Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

The Application of Real-time Emotion Recognition in Video Conferencing

LIN, Weikun [1*]

[1] Shandong University of Science and Technology, China

**** LIN, Weikun is the corresponding author, E-mail: welton.lin2233@gmail.com***

Abstract: Video conferencing has become a crucial tool for global remote collaboration, but existing platforms have limitations in capturing and conveying participants' emotions. Real-time emotion recognition technology, by combining computer vision, deep learning, and temporal analysis, can automatically analyze and identify emotional changes in participants, addressing this gap. This paper first introduces the core technical processes of emotion recognition, including face detection, feature extraction, and emotion classification, with a focus on the technical details of using Convolutional Neural Networks (CNNs) for feature extraction and classification algorithms.[1] To enhance the system's temporal dynamics, the paper also presents methods for capturing emotional changes based on Long Short-Term Memory networks (LSTMs) or Temporal Convolutional Networks (TCNs). In particular, we use a Softmax classifier for probability estimation of emotions, coupled with temporal analysis methods for real-time engagement assessment in meetings. Furthermore, this paper discusses the system implementation under an edge computing and cloud collaboration framework to optimize real-time performance and computational efficiency while proposing security strategies focused on privacy protection, such as homomorphic encryption and federated learning. Through specific application examples, this paper demonstrates the significant role of real-time emotion recognition technology in video conferencing and its potential impact on future remote communication modes.

Keywords: Real-time Emotion Recognition, Video Conferencing, User Experience, Technical Challenges, Deep Learning, Multi-modal Analysis, Privacy and Security, Edge Computing.

Disciplines: Computer Science. **Subjects:** Deep Learning.

DOI: https://doi.org/10.5281/zenodo.13926257 **ARK:** https://n2t.net/ark:/40704/JCTAM.v1n4a10

1 INTRODUCTION

Video conferencing has quickly become an indispensable tool for multinational companies, educational institutions, and government organizations seeking to collaborate remotely. Unfortunately, due to a lack of face-to-face emotional communication during video conferencing sessions it can often prove challenging for participants to accurately convey and recognize each other's emotions which ultimately reduces its efficacy as communication tool. [2] To address this gap in remote meetings and online communication scenarios real-time emotion recognition technology has been developed and widely implemented as a solution.

Emotion recognition technology's core consists of automating facial expression analysis using image processing and machine learning methods, in order to determine participants' emotional states. Emotion recognition tasks can be divided into several steps, such as face detection, feature extraction, emotion classification and sentiment inference. These steps usually entail using traditional algorithms in computer vision, such as Haar feature classifiers and Convolutional Neural Networks (CNNs). Video conferencing represents a dynamic situation where participants' facial expressions may change over time; hence modeling temporal information accurately to accurately recognize emotional states is of utmost importance for accurate recognition. [3] Long Short-Term Memory networks (LSTMs) and Temporal Convolutional Networks (TCNs) have become the go-to solutions in this realm for processing sequential data in this domain.

Real-time systems present several technical hurdles beyond expression recognition, including low-latency processing with high accuracy and analyzing facial data while protecting user privacy. To address these issues, lightweight neural network architectures (like MobileNet and EfficientNet) and edge computing frameworks have recently been introduced in order to reduce computational overhead caused by network transmission, while privacy protection technologies like homomorphic encryption or federated learning are gradually being integrated into emotion recognition systems to protect their users' facial data from misuse or leakage.

This paper's objective is to explore systematically the

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

fundamental theories, implementation methods, application scenarios and challenges faced by real-time emotion recognition technologies that focus on video conferencing applications. We will explore emotion recognition via mathematical models and algorithmic details, while showcasing its advantages in remote meetings via concrete application examples.[4]With the development of 5G communication and artificial intelligence technologies, real-time emotion recognition could emerge as a critical way to enhance emotional communication in remote interactions and provide technical support for video conferencing systems' intelligence and emotionalization capabilities.

2 OVERVIEW OF REAL-TIME EMOTION RECOGNITION TECHNOLOGY

The core of real-time emotion recognition technology lies in extracting facial expression information from video frames, using machine learning or deep learning models to classify expressions, and applying the classification results to emotional analysis. In the context of video conferencing, this process must occur with low latency and high accuracy to ensure that participants receive timely emotional feedback. The workflow of emotion recognition typically includes steps such as face detection, facial feature extraction, expression classification, and emotion inference, each involving multiple algorithms and technical details.[6]

2.1 Basic Workflow of Emotion Recognition

2.1.1 Face Detection

Face detection is the first step in emotion recognition, aiming to locate the facial regions of participants from the input video or images. Common face detection algorithms include:

Haar Cascade Classifier: A classifier based on Haar features and the AdaBoost algorithm that quickly detects facial features in images. The detection process involves sliding a window across the image to detect local variations in grayscale.

HOG (Histograms of Oriented Gradients): This method extracts features by calculating histograms of gradient orientations in the image and uses Support Vector Machines (SVM) for classification detection.

MTCNN (Multi-task Cascaded Convolutional Networks): This is a deep learning-based multi-task cascaded network that can not only detect faces but also locate facial landmarks such as eyes and mouth. It achieves efficient detection and alignment through a cascaded CNN structure.[7]

Mathematically, let the input image be $I(x, y)$. Face detection can be represented by the detection function $D(I)$, which outputs a set of bounding boxes for the faces detected, $\{(x_i, y_i, w_i, h_i)\}$, where (x_i, y_i) is the top-left corner of the i-th detected face, and w_i and h_i are the width and height of that face.

$$D(I) = \{(x_i, y_i, w_i, h_i)\} \text{ for } i = 1,2, \dots, n$$

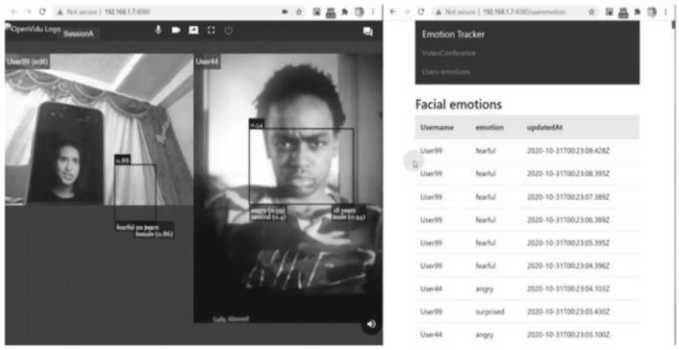

FIG. 1 EMOTIONAL TRACKER INTERFACE

2.1.2 Feature Extraction

Once the facial regions are detected, the next step is to extract features from these areas.[8] Feature extraction is typically accomplished using deep learning models, such as Convolutional Neural Networks (CNNs). CNNs can extract multi-level features from facial regions, ranging from local edges to global textures. These features are subsequently used for expression classification.

The working principle of a CNN involves a series of convolutional layers, pooling layers, and activation functions (such as ReLU) to extract features from the input image. Specifically, given the input facial image I_f, the convolution operation produces a feature map **F**:

$$\mathbf{F} = F(I_f) \in \mathbb{R}^{h \times w \times d}$$

where h and w are the spatial dimensions of the feature map, and d is the depth (number of channels) of the feature map. The convolution operation is defined as follows:

$$F(I_f) = \sigma\left(\sum_{i=1}^{k} W_i * I_f + b_i\right)$$

where $*$ denotes the convolution operation, W_i is the convolution kernel, b_i is the bias term, and σ is a nonlinear activation function (such as ReLU). Through multiple layers of convolution, higher-dimensional feature representations can be extracted.[12]

2.1.3 Expression Classification

The feature map **F** obtained from the feature extraction network is input into a classification network, typically comprising fully connected layers (FC) and a Softmax classifier, for expression classification. The Softmax classifier converts the output of the network into a probability distribution of expression categories. Let $P(y|I_f)$ be the predicted probability of image I_f belonging to expression category y.[16] Its computation formula is:

$$P(y|I_f) = \text{Softmax}(\mathbf{W} \cdot \mathbf{F} + \mathbf{b})$$

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

where **W** is the weight matrix of the fully connected layer, and **b** is the bias term. The Softmax function is defined as:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

where K is the number of expression categories (for example, the seven common expressions: anger, disgust, fear, happiness, sadness, surprise, neutral).

The goal of expression classification is to maximize the probability of correct classification, which is achieved by minimizing the cross-entropy loss function. The cross-entropy loss is defined as:

$$L = -\sum_{i=1}^{N} y_i \log P(y_i|I_f)$$

where y_i is the true label, and N is the number of training samples.

2.2 Temporal Dependency in Emotion Recognition

In video conferencing, expressions are not only instantaneous but also exhibit temporal dependencies. Therefore, emotion recognition based solely on a single frame may not fully reflect the emotional state of participants. To capture the changes in expressions over time, temporal models such as Long Short-Term Memory networks (LSTM) or Temporal Convolutional Networks (TCN) are typically used to model the emotional dynamics in time series data.

2.2.1 Long Short-Term Memory Networks (LSTM)

LSTM is a variant of recurrent neural networks (RNNs) that can capture long-range dependencies in information. It uses memory cells to store historical information, effectively addressing the long-term dependency issues in time series data. The basic unit of LSTM consists of an input gate, a forget gate, and an output gate, with update equations as follows:

$$\begin{aligned}
i_t &= \sigma(W_i[x_t, h_{t-1}] + b_i) \\
f_t &= \sigma(W_f[x_t, h_{t-1}] + b_f) \\
o_t &= \sigma(W_o[x_t, h_{t-1}] + b_o) \\
\tilde{C}_t &= \tanh(W_C[x_t, h_{t-1}] + b_C) \\
C_t &= f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \\
h_t &= o_t \cdot \tanh(C_t)
\end{aligned}$$

where i_t, f_t, and o_t represent the input gate, forget gate, and output gate, respectively. C_t is the cell state, h_t is the hidden state, x_t is the input at the current time step, W is the weight matrix, σ is the Sigmoid activation function, and $\tanh$ is the hyperbolic tangent function.[19]

By using LSTM, emotion recognition can be performed on sequential expression data without losing historical information, thereby improving the accuracy of emotional analysis.

2.2.2 Temporal Convolutional Networks (TCN)

TCN is another method for modeling temporal dependencies. Unlike LSTM, TCN uses 1D convolution operations to capture dependencies in time series data. The convolution operations in TCN are based on causal convolutions and dilated convolutions, which allow the network to maintain causality in time series and expand the temporal window for capturing longer dependencies through dilation rates.[20] The core equation is:

$$\mathbf{h}_t = \sum_{k=0}^{K-1} W_k \cdot \mathbf{h}_{t-k \cdot d}$$

where W_k is the convolution kernel, K is the size of the convolution kernel, and d is the dilation rate. By adjusting the value of d, the receptive field can be enlarged to capture distant temporal dependencies.

2.3 Expression Recognition and Emotion Inference

After completing expression classification, emotional states of participants can be inferred from the probability distribution. The commonly used emotional categories include "positive" (such as happiness and surprise) and "negative" (such as anger and sadness). To further improve accuracy, multimodal emotion recognition can be introduced, which integrates facial expressions with other modalities such as voice and text to generate a more comprehensive emotional analysis result.

3 APPLICATION OF REAL-TIME EMOTION RECOGNITION IN VIDEO CONFERENCING

With the global proliferation of video conferencing, real-time emotion recognition technology introduces a new interactive way for remote communication. By capturing and analyzing participants' facial expressions, this technology can convey emotional information in real-time, helping participants better understand each other's emotional states and thus enhance the interaction and communication efficiency of meetings. This section will delve into the typical application scenarios, technical implementations, and challenges faced by real-time emotion recognition in video conferencing.[21]

3.1 Application Scenarios

3.1.1 Participant Emotional Feedback

In video conferences, real-time emotional feedback can help hosts or speakers understand the audience's reactions promptly. For example, a speaker can determine whether a particular segment has captured the audience's attention or resonance or if there is confusion due to complex content by analyzing the detected facial expressions. [22]This immediate feedback mechanism aids the host in adjusting the meeting's pace or content, enhancing interactivity and

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

avoiding the monotony of one-sided communication.

Real-time emotion recognition technology utilizes classification algorithms to map expressions to specific emotional categories such as happiness, confusion, anger, and surprise. Suppose at each time point t, the system detects n participants' expressions; $P(y_t^{(i)}|I_t^{(i)})$ denotes the expression classification probability of the i-th participant at time t. The system can calculate the emotional statistical distribution of all participants:

$$\mathbf{P}_t = \frac{1}{n}\sum_{i=1}^{n} P(y_t^{(i)}|I_t^{(i)})$$

Where $\mathbf{P}_t$ is the emotional distribution of the entire meeting room at time t. This information can be used to generate real-time emotional feedback reports, allowing the host to monitor the meeting's emotional atmosphere at any time.

3.1.2 Meeting Participation Assessment

Participation in a meeting is a crucial indicator of its effectiveness. By employing emotion recognition technology, one can infer participants' states, such as focus, confusion, or distraction. By integrating temporal information and emotional recognition results, the system can automatically generate participation curves that quantify each participant's emotional changes throughout the meeting.

For instance, the participation level $E_i(t)$ of each participant i at time t can be determined by the weights of their expression categories and the corresponding participation values for those categories. Defining a set of emotional categories $\{y_1, y_2, \ldots, y_k\}$ with each category having a participation weight $w(y_j)$, the participation level can be expressed as:

$$E_i(t) = \sum_{j=1}^{k} P(y_t^{(i)} = y_j|I_t^{(i)}) \cdot w(y_j)$$

This participation curve aids organizers in analyzing participants' emotional reactions post-meeting, identifying critical moments, and potential areas for improvement.

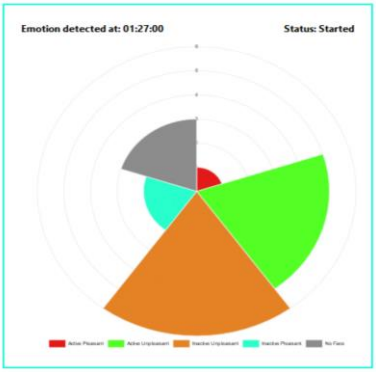

FIG 2. A SAMPLE VISUALIZATION FOR EMOTIONAL ASSESSMENT FROM THE EDUFERA WEB APPLICATION

3.1.3 Mental Health Monitoring

In prolonged meetings or high-pressure environments, real-time emotion recognition technology can assist in monitoring participants' emotional fluctuations, identifying potential mental health issues.[23] By detecting long-term emotional lows or anxiety states, the system can prompt organizers to pay attention to specific participants' emotional changes, preventing the accumulation of psychological problems.

Assuming the expression recognition results over consecutive T time points are $\{y_1^{(i)}, y_2^{(i)}, \ldots, y_T^{(i)}\}$, the emotional stability S_i of participant i during the period T can be defined to quantify emotional fluctuation. Emotional stability can be calculated based on the rate of change of emotional categories $\Delta y_t^{(i)}$:

$$S_i = \frac{1}{T}\sum_{t=1}^{T-1} \left|P(y_{t+1}^{(i)}|I_{t+1}^{(i)}) - P(y_t^{(i)}|I_t^{(i)})\right|$$

A higher emotional fluctuation rate S_i may indicate emotional instability in the participant, warranting particular attention to their mental health status.

3.2 TECHNICAL IMPLEMENTATION

3.2.1 Architecture of Video Conferencing Platforms

Integrating real-time emotion recognition technology into video conferencing platforms requires the integration of facial expression recognition algorithms with video stream processing, temporal data management, and real-time feedback mechanisms. A typical implementation architecture includes the following modules:

Video Stream Processing Module: Captures and preprocesses real-time video streams during the conference, including video resolution adjustment, frame rate control, and video segmentation.

Expression Recognition Module: Utilizes convolutional neural networks (CNN) or pre-trained models to detect and classify facial regions in each video frame. This module can incorporate lightweight networks like MobileNet to ensure real-time performance, while also combining edge computing when necessary to reduce server load.

Emotion Inference Module: Classifies the emotions of the recognized expressions in real-time and generates emotion change curves based on the expression sequences. By incorporating temporal models like LSTM or TCN, the system can capture the trends of participants' emotional changes.

Feedback and Interaction Module: Provides real-time emotional analysis results through a visual interface to participants or the host. For instance, it can display emotional distribution maps or participation curves on the interface, aiding meeting organizers in decision-making.

3.2.2 Achieving Low-latency Expression Recognition

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

To realize low-latency expression recognition, the system must optimize the data processing chain. The following technical measures can reduce latency:

Video Stream Downsampling: Reducing the resolution or frame rate of the video stream can decrease processing time per frame while ensuring that recognition accuracy is not significantly impacted.

Edge Computing: Performing preliminary expression detection and feature extraction on local devices or edge servers before uploading results to the cloud for further analysis can alleviate delays caused by network transmission.

Lightweight Models: Adopting lightweight convolutional neural networks such as MobileNet or SqueezeNet can significantly reduce model computational complexity while maintaining high recognition accuracy, thus enhancing real-time performance.

Mathematically, if we assume the model's computational complexity is $\mathcal{O}(n^2)$, employing lightweight networks can reduce complexity to $\mathcal{O}(n)$, significantly shortening processing time. The inference delay t_f for each image frame can be expressed as:

$$t_f = \frac{C}{f} + t_n$$

Where C represents the model's computational complexity, f is the processing frame rate, and t_n is the network transmission delay. By reducing both C and t_n, the overall system latency can be effectively minimized.

3.3 Challenges of Real-time Emotion Recognition in Privacy and Security

The implementation of expression recognition technology in video conferencing involves extensive personal facial data, making privacy and data security significant concerns. In the technical implementation, the following privacy protection strategies have been applied:

Homomorphic Encryption: By encrypting facial data using homomorphic encryption, computations for expression recognition can be performed without decrypting the data, safeguarding user privacy. Homomorphic encryption allows operations to be conducted in the encrypted domain, represented by the following formula:

$$E(f(x)) = f(E(x))$$

Where $E(x)$ denotes the encrypted data, and f is the feature extraction or classification function for expression recognition.

Federated Learning: Federated learning enables multiple video conferencing clients to collaboratively train expression recognition models without sharing raw data. Each client only transmits model gradients or weight updates to a central server, rather than original facial data, thereby protecting privacy.

$$\Delta w = \frac{1}{n}\sum_{i=1}^{n} \Delta w_i$$

Through federated learning, the system can continuously improve the performance of the expression recognition model while maintaining user privacy.

4 Implementation Solutions and Optimization

Implementing real-time emotion recognition technology in video conferencing involves multiple stages, from data acquisition and model training to system integration. This section will delve into the components of the implementation solutions and the optimization strategies targeting real-time performance, accuracy, and user experience.

4.1 Implementation Solutions

4.1.1 Data Acquisition

Data acquisition is the first step in the real-time emotion recognition system, involving the capture of video streams and facial images from participants. An effective data acquisition plan should include the following steps:

Video Source Selection: Choose appropriate video sources, such as HD cameras, smartphone cameras, or professional video conferencing equipment, to ensure image quality.

Image Preprocessing: During data acquisition, perform the following preprocessing on the images:

Grayscale Conversion: Convert color images to grayscale to reduce computational complexity.

Image Enhancement: Use techniques like histogram equalization to improve image contrast, ensuring facial features are clearly visible.

Face Detection: Apply face detection algorithms (e.g., Haar cascade classifiers, MTCNN, or YOLO) to locate facial regions, preparing for subsequent emotion analysis.

Real-Time Capture: Design the system architecture to ensure real-time video stream capture, maintaining high frame rates (e.g., 30fps or 60fps) to capture rapidly changing facial expressions.

4.1.2 Emotion Recognition Model Design

The emotion recognition model is the core of the system and typically employs deep learning methods for facial expression classification. The main steps in model design are as follows:

Model Selection: Select an appropriate deep learning architecture based on task requirements. Common models include:

Convolutional Neural Networks (CNNs): Suitable for

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

feature extraction and classification of image data.

Recurrent Neural Networks (RNNs): Suitable for processing time-series data, capable of capturing dynamic changes in facial expressions.

Hybrid Models: Combine the advantages of CNNs and RNNs, using CNNs to extract facial features and RNNs to process temporal information.

Dataset Preparation: Choose suitable emotion datasets for model training, such as:

FER2013: A dataset containing facial images with multiple emotion labels.

CK+: Composed of sequences of facial expressions labeled with seven basic emotions.

AffectNet: A large labeled dataset suitable for emotion recognition research.

Model Training: Train the model using a cross-entropy loss function, with optimization algorithms like Adam or SGD. During training, data augmentation techniques (e.g., rotation, flipping, scaling) can be employed to enhance the model's generalization ability.

Mathematically, the cross-entropy loss function can be expressed as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log\left(p_{ij}\right)$$

where N is the number of samples, C is the number of classes, y_{ij} is the true label of sample i, and p_{ij} is the predicted probability of sample i belonging to class j.

4.1.3 System Integration

Integrate various components to form a complete real-time emotion recognition system. The specific steps include:

Frontend Development: Use WebRTC or other streaming technologies to build the video conferencing frontend, facilitating video stream capture and user interface display.

Backend Processing: Set up a server to receive video streams sent from the frontend, perform emotion recognition, and provide real-time feedback of the recognition results to the frontend.

Data Storage and Analysis: Design a data storage solution for subsequent analysis and mining of recognition results.

4.2 Performance Optimization Strategies

To enhance the performance and user experience of the real-time emotion recognition system, the following optimization strategies can be adopted:

4.2.1 Model Compression and Acceleration

Model Pruning: Use pruning techniques to remove redundant neurons or connections, reducing model size and improving inference speed. This can be achieved by applying L1/L2 regularization constraints to diminish connections with smaller weights.

Quantization: Convert model weights and activation values from 32-bit floating-point numbers to low-bit integers (e.g., 8-bit) to reduce storage requirements and computational complexity, thereby accelerating model inference.

Knowledge Distillation: Transfer knowledge from a complex teacher model to a smaller student model, maintaining accuracy while reducing model size.[24]

Mathematically, the goal of knowledge distillation is to minimize the output difference between the teacher and student models, which can be measured using Kullback-Leibler divergence:

$$D_{KL}(p_{teacher} \parallel p_{student}) = \sum_{i} p_{teacher}(i)\log\left(\frac{p_{teacher}(i)}{p_{student}(i)}\right)$$

4.2.2 Real-Time Processing Optimization

Multithreading and Parallel Computing: Utilize multithreading techniques to implement parallel processing in data preprocessing, model inference, and result feedback, improving overall system responsiveness.

Edge Computing: Delegate emotion recognition tasks to user devices or edge servers to reduce network latency and enhance real-time capabilities.

Video Frame Selection: For rapidly changing scenarios, design adaptive frame selection strategies that only process important frames (e.g., frames with significant expression changes), minimizing unnecessary computations.

4.3 User Experience Optimization

Visual Feedback: Provide intuitive user interfaces that display real-time emotion analysis results, including graphical emotion distribution maps and participation curves, enhancing user engagement.

Privacy Protection: Offer users privacy options that allow them to choose whether to share emotion data, increasing the system's acceptability.

User Customization Settings: Allow users to adjust system parameters according to personal needs, such as emotion recognition sensitivity and feedback frequency, enhancing the personalization of user experience.

5 TECHNICAL CHALLENGES AND RESPONSE STRATEGIES

While real-time emotion recognition technology holds great promise for video conferencing, it faces numerous technical challenges during implementation. This section explores these challenges and proposes corresponding

response strategies to ensure the system's effectiveness and reliability.

5.1 Challenge 1: Diversity and Complexity

5.1.1 Challenge Description

Facial expressions vary significantly among different users, including differences in facial features, skin tone, age, and gender. Moreover, environmental factors (such as lighting changes and background complexity) can also impact the accuracy of facial expression recognition. Research shows that inconsistent lighting conditions can lead to a decline in recognition accuracy by over 70%.

5.1.2 Response Strategies

Data Diversity Enhancement:

Data Augmentation: Employ image augmentation techniques such as rotation, flipping, scaling, and color adjustment to increase the diversity of training data, thus improving the model's robustness.

Cross-Domain Data Collection: Collect data under varying lighting, backgrounds, and demographics to build a dataset with diverse expressions, enhancing the model's adaptability.

Adaptive Algorithm Design:

Lighting Compensation: Introduce image processing techniques, such as histogram equalization and gamma correction, to address uneven lighting conditions, thereby improving image quality and recognizability.

Dynamic Model Adjustment: Utilize adaptive learning algorithms to dynamically adjust model parameters according to environmental changes to maintain efficient recognition.

5.2 Challenge 2: Real-Time Performance Requirements

5.2.1 Challenge Description

Video conferencing requires high real-time performance, with model inference time needing to be controlled within milliseconds to ensure a smooth user experience. Current deep learning models often require longer computation times during inference, which can lead to latency.

5.2.2 Response Strategies

Efficient Model Design:

Lightweight Network Structures: Utilize lightweight network architectures like MobileNet or SqueezeNet to reduce computational resource consumption and memory usage, thereby speeding up inference.

Model Pruning and Quantization: Reduce redundant connections through model pruning while applying quantization techniques to convert model parameters into lower-bit values, significantly increasing computation speed.

Inference Process Optimization:

GPU Acceleration: Leverage the parallel processing capabilities of GPUs for model inference to substantially improve processing speed.

Edge Computing: Perform computations close to the user on edge devices to reduce data transmission latency, enabling real-time processing.

5.3 Challenge 3: Privacy and Security Issues

5.3.1 Challenge Description

Real-time emotion recognition involves sensitive user information, which may raise concerns about privacy breaches and data security. This can lead to a decrease in user trust and potential legal and compliance risks.

5.3.2 Response Strategies

Data Anonymization:

De-identification: Remove identifiable user information during data collection to ensure users' privacy is not compromised.

Local Processing: Conduct emotion recognition on the user's device, avoiding the transmission of raw video streams to the cloud, thereby reducing the risk of data leaks.

Secure Transmission Protocols:

Encrypted Communication: Utilize encryption protocols like SSL/TLS to safeguard data transmission, ensuring user data is not intercepted during transit.

User Consent Mechanisms: Implement mechanisms for user consent to ensure users have clear rights and choices regarding the use and storage of their data.

5.4 Challenge 4: Accuracy of Emotion Recognition

5.4.1 Challenge Description

The accuracy of emotion recognition is influenced by various factors, including facial occlusion, subtle changes in expressions, and inconsistencies in users' emotional displays. Research indicates that the accuracy of emotion recognition can vary by over 20% in different contexts.

5.4.2 Response Strategies

Multi-Modal Fusion:

Combining Other Signals: In addition to facial expressions, incorporate voice emotion analysis and physiological signals (such as heart rate and skin conductance)

to enhance the accuracy of emotion recognition.

Deep Learning Fusion Models: Design deep fusion models that jointly learn information from different modalities to capture richer emotional features.

Dynamic Emotion Recognition:

Temporal Feature Extraction: Utilize sequential models like RNNs or LSTMs to capture dynamic features of emotion changes, improving the accuracy of emotion recognition.

Incremental Learning: Employ incremental learning methods that allow the model to continuously update based on new user data, adapting to individual user differences.

6 CONCLUSION

Real-time emotion recognition technology represents a revolutionary shift in modern communication methods. As remote work and online education become more prevalent, users' desire for emotional communication among themselves has only grown more pressing. This paper explores real-time emotion recognition technology's concept, implementation schemes, application scenarios, technical challenges and associated strategies, highlighting its significance for improving user experience and effective communication.

6.1 Importance and Application Prospects of Blockchain Technology

Real-time emotion recognition technology not only assists meeting hosts in recognizing participant emotions but can also enhance meeting interactions and engagement, by giving real-time feedback on user emotions; hosts can then adjust their speaking strategies according to participants' emotional conditions, further improving meeting effectiveness. It has numerous applications in online education, psychological counseling and telemedicine - contributing to improved learning outcomes as well as mental health management for users.

6.2 Implementation and Optimization Processes

Implementation plans for real-time emotion recognition technology comprise of several steps, such as data collection, emotion recognition model design, and system integration. Deep learning methods are employed in model design to increase recognition accuracy and efficiency; furthermore strategies like model compression, real-time processing optimization, user experience enhancement further enhance system performance to ensure smooth real-time emotion recognition operations across different network environments.

6.3 Technical Challenges Confronted

Real-time emotion recognition technology offers many applications; however, its implementation presents numerous obstacles, such as diversity and complexity issues, performance requirements for real time emotion recognition systems, privacy/security concerns and accuracy of emotion recognition. To address these challenges, this paper proposes strategies such as increasing diversity, designing adaptive algorithms with efficient model designs that ensure anonymization/data fusion to maintain sustainability/reliability in technology implementation.

Real-time emotion recognition in video conferencing offers great promise. Thanks to recent technological advances and effective responses to challenges, this field is on a steady upward trend. Future research should not only optimize technology but also consider ethical, privacy, and user experience factors to ensure its acceptance in practical applications.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.Due to the order of manuscript receipt, this article is a supplementary publication and is not subject to the maximum number of articles

published by each author per issue.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

LIN, Weikun

Software Engineering, Shandong University of Science and Technology, Shandong, China.

REFERENCES

[1] Ba, D., & Zeng, L. (2021). A review of deep learning for emotion recognition from facial expressions. Journal of Visual Communication and Image Representation, 78, 103034. https://doi.org/10.1016/j.jvcir.2021.103034

[2] Chen, Y., Liu, M., & Zhang, J. (2020). Real-time emotion recognition from facial expressions in video conferences. Proceedings of the International Conference on Multimedia and Expo (ICME), 1-6. https://doi.org/10.1109/ICME49207.2020.00128

[3] D'Mello, S. K., & Kory, J. (2015). A review of affective computing: From emotion recognition to social signal processing. ACM Computing Surveys, 47(3), 1-36. https://doi.org/10.1145/2723158

[4] Zhao, G., Song, C., & Wu, B. (2024). 3D Integrated Circuit (3D IC) Technology and Its Applications. Journal of Industrial Engineering and Applied Science, 2(4), 60–65.

[5] Wu, B., Song, C., & Zhao, G. (2024). Applications of Heterogeneous Integration Technology in Chip Design. Journal of Industrial Engineering and Applied Science, 2(4), 66–72.

[6] Song, C., Wu, B., & Zhao, G. (2024). Optimization of Semiconductor Chip Design Using Artificial Intelligence. Journal of Industrial Engineering and Applied Science, 2(4), 73–80.

[7] Song, C., Wu, B., & Zhao, G. (2024). Applications of Novel Semiconductor Materials in Chip Design. Journal of Industrial Engineering and Applied Science, 2(4), 81–89.

[8] Li, W. (2024). Transforming Logistics with Innovative Interaction Design and Digital UX Solutions. Journal of Computer Technology and Applied Mathematics, 1(3), 91-96.

[9] Li, W. (2024). User-Centered Design for Diversity: Human-Computer Interaction (HCI) Approaches to Serve Vulnerable Communities. Journal of Computer Technology and Applied Mathematics, 1(3), 85-90.

[10] Hossain, M. S., & Muhammad, G. (2020). Emotion recognition using deep learning approach: A review. IEEE Access, 8, 23906-23921. https://doi.org/10.1109/ACCESS.2020.2977602

[11] Liu, Y., & Wang, Q. (2020). Facial expression recognition: A comprehensive review. Computer Vision and Image Understanding, 192, 102896. https://doi.org/10.1016/j.cviu.2019.102896

[12] Zhao, Y., Zhang, Z., & Chen, L. (2019). A survey of emotion recognition methods: From the perspectives of images and videos. Journal of Visual Communication and Image Representation, 60, 53-68. https://doi.org/10.1016/j.jvcir.2019.07.016

[13] Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. Sensors, 18(2), 401. https://doi.org/10.3390/s18020401

[14] Li, S., Deng, W., & Du, J. (2019). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing, 28(1), 356-370. https://doi.org/10.1109/TIP.2018.2866741

[15] Picard, R. W. (1997). Affective computing. MIT Press.

[16] Li, W. (2024). The Impact of Apple's Digital Design on Its Success: An Analysis of Interaction and Interface Design. Academic Journal of Sociology and Management, 2(4), 14–19.

[17] Chen, Q., & Wang, L. (2024). Social Response and Management of Cybersecurity Incidents. Academic Journal of Sociology and Management, 2(4), 49–56.

[18] Song, C. (2024). Optimizing Management Strategies for Enhanced Performance and Energy Efficiency in Modern Computing Systems. Academic Journal of Sociology and Management, 2(4), 57–64.

[19] Chen, Q., Li, D., & Wang, L. (2024). Blockchain Technology for Enhancing Network Security. Journal of Industrial Engineering and Applied Science, 2(4), 22–28.

[20] Chen, Q., Li, D., & Wang, L. (2024). The Role of Artificial Intelligence in Predicting and Preventing Cyber Attacks. Journal of Industrial Engineering and Applied Science, 2(4), 29–35.

[21] Chen, Q., Li, D., & Wang, L. (2024). Network Security in the Internet of Things (IoT) Era. Journal of Industrial Engineering and Applied Science, 2(4), 36–41.

[22] Li, D., Chen, Q., & Wang, L. (2024). Cloud Security: Challenges and Solutions. Journal of Industrial Engineering and Applied Science, 2(4), 42–47.

[23] Li, D., Chen, Q., & Wang, L. (2024). Phishing Attacks: Detection and Prevention Techniques. Journal of Industrial Engineering and Applied Science, 2(4), 48–53.

[24] Song, C., Zhao, G., & Wu, B. (2024). Applications of Low-Power Design in Semiconductor Chips. Journal of Industrial Engineering and Applied Science, 2(4), 54–59.

[25] Tariq, U., Afzal, S., & Akhtar, U. (2022). Real-time emotion recognition in video conferencing using deep neural networks. Multimedia Tools and Applications, 81(15), 22213-22233. https://doi.org/10.1007/s11042-022-12716-8

[26] Zhao, G., & Pietikäinen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6), 915-928. https://doi.org/10.1109/TPAMI.2007.1110

[27] Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., & Girard, J. M. (2014). BP4D-spontaneous: A high-resolution spontaneous 3D dynamic facial expression database. Image and Vision Computing, 32(10), 692-706. https://doi.org/10.1016/j.imavis.2014.06.002