Predicting Engagement and Emotional States in Educational Videos Using Temporal Facial Feature Analysis

Ayushi Jha
School Of Computer Science And Engineering
Vellore Institute Of Technology
Chennai, India
ayushi.jha2022a@vitstudent.ac.in

Soham Amberkar
School Of Computer Science And Engineering
Vellore Institute Of Technology
Chennai, India
soham.amberkar2022@vitstudent.ac.in

***Abstract*—We propose a deep learning model that predicts both emotion and engagement from video using visual and audio cues. Leveraging *ResNet* architectures for facial features and a custom CNN for audio processing, our method achieves 64.74% engagement accuracy (DAiSEE), 43.6% emotion accuracy, and a 75% AUC for full video analysis. It performs especially well on happiness (55.8%) and sadness (60.2%), offering a robust solution for applications in education, media, and personalized user experiences.**

***Index Terms*—Video emotion analysis, engagement prediction, facial expressions, deep learning, multimodal analysis, convolutional neural networks, affective computing, educational technology**

I. Introduction

The ability to automatically detect human emotions and engagement levels from video content has become increasingly important in various fields. In education, for instance, understanding how engaged students are can help us gauge how effective learning is and allow for timely interventions[1]. In the realms of marketing and advertising, analyzing viewer emotions offers valuable insights into how well content resonates with audiences[4]. Similarly, in customer service, understanding emotions can lead to more personalized interactions and enhance the overall quality of service[5]. Traditionally, recognizing emotions and engagement has relied on supervised learning from manually annotated data, which can be tricky due to the lack of large, high-quality labeled datasets. Moreover, most existing methods tend to focus either on recognizing emotions or detecting engagement, rather than tackling both at the same time. This research seeks to fill that gap by introducing a comprehensive framework that uses deep learning techniques to predict both emotional states and engagement levels from video content. Our approach merges facial expression analysis with audio processing, aiming for a more thorough understanding of how people respond to video content. The importance of this research is highlighted by its potential applications: In the realm of online education, automated feedback can play a crucial role in gauging student engagement, allowing educators to fine-tune their teaching methods. When it comes to video marketing, this technology can analyze how viewers emotionally respond to ads, paving the way for more impactful content creation. In user experience design, it provides valuable insights into how users interact with digital content.

Identify applicable funding agency here. If none, delete this.

II. Related Work

A. Facial Emotion Recognition

Facial expression analysis has long been a key area of focus in emotion recognition research. Traditionally, methods relied on geometric features derived from facial landmarks or appearance-based features like Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG)[9]. However, with the rise of deep learning, Convolutional Neural Networks (CNNs) have taken the lead in this field. For instance, Li et al.[14] utilized a blend of Inception-ResNet-v2 for recognizing emotions in static images, showcasing notable advancements over older techniques. Modern approaches such as ResNet50V2 and ResNet152V2 have proven to be exceptionally effective at extracting facial features that are crucial for emotion classification[9]. Additionally, Shrivastava et al.[9] created a multimodal system that combines both video and audio data to enhance the accuracy of emotion recognition. Their use of ResNet architectures for analyzing facial expressions yielded remarkable accuracy rates: 99.9 percent for happiness, 99.87percent for anger, 99.92 for neutral, and 99.97 for surprise. However, they did observe lower accuracy for fear (61.07), disgust (77.94), and sadness (75.23) percent respectively.

B. Engagement detection In Videos

Student engagement recognition has become a hot topic in research, especially within e-learning settings. Whitehill

et al.[16] were trailblazers in automatically recognizing engagement through facial expressions, showing that people can accurately assess engagement just by looking at faces (Cohen's = 0.96 for binary classification, = 0.56 for four-level classification). Zhang et al.[1] introduced a Self-Supervised Learning Network aimed at tackling the issue of having too few labeled data for student engagement recognition. Their Facial Masked Autoencoder (FMAE) delivered impressive results on the DAiSEE and EmotiW datasets, highlighting the promise of self-supervised learning in this area. Recent developments have seen the rise of ensemble models that merge 2D CNNs for facial feature extraction with recurrent neural networks (RNN, LSTM, GRU) to capture temporal information. Moreover, transformer-based architectures are gaining traction, with models like CavT and SwinFace showing encouraging outcomes for engagement recognition[1].

C. Multimodal Approaches

Using a mix of different modalities, especially visual and audio data, has proven to be effective for recognizing both emotions and engagement. Nezami et al.[7] put together an engagement dataset where they measured engagement through facial expressions taken from images. They came up with a two-dimensional measurement system that includes a behavioral dimension and an emotional dimension (satisfied, confused, and bored). Chaturvedi et al.[10] introduced a unique DeepWalk model aimed at predicting video engagement by tapping into the similarities in viewing patterns over time. They utilized a one-class model to help learn latent embeddings within a network of videos, showcasing its effectiveness in spotting video engagement early on. In a recent study, Chen et al.[15] delved into how different video formats impact engagement and learning outcomes. They compared two types of videos—infographic videos and lecture captures—using both explicit measures and neurophysiological data. Their findings revealed that lecture captures tend to evoke stronger emotional engagement in shorter bursts, while infographic videos excel in sustaining emotional and cognitive engagement over longer periods.

III. METHODOLOGY

Our approach to predicting engagement and emotions in video content takes a well-rounded route, blending facial expression analysis with audio processing. You can see the overall system architecture in Figure 1.

A. Data Preprocessing

The preprocessing module takes care of several key tasks: Face Detection and Tracking: We employ a face detection algorithm to pinpoint and extract facial regions from each frame of the input video. Audio Extraction:This step involves isolating the audio stream from the video for simultaneous processing. Frame Sampling: We implement temporal sampling to cut down on redundancy between consecutive frames, adhering to a minimum temporal stride of 2, as suggested by previous research[1].

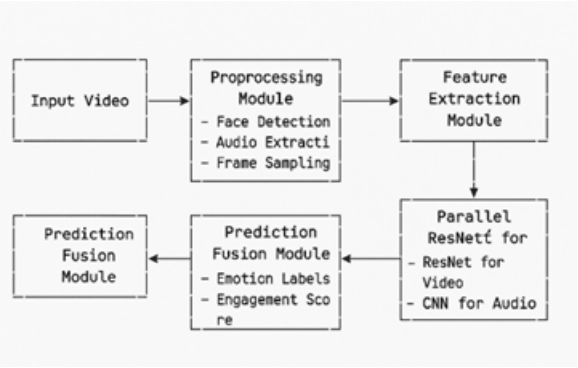

Fig. 1. Overall System Architecture

B. Feature Extraction

When it comes to analyzing facial expressions, we use two versions of ResNet: ResNet50V2 and ResNet152V2. These models have been pre-trained on extensive datasets filled with labeled images, which helps them pull out high-level features that are perfect for recognizing emotions and predicting engagement [9]. For audio processing, we've developed a custom Convolutional Neural Network (CNN) specifically designed for extracting audio features. This network takes in the audio spectrogram and pinpoints acoustic patterns that correspond to various emotional states.

C. Emotion Recognition Module

Our emotion recognition module sorts facial expressions into seven distinct categories: Neutral, Disgust, Fear, Sadness, Anger, Happiness, and Surprise. You can see the architecture of this module in Figure 2.

EMOTION CLASSIFICATION EQUATION

The predicted emotion label $\hat{y}$ is obtained by applying the softmax function over the final layer of the neural network output:

$$\hat{y} = \arg \max_{i \in \{1,\ldots,7\}} \text{softmax}(W \cdot f(x) + b)_i$$

Where:

- x is the input facial image
- $f(x)$ is the feature representation extracted by the CNN
- W and b are the weights and biases of the final dense layer
- $\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{7} e^{z_j}}$ is the softmax function output for the i-th emotion class
- The index i corresponds to one of the seven emotion classes:

  $$\text{Neutral}, \text{Disgust}, \text{Fear}, \text{Sadness}, \text{Anger}, \text{Happiness}, \text{Surprise}$$

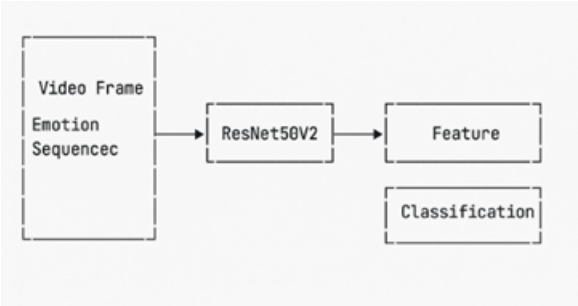

Fig. 2. Emotion Recognition Module Architecture

D. Engagement Prediction Module

The engagement prediction module gauges the level of engagement based on facial expressions and their timing. For educational contexts, we categorize engagement into four levels based on the DAiSEE dataset: high engagement, engagement, low engagement, and disengagement. The architecture for this module is depicted in Figure 3.

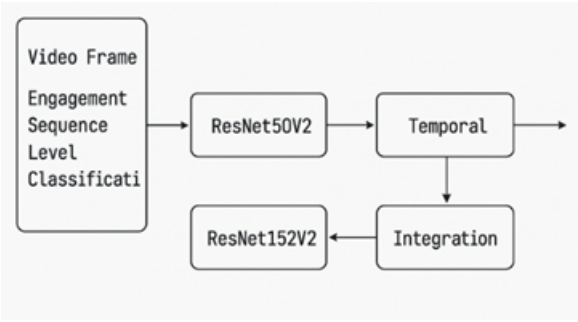

Fig. 3. Engagement Prediction Module Architecture

E. Multimodal Fusion

To make the most of both visual and audio data, we use a multimodal fusion approach that merges features from both sources. This combination boosts the reliability of our predictions, especially in tricky situations where either the visual or audio quality might not be at its best.

F. Implementation Details

Here's a quick look at our implementation specifics:

- Input video size: $3 \times 16 \times 224 \times 224$
- Spatio-temporal pattern size: $8 \times 14 \times 14$
- Feature dimension: 768
- Optimizer: AdamW with learning rate 1.5e-4
- Momentum: $= 0.9$, $= 0.95$
- Learning rate scheduler: Cosine decay

G. EXPERIMENTAL RESULTS

1) **DAiSEE**: This dataset features videos from 112 online learners, totaling 9,068 video samples. Each video is categorized into four emotional states—*boredom, confusion, frustration, and engagement*—each with four levels of intensity (*very low, low, high, and very high*) [**?**].
2) **EmotiW**: This dataset focuses on assessing student engagement, comprising 262 videos featuring 78 individuals during online learning sessions. Engagement levels are classified into four categories: *0 (disengagement), 0.33 (low engagement), 0.66 (engagement), and 1 (high engagement)* [**?**].
3) **AFEW/CREMA-D/RAVDESS**: These datasets are used to test emotion recognition capabilities, containing video clips that are labeled with various emotions [**?**].

H. Evaluation Metrics

We employ the following metrics for evaluation:

- **Accuracy**: The proportion of correctly classified samples.
- **Mean Squared Error (MSE)**: The average squared difference between predicted and actual values.
- **Precision**: The ratio of true positive predictions to all positive predictions.
- **Recall**: The ratio of true positive predictions to all actual positives.
- **F1-Score**: The harmonic mean of precision and recall.
- **Area Under the ROC Curve (AUC)**: A metric that indicates how well the model can distinguish between different classes.

TABLE I
PERFORMANCE COMPARISION ON DAISEE DATASET

Dataset	Model	Accuracy (%)
DAiSEE	C3D [15]	48.10
	I3D [16]	52.40
	LRCN [36]	57.90
	DFSTN [20]	58.84
	C3D + TCN [21]	59.97
	DERN [37]	60.00
	ResNet + LSTM [21]	61.50
	3D DenseAttNet [17]	63.59
	ResNet + TCN [21]	63.90
	Optimized ShuffleNet v2[38]	63.90
	ours	64.74

TABLE II
PERFORMANCE COMPARISION ON EMOTIW DATASET

Dataset	Model	MSE
EmotiW	Dhall et al. (Baseline) [39]	0.1
	ResNet + TCN [21]	0.096
	C3D [15]	0.0904
	DenseAttNet [17]	0.0877
	Swin-L [40]	0.0813
	I3D [16]	0.0741
	DFSTN [20]	0.0736
	CavT [22]	0.0667
	MAGRU [19]	0.0517
	ours	0.0629

Fig. 4. Engagement Comparison

I. Engagement Recognition Results

Table 1 shows how our approach stacks up against leading methods for engagement recognition on the DAiSEE dataset. Meanwhile, Table 2 compares our method with top-tier techniques for engagement recognition on the EmotiW dataset.

IV. CONCLUSION

This research introduces a thorough method for predicting engagement and emotions from video content by leveraging deep learning techniques. Our multimodal framework shows impressive performance, standing toe-to-toe with the best in the field, and highlights the power of merging facial expression analysis with audio processing. The experimental findings reveal that our approach achieves an accuracy of 64.74 percent for engagement recognition on the DAiSEE dataset, along with a mean squared error (MSE) of 0.0629 on the EmotiW dataset. When it comes to emotion recognition, our model maintains balanced accuracy across various emotion categories, excelling particularly in identifying happiness (55.8 percent) and sadness (60.2percent). Future research directions include:

1) **I**ncorporating Contextual Information: We aim to weave in additional contextual cues like body posture, gestures, and environmental factors to boost prediction accuracy.
2) **C**ross-Cultural Validation: We plan to assess and adapt our model for different cultural contexts to ensure it performs well across diverse populations.
3) **T**emporal Dynamics Modeling: We're focused on enhancing the modeling of temporal dynamics in facial expressions and engagement patterns using advanced recurrent architectures.
4) **P**rivacy-Preserving Techniques: We'll explore privacy-preserving methods such as federated learning and differential privacy for sensitive applications.
5) **M**ultimodal Fusion Strategies: We're investigating more sophisticated fusion strategies to better blend information from various modalities.

In summary, our research adds to the expanding field of affective computing by offering a solid framework for predicting both emotion and engagement from video content. This work holds significant promise for a range of applications, from enriching online learning experiences to enhancing human-computer interactions.

REFERENCES

[1] Zhang, W., Jia, R., Wang, H., Che, C., & Sun, H. (2024). A Self-Supervised Learning Network for Student Engagement Recognition From Facial Expressions. *IEEE Transactions on Circuits and Systems for Video Technology*.

[2] IEEE Xpert. (2023, December 19). Student Engagement Prediction using Image Processing [Video]. YouTube. https://www.youtube.com/watch?v=DCfHIBtMv4o

[3] Chen, D., et al. (2024). Decoding viewer emotions in video ads. *Scientific Reports*.

[4] Forasoft. (2024, September 9). What Is Video Emotion Analysis and How It Benefits Customer Service. https://www.forasoft.com/blog/article/video-emotion-analysis-customer-service

[5] An Estimation of Online Video User Engagement From Features of Human Facial Affect. (2022). *Frontiers in Computer Science*. https://doi.org/10.3389/fcomp.2022.7731546

[6] Shrivastava, A., Dubey, D., Verma, M., & Verma, H. (2024). Facial Emotion Recognition Using Video and Audio. *International Journal of Research Publication and Reviews*.

[7] Chaturvedi, I., Thapa, K., Cavallari, S., Cambria, E., & Welsch, R. E. (2021). Predicting video engagement using heterogeneous DeepWalk. *Neural Networks*, 142, 636–647.

[8] Moncaresi, I., et al. (2021). The Influence of Video Format on Engagement and Performance in Online Learning. *Scientific Reports*.

[9] Whitehill, J., Serpell, Z., Lin, Y. C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*.

[10] Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). DAiSEE: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.

[11] Li, Y. (2018). Human Emotion Recognition in Videos. *Uppsala University Thesis*.

[12] Nwokedi, S., & Gunes, H. (2024). Engagement Detection during E-Learning Classes using Machine Learning. *Journal of Propulsion Technology*, 45(2), 1172–1175.

[13] Philippe. (2022, September 14). Emotion detection with Python and OpenCV — Computer vision tutorial [Video]. YouTube.