.edu.cn

https://doi.org/10.18280/ts.410644

Received: 19 May 2024
Revised: 30 October 2024
Accepted: 17 November 2024
Available online: 31 December 2024

Keywords:
smart classroom, behavior analysis, multimodal image recognition, feature construction, feature fusion algorithm

ABSTRACT

With the rapid development of information technology, smart classrooms have become an integral component of higher education reform. Real-time and accurate analysis of student behavior in the classroom has emerged as a key issue for improving teaching quality and promoting personalized learning. Existing classroom behavior analysis methods largely rely on manual observation or single technology approaches, which often suffer from low efficiency and poor accuracy. In recent years, the application of multimodal image recognition technology in intelligent education has received widespread attention. By integrating multiple sensory inputs, such as visual images, student movements, and facial expressions, the comprehensive analysis of student behavior has been shown to significantly improve both the accuracy and timeliness of behavior analysis. However, the effective extraction and fusion of these multimodal features remain a challenge in current research. A behavior analysis system based on multimodal image recognition was proposed in this study for smart classrooms in universities. The main aspects of the research include (a) the construction of multimodal image features for classroom behavior analysis, integrating various information sources such as visual and movement data to accurately capture student behavior, and (b) the design of a multimodal image feature fusion algorithm to optimize data fusion strategies and enhance the system's adaptability and recognition accuracy in complex classroom environments. Experimental results demonstrate that this system achieves efficient and accurate behavior analysis in multiple real classroom settings, providing strong data support for the optimization of smart classrooms and personalized teaching.

1. INTRODUCTION

With the rapid development of information technology, smart classrooms have emerged as one of the key directions for higher education reform [1-3]. In traditional teaching models, interaction between teachers and students primarily occurs through face-to-face methods. In contrast, smart classrooms integrate teaching, learning, and management through digital technologies, efficiently merging these elements [4, 5]. Particularly in the context of higher education, smart classrooms not only enhance the flexibility and interactivity of teaching but also facilitate the realization of personalized learning. However, during this transformation process, a critical challenge is how to accurately and in real time analyze student behavior in the classroom and use this information to optimize the teaching process. Traditional classroom behavior analysis often relies on manual observation and recording, which are inefficient and prone to subjective biases, making it difficult to meet the demands of modern teaching [6-8]. Therefore, intelligent behavior analysis systems based on multimodal image recognition technology have garnered increasing attention, providing new solutions for the management and optimization of smart classrooms.

The development of an effective multimodal image recognition system for behavior analysis in smart classrooms has significant research value [9-13]. On one hand, image recognition technology can automatically monitor student behavior in the classroom, providing teachers with real-time feedback. This helps them understand students' learning states and emotional changes, enabling adjustments to teaching strategies. On the other hand, the integration of multimodal information (such as visual images, motion capture data, facial expressions, etc.) for comprehensive analysis not only improves the accuracy of behavior recognition but also offers data support for personalized teaching [14-17]. Through such a system, the effectiveness of teaching in smart classrooms can be scientifically quantified, thereby enhancing teaching quality. At the same time, it provides educational administrators with a basis for decision-making.

Despite several attempts to apply computer vision technology for classroom behavior analysis, significant limitations remain in current methods [18-21]. Most existing behavior analysis systems rely solely on single-modal image information, lacking effective fusion of multimodal data, which restricts their performance in complex classroom environments. Additionally, many current algorithms fail to meet the practical requirements for recognition accuracy and real-time performance when dealing with large numbers of students and complex backgrounds. The extraction and fusion

of multimodal image features continue to be challenging, with the key issue being how to enhance system efficiency while maintaining recognition accuracy. Thus, designing a multimodal image recognition system that balances both efficiency and accuracy remains a critical problem in this field.

This study focuses on two aspects: First, it investigates a multimodal image feature construction method for behavior analysis in smart classrooms by integrating sensory inputs like visual and motion data. Second, it proposes a multimodal image feature fusion algorithm, which enhances recognition accuracy and real-time performance of the behavior analysis system by optimizing data fusion strategies. These research approaches provide a new technological pathway for smart classroom behavior analysis, supporting data-driven decision-making in intelligent education with significant theoretical value and application potential.

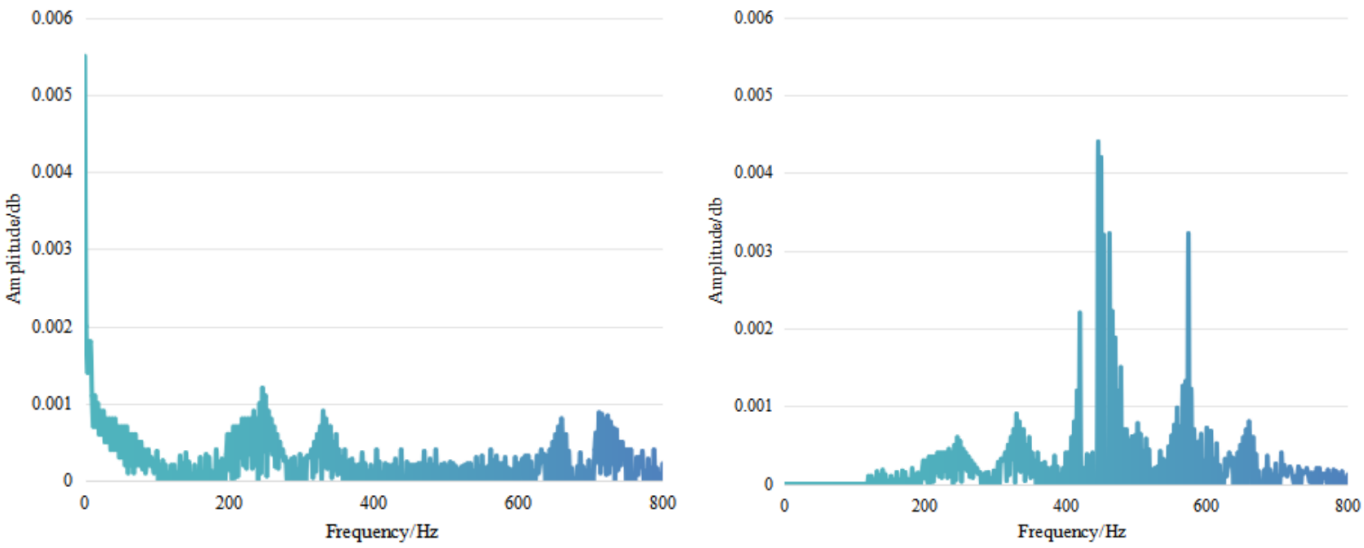
2. MULTIMODAL IMAGE FEATURE CONSTRUCTION

2.1 Construction of speech features

In the research on behavior analysis for smart classrooms in universities, the multimodal speech images employed in this study are derived from two types of speech images: phase-space-based speech images and time-frequency domain-based speech images. Phase-space-based speech images were generated through nonlinear dynamical methods by mapping classroom speech signals into phase space, forming images that encapsulate long-term evolutionary behavioral information. These images reflect the complex dynamic changes of speech signals in the time series, revealing long-term interaction patterns and behavioral trends of students in the classroom. On the other hand, time-frequency domain-based speech images were analyzed using techniques such as Short-Time Fourier Transform (STFT), which combine time, frequency, and energy information to generate images that contain rich speech features. This type of speech image can capture frequency characteristics and energy variations in the students' speech during class, helping to analyze their engagement, emotional responses, and interaction levels.

For multimodal speech images, short-time speech Fourier features were first constructed. The speech signal was preprocessed to obtain the speech sequence B and its sampling frequency D_t. To effectively capture the time-frequency characteristics of the speech signal, the STFT method was employed, where the speech signal is divided into several short time frames. Each speech frame was windowed and then Fourier-transformed to obtain the spectral information for that frame. Through this process, the energy distribution for each frame across time and frequency was derived, resulting in a series of spectrograms. By applying a sliding window technique, the energy density function $|A|^2$ for each frame was calculated, forming a continuous energy distribution matrix C across different time periods and frequencies. These time-frequency domain features were then processed into spectrograms to represent the variation of the speech spectrum over time. In the generated spectrogram, the horizontal axis represents time, the vertical axis represents frequency, and the intensity of the color indicates the energy strength of the corresponding frequency component. The resulting spectrogram can reveal the frequency characteristics and energy changes of students' speech in the classroom, especially during frequent interactions when the speech spectrum undergoes significant dynamic changes over time.

Further, the construction of speech phase-space features for multimodal speech images was conducted in this study. Traditional speech analysis methods typically treat speech as a stationary linear signal over short time intervals. However, this approach overlooks the nonlinear dynamic characteristics of speech signals during long-term evolution. In fact, speech production is a complex nonlinear process involving the nonlinear vibration of the vocal cords and the dynamic changes in characteristics such as frequency and energy during propagation. In a classroom setting, students' speech behaviors not only reflect immediate emotions, cognitive states, and other aspects, but also contain complex information that evolves over time, especially in group interactions and fluctuations in speech-related emotions. Therefore, investigating the nonlinear dynamical characteristics of speech was focused on in this study to reveal the complex temporal and spatial evolutionary patterns of speech signals, thereby enhancing the accuracy of behavior analysis for students in the classroom.



(a) Without phase-space reconstruction

(b) With phase-space reconstruction

Figure 1. Spectrograms of speech images before and after phase-space reconstruction

Phase-space reconstruction is a crucial step in the processing of nonlinear time series because speech data collected in the classroom is often noisy and consists of finite-length time sequences and speech signals themselves exhibit complex nonlinear and time-varying characteristics. According to Takens' embedding theorem, by selecting the appropriate embedding dimension and delay time, phase-space reconstruction can be performed effectively with the same topological structure as the original system, revealing the intrinsic dynamical laws of speech signals. Thus, when applying phase-space reconstruction to process the nonlinear features of speech signals, selecting the appropriate embedding dimension f and delay time s is essential. The embedding dimension f determines the dimensionality of the phase space. A dimension that is too small may fail to capture sufficient complex dynamic information, while a dimension that is too large could introduce redundant information, increasing computational complexity. The delay time s determines the time interval between each data point when constructing the phase space. An appropriate delay time ensures that the reconstructed phase space maximally reflects the evolutionary trends of the speech signal. Figure 1 illustrates the spectrograms of speech images before and after phase-space reconstruction. As seen in the figure, improper selection of delay time may lead to information loss or redundancy. The phase-space vector is as follows:

$$b(u)=a(u),...,a\left(u+(f-1)s\right),1\le u\le v-(f-1)s \tag{1}$$

As indicated by the above equation, the selection of delay time s has a significant impact on the result of phase-space reconstruction. If the delay time is too short, the coordinate components of adjacent time points may become overly similar, lacking independence and failing to provide effective dynamic information. Conversely, if the delay time is too long, the coordinate components may become completely independent, losing essential correlations. Therefore, choosing an appropriate delay time s is crucial. In this study, the autocorrelation coefficient method was employed to determine the delay time, as this method effectively reveals the intrinsic structure of the time series by analyzing the linear correlations between different time points. The autocorrelation function method has a relatively low computational load for low-dimensional time series, making it suitable for real-time processing of classroom speech data. This method can quickly capture the correlations and changing trends of the speech signal. By calculating the autocorrelation function, it is ensured that the delay time s reflects the dynamic evolution of the speech signal while maintaining appropriate independence and correlation. This improves the accuracy of the speech image features during phase-space reconstruction. For a nonlinear time series $a(1),a(2),...,a(v)$, the autocorrelation function is given by:

$$E(s)=\frac{1}{v}\sum_{u=1}^{v-s}a(u)a(u+s) \tag{2}$$

To reveal the temporal correlations between speech data points, it is necessary to further plot the autocorrelation function $E(s)$ as a function of delay time. That is, the focus is on the value of the autocorrelation function $E(s)$ at different delay time s. When the autocorrelation function $E(s)$ reaches a specific value, namely $(1\text{-}e^{-1})$ multiplied by $E(0)$, the corresponding delay time s is considered to be the ideal delay time for phase-space reconstruction. Interviews, as a form of free-flowing verbal exchange, possess a strong level of interactivity and flexibility, making them closer to the natural speech communication found in daily conversations. Therefore, interviews were chosen for delay time calculation, as they more accurately simulate the language behavior patterns of students in the classroom, particularly the nonlinear characteristics of their speech. Initially, the delay time for each recording was calculated using the autocorrelation coefficient method, allowing for the extraction of the intrinsic dynamic characteristics of the speech signal. This method aids in identifying the speaking rhythm, emotional fluctuations, and interaction patterns of students in the classroom. Furthermore, the delay time was divided into five intervals, with the range of 31 to 60 being selected as the primary analysis interval. This choice is based on the fact that this range represents a significant proportion of the calculated results. The underlying logic is that by focusing on the most common delay time range, the influence of noise on delay time selection can be effectively minimized, ensuring that the analysis results are representative. When a delay time of 45 was selected, it was found to balance the independence and correlation of the signal, providing the optimal delay time for phase-space reconstruction. The trajectory matrix of the reconstructed phase space is given by:

$$b=\begin{bmatrix} a_1, & a_{1+s} \\ a_2, & a_{2+s} \\ & \vdots \\ a_{v-s}, & a_v \end{bmatrix} \tag{3}$$

2.2 Construction of facial image features

In the construction of multimodal image features for behavior analysis in smart classrooms, two types of facial images were employed: time-encoded facial images and frequency-domain facial images. The time-encoded facial images were designed based on relevant theories of smart classroom behavior in higher education. These images aim to integrate both the dynamic and spatial information of the subjects, enabling real-time capture of facial expressions, eye movements, and other behavioral features of students. Such image data provides more accurate behavioral recognition, especially during classroom interactions, where it can effectively reflect emotional fluctuations, attention levels, and engagement, thereby offering more intuitive and comprehensive feedback to intelligent teaching systems. The frequency-domain facial images were primarily designed with privacy protection in mind. In this design, facial images were processed in the frequency domain to prevent the direct exposure of the subject's personal identity, while still allowing for the extraction of valuable facial features. This approach ensures the adherence to ethical and privacy requirements in the research.

For the multimodal facial images, the Fourier features from facial images were first constructed. Facial images inherently contain rich information about emotions, expressions, and attention, but this information typically requires image processing techniques for extraction. To facilitate better integration with speech signals and take advantage of the time-frequency analysis of speech, the image data were transformed into time-domain waveforms. This transformation allows the

image features to be mapped into the same analytical framework used for the speech signal, while also enabling the capture of periodic and nonlinear features in the images through frequency-domain processing. By applying the Fourier transform, the image was converted from the spatial domain to the frequency domain, and the inverse Fourier transform was used to reconstruct the time-domain waveform related to the image content. This processed image data retains the dynamic information of facial expressions and behavior, making it more suitable for subsequent speech recognition and multimodal data fusion analysis.

Specifically, the images were initially pre-processed in this study. Subsequently, a two-dimensional Fourier transform was applied to map the images from the time domain to the frequency domain, thereby extracting their spectral features. Further processing involved applying the inverse Fourier transform to each column of the image's spectrogram, converting it into a time-domain waveform. During this process, each column of data was treated as a distinct set of spectral information. The time-domain waveform obtained after the inverse Fourier transform represents the temporal features of the image data in that column. Ultimately, all the resulting time-domain waveform vectors were concatenated to form a complete time-domain waveform, which was then saved in a standard audio file format (e.g., WAV). Once transformed in this manner, the image data are more compatible for further analysis and fusion with speech signals. To extract the Fourier features of the image, the Fourier feature images corresponding to the time-domain waveform were calculated. These feature images can capture significant information about the image in the frequency domain, thereby enriching the multimodal data analysis. Let $d(a,b)$ represent a matrix of size $L\times V$, where a=0,1,2,…,L-1 and b=0,1,2,…,V-1. The Fourier transform of $d(a,b)$ is denoted as $D(i,n)$. The coordinate system of $D(i,n)$ is referred to as the frequency domain, and the frequency-domain matrix is a $L\times V$ matrix defined by i= 0,1,2,...,L-1 and n = 0,1,2,...,V-1. The coordinate system of $d(a,b)$ is referred to as the spatial domain, with the spatial-domain matrix being a $L\times V$ matrix defined by a= 0,1,2,...,L-1 and b=0,1,2,...,V-1. The two-dimensional discrete Fourier transform is expressed by the following formula:

$$D(i,n)=\sum_{a=0}^{L-1}\sum_{b=0}^{V-1}d(a,b)e^{-k2\tau\left(\frac{ia}{L}+\frac{nb}{V}\right)} \tag{4}$$

Further, the spatial-temporal features of facial images in multimodal speech-image analysis were constructed. In this approach, facial images were mapped to images with time information using time encoding, thereby retaining the temporal correlation between consecutive frames. Specifically, nine consecutive facial images were combined into a composite image, which preserves both the spatial features of facial expressions and the temporal changes between adjacent frames. The difference between each image and its adjacent images was computed, and the direction of change was mapped through threshold processing. These differences were then converted into binary encoding, forming the pixel values of the composite image. Specifically, the difference between each image and the eight surrounding images was calculated using the following formula:

$$a[l,v,s]=\sum_{k=1\&k\neq\frac{m+1}{2}}^{1}b[l,v,k]-b\left[l,v,\frac{m+1}{2}\right] \tag{5}$$

To map the direction of facial expression changes in images, a thresholding technique was employed to assign output values:

$$a[l,v,s]=\begin{cases}0, & if\ a[l,v,s]>0\\ 1, & otherwise\end{cases} \tag{6}$$

Finally, the binary encoding was converted according to the sequence of images, and the pixel values of the composite image were computed using the following formula:

$$c[l,v]=\sum_{k=1}^{g}a[l,v,k]*2^{k} \tag{7}$$

This approach not only leverages the strengths of 3D convolutional neural network (CNN) in action recognition but also effectively addresses the issue of neglecting temporal information in traditional image processing methods. It is capable of more accurately capturing the changes in facial expressions and dynamic behaviors of students in the classroom, thereby providing robust data support for behavior analysis in smart classrooms in universities.

3. MULTIMODAL IMAGE FEATURE FUSION ALGORITHM

In the analysis of behavior in smart classrooms in universities, the decision-level speech and image feature fusion algorithm utilized features from both speech and facial image modalities. These features were processed through independent classifiers, and their respective recognition results were fused to enhance the overall recognition accuracy. Figure 2 illustrates the flowchart of behavior recognition in smart classrooms based on decision-level feature fusion.

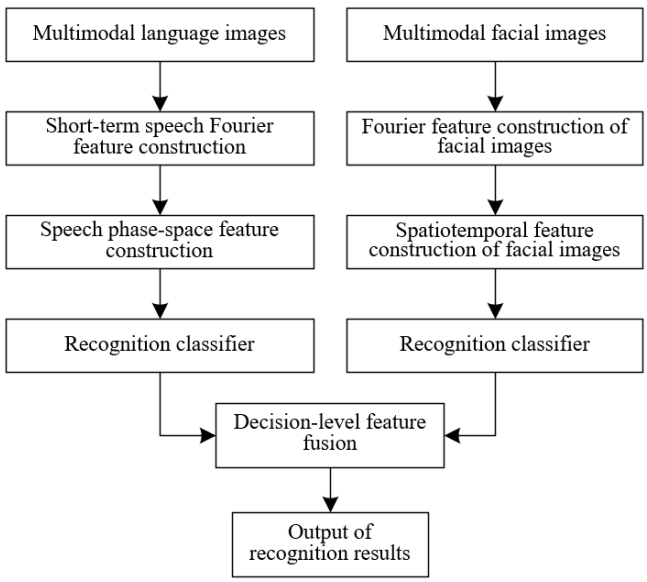


Figure 2. Flowchart for behavior recognition in smart classrooms in universities based on decision-level feature fusion

Initially, the speech and image features were processed separately by dedicated classifiers, which yielded their respective recognition results and corresponding probabilities. Each classifier output a set of recognition probabilities for the

respective modality. Subsequently, based on decision fusion rules, the recognition probability sets of speech and facial images were weighted or combined to obtain a final classification probability set. By comparing these aggregated probabilities, the label category with the highest probability value was selected as the final recognition result. The following formula defines the new classification probability $w_k(a)$ after processing according to the rules:

$$w_k(a) = w_k'(a) \Big/ \sum_{k}^{l} w_k'(a) \tag{8}$$

where, $w'_k(a)$ was calculated by the following formula:

$$w_k'(a) = RULE\left(o_{uk}(a)\right) \tag{9}$$

The following final classification label was obtained:

$$q(a) = ARGMAX\left(w_k(a)\right) \tag{10}$$

To improve the accuracy and robustness of classroom behavior recognition, three decision-making methods were employed: summation, product, and maximum-value decision rules. The summation decision rule combines the classification probabilities generated by the speech and facial image classifiers by adding them together. This approach allows for a weighted average of the classification information from different modalities, reducing the errors that may arise from any single modality. It is particularly suitable for scenarios where strong complementary relationships exist between the modalities. Specifically, $w'_k(a)$ can be represented as:

$$w_k'(a) = \sum_{u=1}^{v} o_{uk}(a) \tag{11}$$

The product decision rule multiplies the same-class classification probabilities from each classifier, emphasizing the synergistic effects of each modality's classification. This rule effectively amplifies consistency when the emotional recognition results of speech and facial images are highly consistent, making it suitable for scenarios where strong consistency between modalities exists. The probability $w'_k(a)$ can be represented as:

$$w_k'(a) = \prod_{u=1}^{v} o_{uk}(a) \tag{12}$$

The maximum-value decision rule selects the highest recognition probability from each modality's classifier for comparison, choosing the final classification. This approach ensures the conservativeness of classification decisions, particularly in cases where certain modalities may have recognition uncertainty or noise, thus avoiding the influence of errors from any single modality on the final result. The probability $w'_k(a)$ can be represented as:

$$w_k'(a) = MAX_u\left(o_{uk}(a)\right) \tag{13}$$

These three decision rules, by considering the weights of different emotional information, the synergistic effects, and the robustness of modality recognition, assist in capturing the multi-dimensional information of students' emotions, behaviors, and engagement in the classroom. This, in turn, contributes to more accurate classroom behavior analysis results.

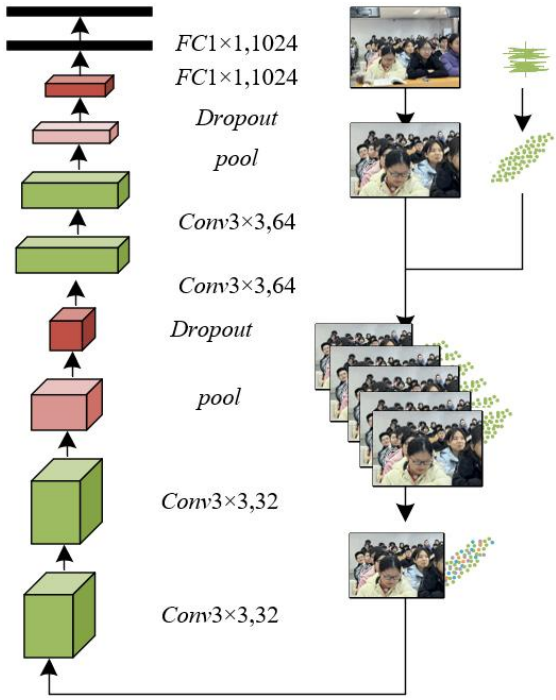


Figure 3. Structure diagram of the multimodal image feature fusion algorithm

In terms of feature fusion, a feature-level fusion method was adopted for the integration of features, wherein the features of both speech and facial images were directly combined into a unified feature vector for emotional recognition. The feature vectors obtained from different modalities were concatenated at the feature layer, forming a new composite feature vector. This process enables a unified representation of the information from different modalities. Through this feature-level fusion, the multidimensional behavioral characteristics of students, including emotions, engagement, and attention in the classroom, can be more effectively captured. The changes in speech and facial images complementarily provide comprehensive information about the students' emotional states. Finally, the fused feature vector was processed through a classification algorithm for emotional recognition, resulting in the final classification of emotional and behavioral states. The advantage of this method lies in its simplicity and efficiency. It directly integrates multimodal features, avoiding the need for multiple rounds of decision-making or complex computations, and significantly improves recognition accuracy. This is particularly beneficial in dynamic classroom environments, as it allows for the real-time reflection of students' emotional fluctuations and behavioral patterns. Figure 3 illustrates the structure of the multimodal image feature fusion algorithm.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Based on the experimental results in Table 1 and Figure 4, the influence of different decision rules on multimodal image behavior recognition performance, as well as the effectiveness of the feature-level fusion method, can be observed. The data

in the table shows the recognition accuracy for various behavioral categories such as "focused listening," "drowsiness/fatigue," and "active questioning" under different decision rules. Among all decision rules, the summation decision rule exhibits relatively balanced performance across most behavior categories, particularly in the "drowsiness/fatigue" and "nervous/anxious" categories, where recognition accuracy reaches 68.25% and 75.64%, respectively, slightly outperforming the other rules. However, considering the overall average performance, the product decision rule yields the best results, with an overall recognition accuracy of 78.13%. It shows particularly strong performance in the "focused listening" and "low mood/dissatisfaction" categories, achieving 87.52% and 79.26%, respectively. This suggests that by accumulating feature information from different modalities, the model is better able to recognize and classify behavioral features, particularly in complex behavioral patterns where the accumulated decision rule exhibits strong advantages.

Table 1. Performance comparison of multimodal image behavior recognition using feature-level fusion and different decision rules (%)

	Focused Listening	Drowsiness / Fatigue	Active Questioning	Low Mood / Dissatisfaction	Distracted / Daydreaming	Nervous / Anxious	Average
Feature-level fusion	83.25	66.58	72.15	78.52	83.21	73.26	76.16
Summation decision rule	83.25	68.25	72.62	78.62	82.15	75.64	76.75
Product decision rule	87.52	68.26	72.36	79.26	84.26	77.15	78.13
Maximum-value decision rule	82.36	66.36	69.26	81.25	83.26	73.26	75.95

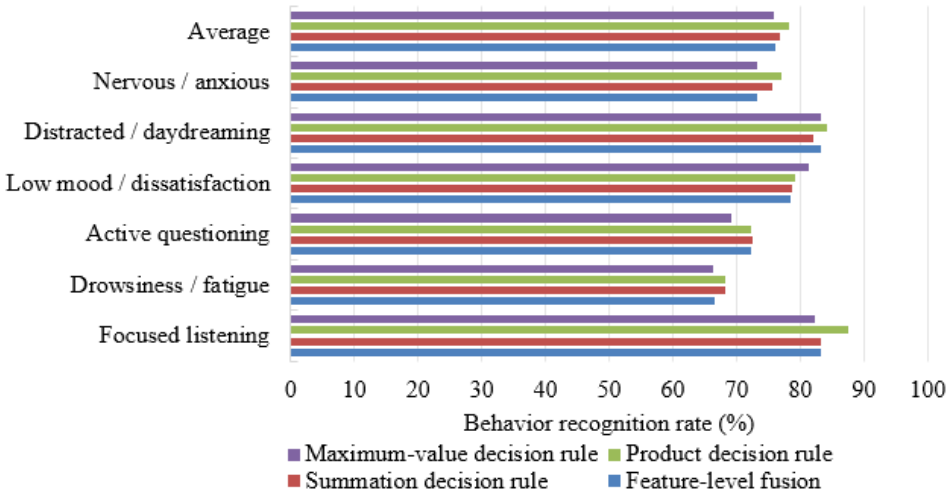


Figure 4. Histogram comparison of multimodal image behavior recognition using feature-level fusion and different decision rules (%)

Table 2. Comparison of unimodal and bimodal behavior recognition (%)

	Speech Features	Facial Features	Fusion Features
Focused listening	81.26	58.96	87.52
Drowsiness / fatigue	58.98	58.64	67.15
Active questioning	66.98	52.26	73.26
Low mood / dissatisfaction	63.15	76.62	78.26
Distracted / daydreaming	81.58	71.26	84.51
Nervous / anxious	68.99	62.35	76.26
Average	71.26	62.65	77.59

It can be inferred that feature-level fusion and optimized decision rules play a crucial role in system performance. Under the efficient feature-level fusion method, multiple sensory inputs are integrated, offering more accurate capture and recognition of student behavior. The multimodal image feature fusion algorithm proposed in this study enhances behavioral analysis accuracy and real-time performance by effectively integrating information from various modalities. As seen from the table, the product decision rule improves the system's robustness and accuracy through the accumulation of weighted information, and it performs especially well in complex classroom environments, where it adapts better to various behavioral patterns.

Based on the data from Table 2 and Figure 5, a significant improvement in behavior recognition accuracy is demonstrated through the comparison between unimodal and bimodal behavior recognition. By comparing the recognition performance of speech features, facial features, and their fusion features, it is evident that the fusion features outperform the individual modalities in most behavior categories.

Specifically, speech features exhibit higher recognition accuracy in the "focused listening" and "distracted/daydreaming" categories, with accuracies of 81.26% and 81.58%, respectively. On the other hand, facial features perform best in the "low mood/dissatisfaction" category, achieving an accuracy of 76.62%. However, when speech and facial features are fused, recognition accuracy improves across the board, with "focused listening" rising from 81.26% to 87.52%, and "drowsiness/fatigue" increasing from 58.98% to 67.15%. This suggests that the fusion of speech and facial features provides a more comprehensive reflection of students' behavioral states, thereby enhancing both the accuracy and robustness of recognition. The experimental results highlight that the advantage of bimodal fusion features lies in the complementary nature of speech and facial information, which enables more accurate recognition of students' behavior patterns. This is particularly evident in the recognition of behaviors such as "focused listening" and "low mood/dissatisfaction".

Based on the accuracy comparison results for speech recognition, facial recognition, and combined speech and facial behavior recognition tasks in Table 3 and Figure 6, the proposed method can be analyzed from multiple perspectives. The proposed method outperforms most of the comparison methods in all three tasks.

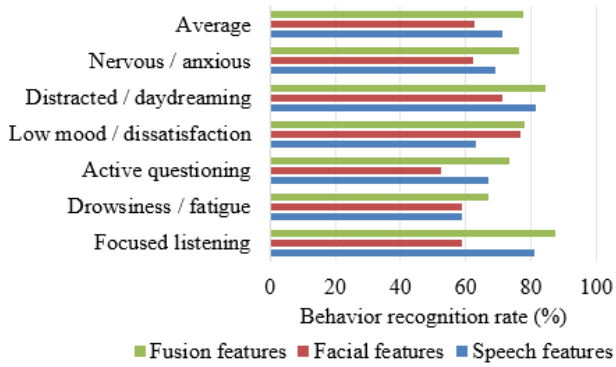


Figure 5. Histogram comparison of behavior recognition performance for six categories between unimodal and bimodal recognition

Table 3. Accuracy comparison of the proposed method with other methods

	Speech Recognition	Facial Recognition	Integrating Speech and Facial Behavior Recognition
Weighted voting method	34.15	24.56	66.28
Self-attention network	32.62	36.89	72.61
Generative adversarial network	61.25	53.21	72.58
Dual-stream network	54.25	38.95	55.69
Multimodal embedding network	39.58	57.26	75.62
Multimodal graph convolutional network	57.26	52.31	62.32
Cross-modality alignment network	73.26	41.23	77.89
Vision-language transformer	57.89	54.56	61.23
Proposed method	71.23	62.36	78.26

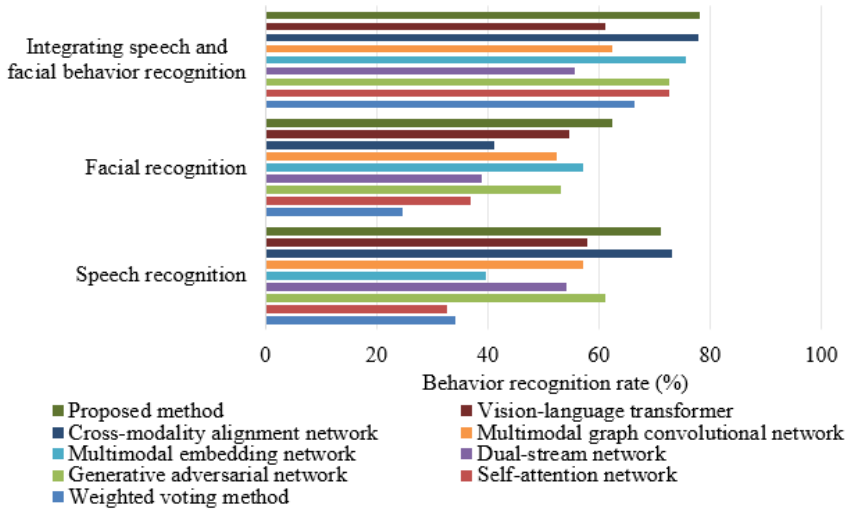


Figure 6. Histogram comparison of the behavior recognition accuracy between the proposed method and other methods

In the speech recognition task, the accuracy of the proposed method is 71.23%, significantly surpassing other methods, particularly the weighted voting method (34.15%) and the self-attention network (32.62%). This result indicates the significant advantage of the multimodal image feature fusion algorithm proposed in handling speech data, as it can more accurately capture speech features. For the facial recognition task, the accuracy of the proposed method is 62.36%, which is also higher than most methods, especially when compared to the dual-stream network (38.95%) and the generative adversarial network (53.21%). This demonstrates the superiority of the proposed method in facial recognition, likely attributed to its optimization in image feature fusion. Most notably, in the combined speech and facial behavior recognition task, the accuracy of the proposed method reaches 78.26%, surpassing all other methods, including the cross-modality alignment network (77.89%) and the multimodal embedding network (75.62%). This suggests the method's strong capability in handling multimodal data, such as the combination of speech and facial information. From the

analysis results, it is evident that by integrating multiple sensory inputs, the proposed method can precisely capture students' behavioral characteristics in the classroom, showing stronger adaptability and robustness, particularly in complex environments. The multimodal image feature fusion algorithm proposed enhances system performance through optimized data fusion strategies, enabling more efficient data processing when multiple modalities are combined, thus improving recognition accuracy. In practical applications, this algorithm can significantly enhance the performance of the behavior analysis system, particularly in the fusion of multimodal information, showing substantial potential for future applications.

5. CONCLUSION

The primary research focus of this study is on behavior analysis in smart classrooms in universities, with a particular emphasis on exploring how multimodal data fusion can enhance system accuracy and real-time performance. Specifically, a multimodal image feature construction method was proposed, aimed at accurately capturing students' behavioral characteristics in the classroom by integrating multiple sensory inputs, such as visual and motion data. This method effectively consolidates information from various perceptual channels, providing a more comprehensive and precise foundation for behavior analysis. Additionally, a multimodal image feature fusion algorithm was introduced, optimizing the data fusion strategy to significantly improve the adaptability and robustness of the behavior analysis system in complex classroom environments. This algorithm fully exploits the relationships between different modalities during the fusion process, thereby enhancing both the accuracy and real-time responsiveness of behavior recognition.

The results indicate that the proposed method demonstrates exceptional performance in speech recognition, facial recognition, and behavior recognition tasks involving the fusion of speech and facial information. Notably, in the behavior recognition task that integrates speech and facial data, an accuracy of 78.26% was achieved, which is significantly higher than that of other comparison methods. This suggests that the multimodal image feature fusion algorithm proposed can effectively enhance the overall system performance when handling complex multimodal data, particularly when speech and facial information are combined, resulting in significant improvements in behavior recognition accuracy and reliability. Furthermore, the method not only enhances the accuracy of behavior recognition but also strengthens the system's real-time capabilities, enabling it to quickly respond to changes in students' behavior in the classroom, thus holding substantial practical application value.

Future research could further expand and deepen these findings in several areas. First, the multimodal data fusion strategy could be further optimized by exploring the effective integration of additional modalities, such as biological signals and environmental sensor data, to further improve system accuracy and robustness. Second, by incorporating advanced technologies like deep learning, the computational efficiency of the algorithm could be optimized, reducing computational complexity, thereby improving the real-time capabilities and scalability of the system. Moreover, future research could extend the proposed method to a broader range of applications, such as smart homes and public safety monitoring, to validate its adaptability and performance in various environments. Lastly, by expanding the dataset to encompass more diverse classroom scenarios and student behaviors, the generalization ability of the algorithm could be enhanced, further advancing the development of the smart classroom behavior analysis system.

FUNDINGS

This paper was supported by "Double first-class" special project of Chang'an University, research on engineering graphics teaching system based on Internet (Grant No.: 300104281221) and Curriculum Construction of Chang'an Academy (Grant No.: 300207243520).

REFERENCES

[1] Wang, S.Y., Cheng, L.M., Liu, D.Y., Qin, J.Q., Hu, G.H. (2022). Classroom video image emotion analysis method for online teaching quality evaluation. Traitement du Signal, 39(5): 1767-1774. https://doi.org/10.18280/ts.390535

[2] Dai, Z., Sun, C., Zhao, L., Zhu, X. (2023). The effect of smart classrooms on project-based learning: A study based on video interaction analysis. Journal of Science Education and Technology, 32(6): 858-871. https://doi.org/10.1007/s10956-023-10056-x

[3] Jiang, L., Lu, X. (2023). Analyzing and optimizing Virtual Reality classroom scenarios: A deep learning approach. Traitement du Signal, 40(6): 2553-2563. https://doi.org/10.18280/ts.400618

[4] Selim, H.M., Eid, R., Agag, G. (2020). Understanding the role of technological factors and external pressures in smart classroom adoption. Education+ Training, 62(6): 631-644. https://doi.org/10.1108/ET-03-2020-0049

[5] Ma, X., Xie, Y., Yang, X., Wang, H., Lu, J. (2024). Structural model construction and analysis for teacher-student interaction in smart classroom based on the development of higher-order thinking. Education and Information Technologies, 29(16): 21691-21717. https://doi.org/10.1007/s10639-024-12733-9

[6] Lu, K., Yang, H.H., Shi, Y., Wang, X. (2021). Examining the key influencing factors on college students' higher-order thinking skills in the smart classroom environment. International Journal of Educational Technology in Higher Education, 18: 1. https://doi.org/10.1186/s41239-020-00238-7

[7] Aguilar, J., Altamiranda, J., Diaz, F. (2020). Specification of a managing agent of emergent serious games for a smart classroom. IEEE Latin America Transactions, 18(1): 51-58. https://doi.org/10.1109/TLA.2020.9049461

[8] Zhan, Z., Wu, Q., He, W., Cheng, S., Lu, J., Han, Y. (2021). K12 teacher-student interaction patterns in the smart classrooms. International Journal of Innovation and Learning, 29(3): 267-286. https://doi.org/10.1504/IJIL.2021.114511

[9] Fu, S. (2022). A reinforcement learning-based smart educational environment for higher education. International Journal of e-Collaboration, 19(6): 1-17. https://doi.org/10.4018/IJeC.315019

[10] Zhan, Z., Wu, Q., Lin, Z., Cai, J. (2021). Smart classroom

environments affect teacher-student interaction: Evidence from a behavioural sequence analysis. Australasian Journal of Educational Technology, 37(2): 96-109. https://doi.org/10.14742/ajet.6523

[11] Aguilar, J., Buendia, O., Pinto, A., Gutiérrez, J. (2022). Social learning analytics for determining learning styles in a smart classroom. Interactive Learning Environments, 30(2): 245-261. https://doi.org/10.1080/10494820.2019.1651745

[12] Liu, Q., Zheng, X., Liu, Y., Wu, L., Zhang, S., Zhang, N., Wang, Q. (2024). Exploration of the characteristics of teachers' multimodal behaviours in problem-oriented teaching activities with different response levels. British Journal of Educational Technology, 55(1): 181-207. https://doi.org/10.1111/bjet.13332

[13] Noda, K., Arie, H., Suga, Y., Ogata, T. (2014). Multimodal integration learning of robot behavior using deep neural networks. Robotics and Autonomous Systems, 62(6): 721-736. https://doi.org/10.1016/j.robot.2014.03.003

[14] Fu, X., Zuo, C., Yan, H. (2017). Multimodal quantitative phase and fluorescence imaging of cell apoptosis. In Fifth International Conference on Optical and Photonics Engineering, Singapore, pp. 1044920. https://doi.org/10.1117/12.2270794

[15] Borg, E. (2015). Classroom behaviour and academic achievement: How classroom behaviour categories relate to gender and academic performance. British Journal of Sociology of Education, 36(8): 1127-1148. https://doi.org/10.1080/01425692.2014.916601

[16] Zheng, Q., Chen, Z., Wang, M., Shi, Y., Chen, S., Liu, Z. (2024). Automated multi-mode teaching behavior analysis: A pipeline based event segmentation and description. IEEE Transactions on Learning Technologies, 17: 1677-1693. https://doi.org/10.1109/TLT.2024.3396159

[17] Pas, E.T., Cash, A.H., O'Brennan, L., Debnam, K.J., Bradshaw, C.P. (2015). Profiles of classroom behavior in high schools: Associations with teacher behavior management strategies and classroom composition. Journal of School Psychology, 53(2): 137-148. https://doi.org/10.1016/j.jsp.2014.12.005

[18] Lin, J., Li, J., Chen, J. (2022). An analysis of English classroom behavior by intelligent image recognition in IoT. International Journal of System Assurance Engineering and Management, 13(Suppl 3): 1063-1071. https://doi.org/10.1007/s13198-021-01327-0

[19] Greer, B.D., Neidert, P. L., Dozier, C.L., Payne, S.W., Zonneveld, K.L., Harper, A.M. (2013). Functional analysis and treatment of problem behavior in early education classrooms. Journal of Applied Behavior Analysis, 46(1): 289-295. https://doi.org/10.1002/jaba.10

[20] Akila, D., Garg, H., Pal, S., Jeyalaksshmi, S. (2024). Research on recognition of students attention in offline classroom-based on deep learning. Education and Information Technologies, 29(6): 6865-6893. https://doi.org/10.1007/s10639-023-12089-6

[21] Wilson, A.N., Dzugan, E., Hutchinson, V.D. (2022). Using a nonconcurrent multiple-baseline across-participants design to examine the effects of individualized ACT at school. Behavior Analysis in Practice, 15(1): 141-154. https://doi.org/10.1007/s40617-021-00558-8