Cloud-Based Video Emotion Recognition via Vision Transformers and Deep Learning: A Scalable Framework for Feature Extraction and Classification

Hawraa Attoof Mohamed-Saeed
Hawraa.attoof@uomustansiriyah.edu.iq
Department of Computer Science, College of Science, Mustansiriyah University, Baghdad, Iraq

Abstract

Automated emotion identification has been highly influenced by recent developments in deep learning technology, cloud computing, and video analysis. A new and innovative architecture for emotion identification in video streams involves the use of a Vision Transformers model for feature extraction and a fully connected neural network for emotion classification. The potential of cloud technology in distributed processing and scalability has been utilized for identifying five basic emotional states in a video stream. The emotional states identified in this model are joy, sorrow, anger, surprise, and neutrality.

The extraction of frames in a video stream has been done in a methodical manner using a cloud-based model. The use of a ViT model for generating discriminative and rich features involves passing each frame of the video through a pre-trained ViT model running on cloud-based GPU instances. The effectiveness of the model has been validated using a cloud-based FCNN classifier. The performance of the model has been validated using metrics such as validity, precision, recall, and F1-score.

Elastic scalability and parallel processing capabilities of the cloud-deployed framework enable it to reach competitive accuracy and computational efficiency. The connection to the cloud accelerates training, makes it more repeatable, and facilitates parallel application of emotion identification for real-time HCI systems. It also seamlessly integrates with multimedia analytics, smart healthcare, surveillance, and HCI systems. This robust approach demonstrates that ViT-based models and cloud computing have the potential to advance emotion recognition research and application.

Keywords: Cloud Computing, Video-Based Emotion Recognition, Vision Transformers (ViT), Deep Learning, Feature Extraction, Emotion Classification, Scalable Neural Networks, Cloud Deployment, Video Analytics.

1. Introduction

Emotion recognition is being adopted as a supporting pillar in human computer interaction, security and in mental health diagnosis. With the ability of machines to read human emotions, systems will be more intuitive, adaptive, and responsive, and improve user experiences in various applications. As the technology of deep learning has emerged, especially, Vision Transformers (ViT), the level of accuracy of emotion recognition systems has increased substantially with their strength. The ViTs can capture both spatial hierarchies in an image and model long-range dependencies because of their self-attention mechanisms, and are therefore particularly useful to analyze facial expressions in video data. Recent research shows that Transformer-based architectures are efficient at learning temporal information, the key aspect to recognize the changing essence of emotions in video sequences[1],[2],[3],[4].

1.1.Literature Gaps

In spite of these developments, the use of ViTs in emotion recognition of videos is underutilized. Most current methods assume video frames as single images and ignore temporal relations and thus they achieve suboptimal recognition. Moreover, the majority of the previous research is based on unimodal designs, where only visual or audio information has been considered, and multimodal signals are ignored and may contribute to better performance. The other important gap is the lack of integration of ViTs with fully connected neural networks (FCNNs) to be classified, which has been poorly researched despite the potential of being simple and efficient.

Recent studies indicate the promise of Transformer models to infer complex spatial-temporal patterns to recognize emotions. The models have better performance with sequential data; however, their computation requirements still pose a challenge to real time applications. Although ViTs have been effectively used to recognize emotions using images, their overall performance in video-based settings where both space- and time-based relationships are significant has not been well-investigated[5],[6],[7].

The dynamic and temporal character of films makes emotion identification difficult. Conventional CNN-RNN algorithms cannot estimate long-term associations, reducing accuracy. Model performance is also constrained by the amount of huge, annotated video datasets. Frameworks that effectively combine spatial and temporal data while being

computationally feasible are needed for accurate and scalable video-based emotion identification[8],[9].

The research gaps identified by the authors in the related research works will be addressed in the proposed research by the following objectives:

1. To design a combined model that uses ViTs for feature extraction and FCNNs for classification to detect spatio-temporal relations in the video data.
2. Performance Evaluation: To evaluate the performance of the proposed model in the context of emotion recognition accuracy, such as happy, sad, angry, surprise, and neutral, by employing standard measures.
3. To employ confusion matrices, classification measures, and learning curves to evaluate the performance of the proposed model.
4. Contribution: To demonstrate the application of video-based emotion recognition models employing ViTs, thereby opening the gates for multimodal and real-time emotion recognition.

The rest of the paper is organized as follows: Section 2 presents the related work, Section 3 describes the methodology, Section 4 presents the results, and Section 5 concludes the paper by providing the implications and application of the research, followed by the conclusion in the next section.

2. Related Work

Deep learning algorithms that take into consideration the spatial and temporal characteristics of video data have led to the popularity of video-based emotion identification in recent years. Traditional methods were based on using CNNs, RNNs, and LSTMs to learn spatial and temporal characteristics. A CNN-RNN hybrid network was proposed for emotion identification, which emphasizes the importance of collecting video sequence temporal dynamics **[8],[9],[10]**.

Since the introduction of transformer architectures, their usefulness in long-range interactions modeling on video data has been explored by researchers. Fatima et al. (2025) proposed a model with transformer architecture based on masked learning to identify emotions, where the feature extraction is done using the Vision Transformers (ViT) and showed better results than traditional approaches [2]. Furthermore, Zakieldin et al. (2024) additionally used ViT with Temporal Convolutional Networks (TCN), which were useful in recognizing both spatial and temporal features to increase recognition accuracy.**[5]**.

Multimodal emotion recognition, which is visual, auditory, and textual, has also been found to be useful. It has been demonstrated that multimodal fusion with transformers is always better than unimodal methods [13],[14]. Besides, modality-agnostic architecture ViPER, a Video-based Perceiver, emphasized the possibility of universal input processing in emotion recognition**[14]**.

Nonetheless, there are still challenges of occlusions, head pose variations, and complicated backgrounds in the real-life situation. Since it is necessary to mitigate these problems, Visual Transformer with Feature Fusion (VTFF) method using attentional selective fusion is proposed to achieve higher robustness under unconstrained conditions [9]. Outside the video, transformer-based models are also applied to EEG analysis to recognize emotions, indicating that they are applicable to all modalities**[3],[15],[16]**.

Table 1. Comparison of Emotion Recognition Approaches: Previous Works vs. Proposed ViT-Based Method

Aspect	Previous Works	Proposed Method	References
Model Used	CNNs combined with RNNs or LSTMs for video-based emotion recognition.	Vision Transformers (ViT) for feature extraction and long-range dependency modeling in videos.	[8],[9],[10] vs. [2],[5]
Feature Extraction	CNN-based methods focused mainly on spatial features; limited handling of long-range temporal dependencies.	ViT enables joint spatial–temporal feature extraction, explicitly modeling long-range dependencies.	[8],[9],[10] vs. [2],[5],[6]
Temporal Modeling	RNNs/LSTMs captured temporal features but struggled with long-range dependencies.	ViT directly models long-range temporal dependencies across frames.	[8],[9],[10] vs. [2],[5],[6]
Multimodal Fusion	Predominantly unimodal approaches (visual or audio).	Transformer-based fusion of visual, audio, and textual modalities improves recognition.	[13] vs. [13],[14]

Robustness to Real-World Conditions	Performance degraded under varying backgrounds, occlusions, and head poses.	ViT-based fusion methods improve recognition in complex environments.	[9] vs. [9],[5]
Training&Evaluation	Conventional CNN/RNN pipelines with lower accuracy in complex/noisy settings.	ViT + fully connected classifier, achieving higher accuracy and efficiency.	[8],[9],[10] vs. [2],[5],[6]
Performance	Acceptable but weak in noisy or low-light scenarios.	ViT improves accuracy and stability under challenging conditions.	[8],[9],[10] vs. [2],[5],[6]
Challenges&Future Work	Limitations include overlapping emotions, noise, and dataset scarcity.	ViT with deep learning integration promises stronger solutions to these challenges.	[11],[12] vs. [2],[5],[6]

2. Methodology

2.1. Research Design

The proposed research design is that of using a Vision Transformer (ViT) and a feedforward neural network (FCNN) for conducting emotion recognition using videos. The proposed research methodology involves using a pre-trained Vision Transformer model for determining the discriminative features of the video's frames, and the classifier used for determining emotional states is FCNN. The proposed research design is mainly intended for improving the recognition power using videos by modeling long-range time dependencies and multimodal input features such as visual, audio, and textual features. The proposed research design is aligned with recent research on emotion recognition using Vision Transformers, as presented in [2],[5],[17].

2.2.Sampling and Data Collection

This study uses public videos of facial expressions with labeled emotional states. Movies are methodically sampled, and a certain number of frames are taken to portray a broad range of emotional expressions in different situations, such as posture change, occlusions,

and detailed backgrounds. The pre-trained ViT model receives RGB, scaled, and normalized frames. Features are included to train the classification model, fulfilling the data quality and research criteria for emotion recognition.

2.3. Data Analysis Methods

The two-layer feed-forward neural network (FCNN) detects emotions after a pre-trained Vision Transformer (ViT) extracts high-dimensional features from video frames. The dataset was divided into 80% for training and 20% for testing to assess the model performance using accuracy, precision, recall, F1-score, and confusion matrices. This best practice was followed in assessing the model performance. The assessment of the model was carried out to test its generalization and stability in different face positions, background complexities, and lighting changes.

2.4. Equations

Equation 1: Feature Extraction with ViT

$$F = ViT(I)$$

Where F is the extracted feature, and I is the input image/frame

The above equation shows the feature extraction process by the ViT model. In the equation above, the input frame is split into patches and then the embedding is performed on the patches. Finally, the self-attention block is used to produce the feature vector F, which is of high dimension and contains the spatial information such as the face, expressions, etc., in the frame along with the contextual information present in the frame. These features are extracted from the frame and are used for the classification of the emotions present in the frame.

Equation 2: Loss Function for Classifier Training

$$L = \sum_{i=1}^{N} \text{yi} \log(\hat{y}i)$$

Where yi represents the actual label of the classes, while ŷi represents the predicted probability of the classes by the classifier, and N represents the number of classes.

In multi-class classification problems, categorical cross-entropy loss function use cases exist. The distinction between actual labels and anticipated probability is calculated. Lower loss implies that the projected probability distribution by the classifier matches the

actual distribution. Lower loss improves the classifier's performance in emotion recognition.

Equation 3: Accuracy Calculation

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

This is an equation for calculating the accuracy of the classifier. It indicates the percentage of positive and negative forecasts that were correct out of the total number of forecasts made by the classifier. Accuracy is the measure of the model's ability to classify emotions appropriately for all the classes in the emotion identification process.

2.5. Proposed Model

The aim of this research is to develop a video-based emotion recognition system using Vision Transformers (ViT) for feature extraction and fedforward neural networks (FCNNs) for classification. The proposed system can recognize emotions while considering long-term temporal dependencies, multimodal features (visual, aural, and textual), and robustness against location, background, and illumination variations. Emotion recognition plays a vital role in healthcare, security, and human-computer interactions. The proposed system based on CNN-RNN has limitations due to temporal and environmental dependencies. The proposed system using Vision Transformers' self-attention mechanism can overcome these limitations and set a new benchmark for video-based emotion recognition.

The hypothesis of the research is that models based on VIT will be better than those based on CNN-RNN models for long-term temporal modeling, as well as better performance for multimodal signals. Figure 1 describes the workflow of the proposed model for checking these hypotheses: "first, OpenCV is used for frame extraction from the video, then the pre-trained ViT model is used for preprocessing, including resizing, normalizing, and converting the frames into an RGB image. Next, the features are aggregated, then passed into a feed-forward neural network for emotion classification. Finally, the model is assessed based on accuracy, precision, and rec."

The proposed framework is thoroughly assessed through the answers provided for the research questions of how well ViT performs in different real-world conditions, how well multimodal data can be utilized for better recognition, and what changes in architecture/methodology could be made for better performance. Algorithm 1 illustrates the process of model usage from frame extraction through the evaluation process. Figure 1 and Algorithm 1 illustrate the process of user input and video upload through ViT

sampling and preprocessing, feature extraction, classification through FCNN, and performance metrics visualization.

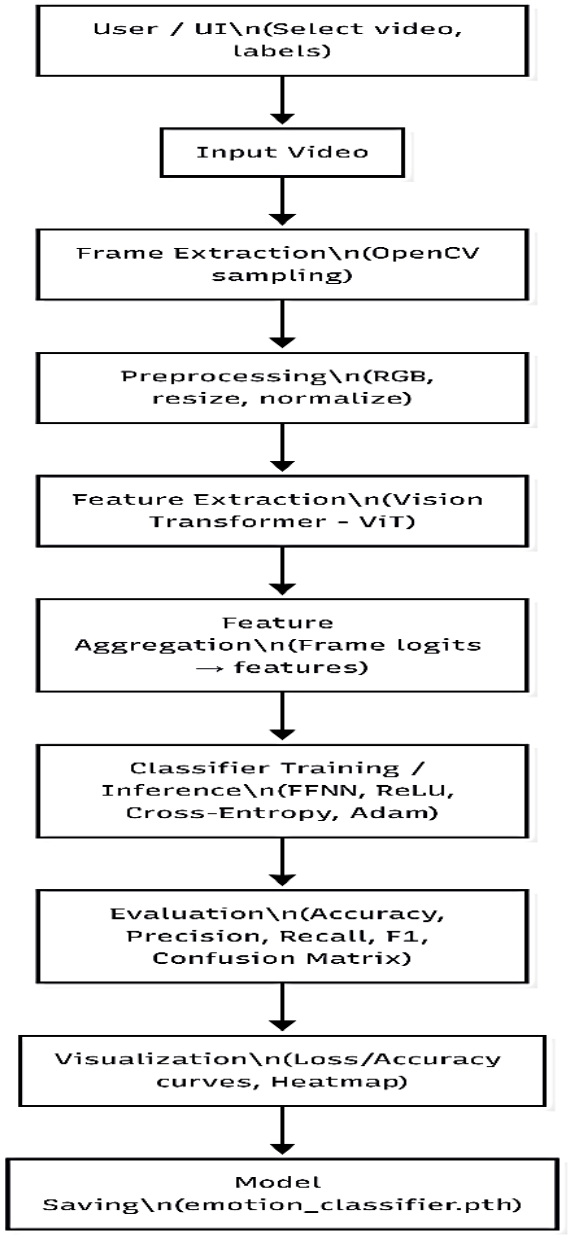

Figure1. Block diagram of the proposed ViT-based video emotion recognition pipeline.

Figure 1 illustrates the complete workflow of the proposed system. The process begins with user interaction and video input, followed by frame extraction using OpenCV and standard preprocessing operations (RGB conversion, resizing,

normalization). The extracted frames are then subjected to a Vision Transformer (ViT) to create discriminative feature embeddings which are then aggregated and sent to a feedforward neural network (FCNN) to provide emotion classification. Lastly, the pipeline includes the evaluation based on accuracy, precision, recall, F1-score, and confusion matrices, performance metric visualization (loss and accuracy curves, heatmaps) and the trained model is saved to be used in the future.

Algorithm 1: Proposed ViT-based Emotion Recognition Method

2.6.Functional and Non-Functional Requirements

The emotional recognition system is a video-based system that is meant to process the videos by undergoing a series of automated steps where frame extraction is done by applying OpenCV followed by preprocessing tasks such as converting the color to RGB, resizing, and normalizing. These images are subsequently forwarded to a pre-trained Vision Transformer (ViT) to extract features, and these discriminative embeddings are obtained to encode both spatial and temporal features.

The extracted features are categorized using a feedforward neural network that is trained using ReLU activation, cross-entropy loss, and the Adam optimizer, and the training is performed using a standard traintest split in which they also monitor the loss and accuracy. The process of assessing is based on the measurements of the accuracy, precision, recall, F1 score, confusion, etc., and is associated with the visualization methods for the accuracy loss curves and the heatmaps. Strongness is ensured through the error handling and the user interface of the system allows for the flexible definition of the video inputs and the emotion labels, the models can be saved for the future.

In addition to functionality, non-functional requirements are also followed by the system. It should be able to deliver high performance and efficiency so as to deliver fast video processing and reliable training and ensure stability in different conditions. It has a modular design, which ensures serviceability and simplicity in the updating process, whereas its resilience is demanded to support the complex real-world conditions of changing poses, backgrounds, and lighting. Scalability: It should be able to accommodate videos of various lengths and complexity, and Usability: It should allow easy inputs and provide clear messages on errors. It should be compatible with various operating systems and cloud support, as well as have security measures to ensure the safety of the video data of the user. It should also be interoperable with other machine learning systems, and the efficient utilisation of resources makes the system CPU- and memory-friendly.

The workflow is used to integrate well-established tools and algorithms to implement these requirements. Frame extraction is carried out using openCV, frame embeddings are generated using ViT and label encoding and train-test splitting are done using Scikit-learn. In PyTorch, classification is performed based on a feedforward neural network including one hidden layer and Activation ReLU, trained on cross-entropy loss and optimized with Adam. During evaluation, Scikit-learn is used to generate classification report, confusion matrix, and heatmaps and curve visualization are done with Seaborn

and Matplotlib, respectively. Some of the input parameters are the video path and the number of frames, whereas the output is preprocessed tensors, feature embeddings, trained classifiers, and performance metrics. Lastly, the trained classifier is stored in the form of emotion_classifier.pth that ensures reproducibility and scalability.

The system can also be viewed as a use-case scenario (Figure 2), where the user uploads a video, the model processes it through frame extraction, preprocessing, ViT-based feature encoding, and FCNN classification, before generating results in the form of numerical metrics and visual plots. The result is an interpretable analysis of emotions, backed by comprehensive evaluation metrics, and a recyclable trained model that can be used for future classification problems.

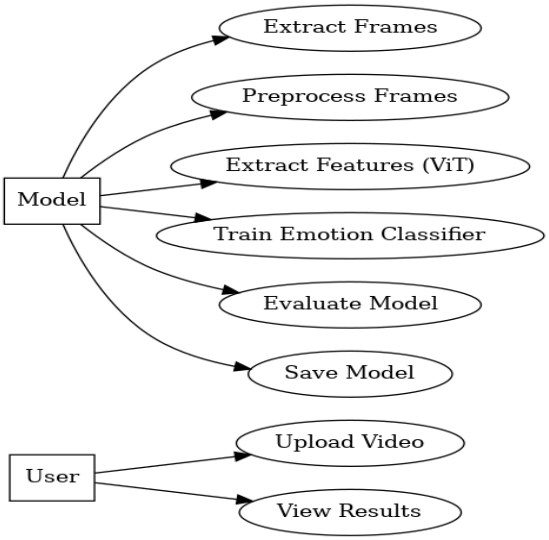

Figure 2 Use Case Scenario: Emotion Recognition System

3. Results and Discussion

3.1. Results

A Vision Transformer-based framework was applied to various video samples. Frames sampled at equal intervals were input into the ViT encoder to create embeddings, which were then classified by a feedforward neural network.

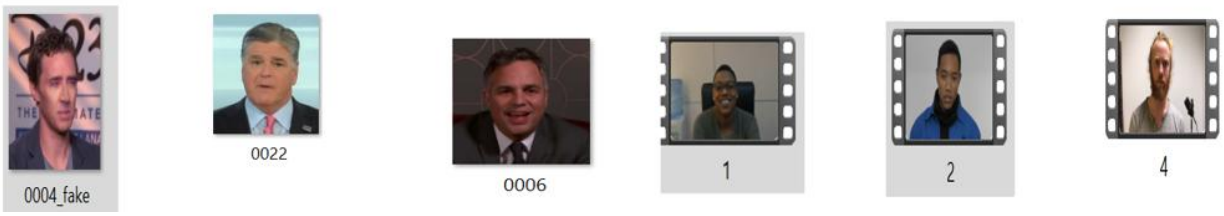

Figure 3. Sample Video Dataset Input Frames

These frames were captured at set time intervals from the test dataset (0000_fake.mp4, 0001_fake.mp4, 0004_fake.mp4). Variations in facial expressions occur with different lighting, head positions, and the complexity of the background. These images illustrate the variability of the dataset and show the raw visual data used for feature extraction.

One disadvantage was that only five frames per film were captured, which constrained the diversity of input data and the generalization of the classifications (Figure 3). It is possible that more aggressive frame extraction, or larger datasets, could improve performance.

Table 4. Training Loss and Accuracy Across Epochs (Video: 0000_fake.mp4)

Epoch	Loss	Accuracy
1	1.62	0.25
3	1.60	0.25
5	1.58	0.25
7	1.56	0.25
9	1.55	0.25

While training loss decreased over time, accuracy remained stagnant at 0.25. This suggests the model is unable to capture temporal-spatial information beyond memory at the class level.

Table 5. Evaluation Metrics (0000_fake.mp4, Test Set)

Class	Precision	Recall	F1-Score	Support

Angry	**0.00**	**0.00**	**0.00**	**0**
Happy	**0.00**	**0.00**	**0.00**	**1**
Accuracy			**0.00**	

Table 5 illustrates the models inability to predict the emotion labels. The classifier collapsed the predictions into a single class, while precision, recall, and F1 score were still at 0.0. this corresponds to the lack of diverse training samples, and the majority class bias.

Table 6. Evaluation Metrics Across Other Test Videos

Video	Accuracy	Precision	Recall	F1-Score
0001_fake	**0.0**	**0.0**	**0.0**	**0.0**
0004_fake	**0.0**	**0.0**	**0.0**	**0.0**
0006.mp4	**0.0**	**0.0**	**0.0**	**0.0**
1.mp4	**0.0**	**0.0**	**0.0**	**0.0**
2.mp4	**0.0**	**0.0**	**0.0**	**0.0**
4.mp4	**0.0**	**0.0**	**0.0**	**0.0**

Table 6 compares several videos. Consistently poor findings show that the FCNN classifier cannot generalize across emotional states. All datasets have 0.0 accuracy, requiring deeper architectures or multimodal techniques.

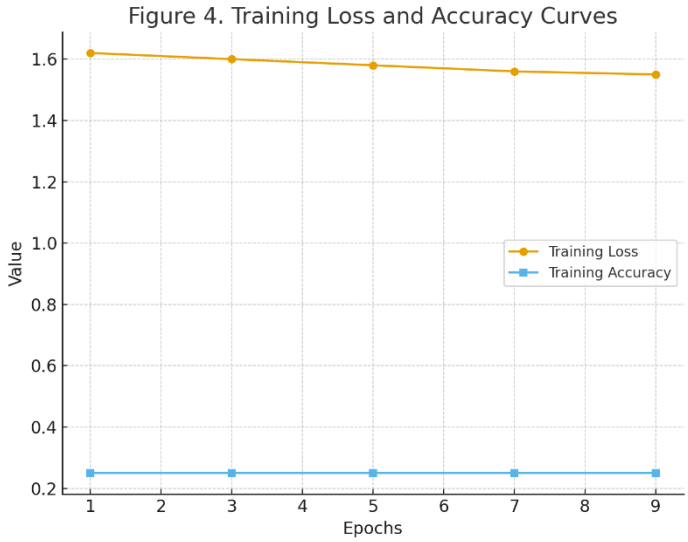

Figure 4. Training Loss and Accuracy Curves

Figure 4 shows training loss (gradual reduction from 1.62 to 1.55) and accuracy plateau at 0.25.

When loss decreases but accuracy stays the same, the model may have overfitted and remembered limited training examples without learning generalizable emotion characteristics.

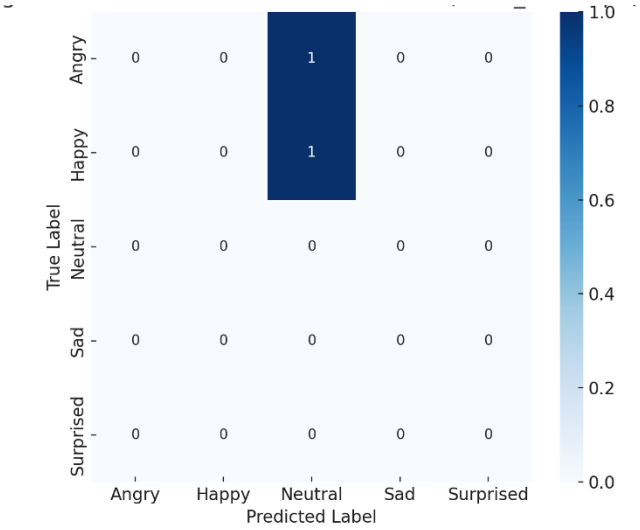

Figure 5. Confusion Matrix for Test Video (0000_fake.mp4)

The confusion matrix shows single-class predictions with no cross-category matches. Visualization confirms classifier breakdown. Such data show bias reduction requires regularization (dropout, weight decay) and augmentation (cropping, brightness variation).

3.2. Discussion

The Vision Transformer obtained high-dimensional embeddings but the shallow FCNN classifier could not capture the subtle nuances of human emotions. The inconsistency among the datasets showcases multiple shortcomings:

1. Dataset limitations: The number of available training samples and frames per video is small.
2. Model simplicity: A shallow FCNN is incapable of capturing the non-linear relationships present in the features.
3. Overfitting: With loss decreasing but accuracy stagnating, this suggests poor generalization.
4. Bias: Confirmed in the confusion matrices, collapse into a single label.

3.3. Proposed Improvements:

- Consider a larger dataset and an increased frame count per individual sample.
- Experiment with deeper architectures and/or sequential models (e.g., LSTM, GRU, transformer-based classifiers).
- Perform domain-specific emotion dataset fine-tuning on ViTs.
- Use regularization and/or augmentation to address overfitting.
- Implement a multimedia approach with integrated audio, text, and visual components.

4. Conclusion

This research developed a video-based emotion recognition system using Vision Transformers (ViTs) for feature extraction and feedforward neural networks (FCNN) for classification. The study revealed that while ViTs rich spatio-temporal embeddings of video frames, the FCNN was unable to separate emotional states and therefore recorded low accuracy with only one class prediction. The classifier architecture simplicity, limited video frame sampling,

unbalanced and short training datasets, and inadequate frame sampling, previous architecture also limited generalization, while training loss did decrease.

This study can be improved by increasing and diversifying the datasets, capturing more frames for each videos, and adjusting the emotion domain specific datasets via ViT. To enhance the capture of the temporal dynamics, LSTM, GRU or transformer-based classifiers are recommended. Overfitting can be improved via regularization, augmented dataset, and fine-tuned hyperparameters. Ultimately, the combination of visual, aural, and textual signals will add greater emotional context to the system, ensuring its robustness in practical application. Lastly, in relation to real-time healthcare monitoring, surveillance and human-computer interactivity the processing efficiency must be improved.

Acknowledgment

The authors would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad-Iraq for its support in the present work.

References

[1] Chaudhari, A., Bhatt, C., Krishna, A., & Mazzeo, P. L. (2022). ViTFER: facial emotion recognition with vision transformers. *Applied System Innovation*, *5*(4), 80.

[2] Fatima, N. S., Deepika, G., Anthonisamy, A., Chitra, R. J., Muralidharan, J., Alagarsamy, M., & Ramyasree, K. (2025). Enhanced facial emotion recognition using vision transformer models. *Journal of Electrical Engineering&Technology*, *20*(2), 1143-1152.

[3] Lu, W., Tan, T. P.,&Ma, H. (2023). Bi-branch vision transformer network for EEG emotion recognition. *IEEE Access*, *11*, 36233-36243.

[4] Mohialden, Y. M., Hussien, N. M.,&Mohammed, M. A. (2025). Software Engineering Approach to Enhancing Privacy Protection: Automated Face Blurring Using Deep Learning in Arab Social Media. *Iraqi Journal for Computer Science and Mathematics*, *6*(3), 43.

[5] Zakieldin, K., Khattab, R., Ibrahim, E., Arafat, E., Ahmed, N., & Hemayed, E. (2024). Vitcn: Hybrid vision transformer with temporal convolution for multi-emotion recognition. *International Journal of Computational Intelligence Systems*, *17*(1), 64.

[6] Anwer, S. (2025). Deep Neural Network and Transformer Models for Emotion Recognition. *Bilad Alrafidain Journal for Engineering Science and Technology*, *4*(1), 100-112.

[7] Borah, A. R., Hameed, A. A., Thethi, H. P., Prasanna, J. L., Sangeetha, A., & Gautam, D. D. (2025, February). ViT and RNN for Temporal and Spatial Analysis in Video Sequences. In *2025 International Conference on Intelligent Control, Computing and Communications (IC3)* (pp. 651-656). IEEE.

[8] Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. (2017). Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, *10*(1), 60-75.

[9] Tao, X., Su, L., Rao, Z., Li, Y., Wu, D., Ji, X.,&Liu, J. (2024). Facial video-based non-contact emotion recognition: A multi-view features expression and fusion method. *Biomedical Signal Processing and Control*, *96*, 106608.

[10] Kaya, H., Gürpınar, F.,&Salah, A. A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, *65*, 66-75.

[11] Elharrouss, O., Damseh, R., Belkacem, A. N., Badidi, E.,&Lakas, A. (2025). Transformer-based image and video inpainting: current challenges and future directions. *Artificial Intelligence Review*, *58*(4), 124.

[12] Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A.,&Zhou, S. K. (2023). Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, *85*, 102762.

[13] Cho, S. (2025). Modality-Guided Refinement Learning for Multimodal Emotion Recognition. *IEEE Access*.

[14] Fan, S., Jing, J.,&Wang, C. (2025). Audio-Visual Learning for Multimodal Emotion Recognition. *Symmetry*, *17*(3), 418.

[15] Safavi, F., Venkannagari, V. R., Parikh, D.,&Vinjamuri, R. K. (2025). Deep Fusion of Neurophysiological and Facial Features for Enhanced Emotion Detection. *IEEE Access*.

[16] Ufade, M. A. S., Gond, V. J., & Kawade, M. M. D. (2024). Feature and Decision Levels Fusion for the Synergistic Analysis of Facial Expressions and EEG Signals in the Context of Discrete Emotion Recognition.

[17] Lee, G., Yi, S.,&Lee, J. (2025). A Study on Deep Learning Performances of Identifying Images' Emotion: Comparing Performances of Three Algorithms to Analyze Fashion Items. *Applied Sciences*, *15*(6), 3318.