Theory and Practice of Science and Technology
2025, VOL. 6, NO. 8, 208-213
DOI: 10.47297/taposatWSP2633-456938.20250608

Research on an Attention Detection Model in Speech Scenarios Based on Computer Vision

Zeming Hu

Wuhan Fiberhome Information Integration Technology Co., Ltd, Wuhan, Hubei, 430010 ,China

ABSTRACT
Online education and remote meetings make it difficult for speakers to gauge audience attention in real time; without direct eye contact and interactive feedback, communication effectiveness often suffers. Studies show that in face-to-face communication only about 7% of information comes from verbal content, 38% from voice tone, and as much as 55% comes from facial expressions and body language (nonverbal cues). Therefore, a vision-based model that can assess audience attentiveness via visual signals without requiring wearables is highly valuable. Building on prior work, this paper systematically reviews key background and methods, and proposes a temporal multi-modal model fusing facial expression and body posture features. We train and evaluate the model on public datasets such as DAiSEE and on self-recorded TED-style presentation videos. The fusion model achieves 86.3% accuracy and a macro-average F1 score of 0.84 in a four-class attention classification task, significantly outperforming a baseline model. Additionally, extensive experiments validate the model's robustness and applicability. Finally, we discuss potential application scenarios, ethical/privacy issues, and future development trends for attention detection technology in smart education and human–computer interaction.

KEYWORDS
Attention detection; Computer vision; Human–computer Interaction; Speech feedback; Multi-modal fusion; Pose estimation

1 Introduction

1.1 Background and motivation

Human communication relies not only on spoken language but also heavily on nonverbal cues such as eye contact, facial expressions, body posture, and gestures. Related research indicates that in face-to-face communication, over half of the conveyed information is delivered through nonverbal channels. With accelerating digitalization, classroom teaching, corporate training, and academic conferences are increasingly moving online, dramatically changing the interaction between speakers and audiences. Without in-person eye contact, a speaker cannot promptly discern whether the audience is understanding or drifting off; likewise, the audience cannot easily use facial feedback to influence the speaker's pace. This information asymmetry weakens communication effectiveness, leading to reduced learning efficiency and low meeting engagement.

To improve virtual interaction, researchers have proposed methods to monitor attention using EEG, eye trackers, pressure sensors, and other devices. However, these solutions are often costly, cumbersome, or invasive, making them unsuitable for large-scale use. In contrast, computer vision techniques can leverage ordinary cameras to capture facial expressions, head pose, and body movements and infer the user's focus level. Vision-based methods offer low cost and flexible deployment, opening up new possibilities in fields like education, healthcare, driver safety, and human–computer interaction.

1.2 Research challenges

Despite progress in vision-based attention detection, several challenges persist in live presentation scenarios: (1) Lack of data – few public datasets cover natural presentations, especially with multi-level attention labels; (2) Complex signals – no single visual cue fully captures an audience's attention state; (3) Generalization – models trained on one setting often struggle in others (different environments or audiences); (4) Privacy – continuous facial monitoring raises ethical concerns.

1.3 Article structure

The rest of the paper is organized as follows. Section 2 provides theoretical background and related work, including key datasets and existing methods. Section 3 details the data collection, preprocessing, feature engineering, and the design of our multi-modal fusion model. Section 4 presents the experimental setup, results (including baseline comparisons and ablation studies), and analysis. Section 5 explores applications of this research (in education and other domains) and discusses ethical considerations and future directions. Section 6 concludes the paper.

2 Theoretical background and related work

2.1 Attention and nonverbal cues

Nonverbal signals often convey more than spoken words in communication. For example, facial expressions, eye contact, and body posture account for a majority of information transfer in person-to-person interaction. However, in virtual presentation settings these cues are greatly reduced—cameras may only show the speaker's face, subtle expressions can be missed due to video quality, and many audience members turn off their video for privacy. Lacking such feedback cues, a speaker often cannot tell if listeners are engaged or not. This challenge motivates the development of computer vision systems to automatically monitor audience attentiveness in real time as a surrogate for direct observation.

2.2 Datasets

We used two public datasets for model development: DAiSEE (over 9,000 video clips with four engagement levels) and EmotiW (≈2,000 classroom clips labeled with low/medium/high engagement). We also collected a TED-style presentation dataset (~2 hours of video from 20 volunteer audience members) with each person's attention labeled at four levels (via combined self-reports and expert ratings) to evaluate the model in real presentation conditions.

2.3 Related work on attention detection

Early vision-based attention methods often focused on a single cue like gaze direction or blink rate and sometimes required specialized hardware (e.g. eye trackers). These approaches were straightforward but limited by equipment and controlled conditions. Recent methods fuse multiple cues (facial expressions, head pose, body movements, etc.) using deep learning. Techniques have evolved from rule-based algorithms to deep neural networks, inputs have expanded from a single modality to multi-modal fusion, and models have progressed from simple classifiers to temporal architectures (LSTMs, TCNs, Transformers), with steadily improving performance.

We leverage existing computer vision tools to extract features: MediaPipe FaceMesh for facial landmarks (468 3D landmarks) in real time on mobile devices, and OpenPose for multi-person pose estimation (body, hand, face keypoints) in a single image. These techniques help capture each audience member's facial expressions and body movements. For classification, we use the ensemble learning model XGBoost, a gradient-boosted tree algorithm that includes various optimizations (handling missing values, regularization, etc.) and is widely used as a baseline or feature fusion component in attention models.

3 Data and method

3.1 Data collection and labeling

3.1.1 Public data

We first utilized the DAiSEE dataset and the EmotiW student engagement subset as the primary training data. In preprocessing, we converted the raw videos into frame sequences (25–30 fps) and discarded segments where faces were missing or frames were blurry. Since most videos in DAiSEE feature a single individual, we additionally merged some videos into composite clips containing multiple people to simulate a group audience view of a presentation.

3.1.2 Self-recorded presentation data

We recorded a new dataset in a conference room setting. Twenty adult volunteers (ages 22–35, balanced gender) attended five TED-style talks as audience members. Multiple cameras (front, side, wide-angle) recorded each session to capture every audience member's face and posture. Each talk lasted ~15–20 minutes, during which audience members behaved naturally. After each talk, the audience self-rated their attention over time via a questionnaire, and three researchers independently watched the videos and rated each person's attention on a four-level scale. The average of the three expert ratings was used as the final attention label. Combining self and external evaluations helped reduce labeling bias.

3.2 Video sample construction

We segmented each recording into short temporal samples using a sliding window of 2 seconds with a 1-second overlap (50% overlap). This window length balances capturing short-term dynamics with increasing the number of

training samples. For each 2-second window, we used FaceMesh to detect facial landmarks and OpenPose to detect body keypoints in each frame, then associated each person's face features with their corresponding body features. If a frame had an occluded face or a keypoint detection failure, we interpolated from the previous frame's keypoints to maintain continuity of the feature sequence. Using this strategy, we obtained roughly 450,000 samples (≈89% from DAiSEE, ~50,000 from our data, and the rest from EmotiW).

3.3 Feature engineering

3.3.1 Facial expression features

Within each 2-second sample (~50–60 video frames per face), we extract several types of features from the 468 3D facial landmarks (detected by FaceMesh). Local motion features: we track key facial landmark points (around the eyebrows, eyes, mouth corners) across consecutive frames to measure dynamic movements like blinks, eyebrow raises, and mouth movements. Global shape features: we divide the facial landmarks into regions and apply Principal Component Analysis (PCA) to the coordinates in each frame, yielding a low-dimensional representation of the overall facial configuration (which reflects expressions of interest, boredom, etc.). Eye-gaze features: using the eye region landmarks, we estimate the viewer's gaze direction (horizontal and vertical angles); by considering camera geometry, we infer whether the person's eyes are directed toward the speaker.

3.3.2 Posture features

OpenPose provides 25 upper-body keypoints (head, neck, shoulders, torso, arms, etc.) for each person. From these, we derive: Head pose: the pitch, yaw, and roll angles of the head (computed from head and shoulder keypoints) to determine if the person is facing the speaker or looking away/down. Movement frequency: counts of nods, head shakes, or hand raises within the window (detected via keypoint angle changes). Frequent nodding generally indicates understanding, while repeated fidgeting or head-shaking may indicate confusion or distraction. Posture angle: the angle of the spine relative to the seat (and of the upper arms relative to the torso) to distinguish upright versus slouched sitting. An upright posture tends to correlate with attentiveness, whereas slouching suggests fatigue or low attention.

3.3.3 Temporal normalization

To incorporate temporal information, we concatenate each sample's facial feature sequence and posture feature sequence along the time dimension to form a combined feature matrix. We apply standard normalization (zero-mean, unit-variance) to these features and add positional encodings to indicate each time step's index, so that the model can learn the temporal order of events within the 2-second window.

3.4 Model architecture

3.4.1 Baseline model: XGBoost

As a baseline, we reduced each 2-second sample to a single feature vector (by averaging features over time) and performed four-class classification using XGBoost. We chose XGBoost because it handles missing values, captures non-linear relationships, and generally offers strong performance out-of-the-box】. We tuned hyperparameters via grid search: maximum tree depth = 8, learning rate = 0.1, subsample ratio = 0.8, and 500 boosting rounds. We also employed early stopping (monitoring validation F1_macro) to prevent overfitting.

3.4.2 Multi-modal fusion model

To better capture temporal interactions between facial and postural cues, we designed a multi-modal fusion model that combines Convolutional Neural Networks (CNNs) and a Transformer encoder:

· Feature encoding: Two parallel one-dimensional CNN streams encode the facial expression sequence and the posture sequence, respectively. Each stream has 3 convolutional layers (kernel size 5) with 64, 128, 256 filters, each followed by max pooling, to extract local temporal patterns from each modality.

· Modality fusion: The two encoded feature sequences (facial and posture) are concatenated along the time axis and fed through a fully connected layer to project them into a common feature space. This sequence is then input to a Transformer encoder (2 layers, 4 attention heads) which uses multi-head self-attention to learn long-range dependencies across and within modalities.

· Classification: The Transformer's output sequence (per time step features) is averaged over time to produce a single feature vector. This vector passes through a dense layer and a softmax to classify the attention level (four classes). We train

the fusion model using cross-entropy loss and the Adam optimizer (initial learning rate 1×10^−4, batch size 64).

3.4.3 Real-time inference and feedback

For real-time deployment, we implemented the model in a pipeline with separate threads for detection, evaluation, and visualization. The detection module continuously obtains facial landmarks and pose keypoints from the camera feed (via FaceMesh and OpenPose). The evaluation module computes the attention level for each 2-second window of data (processing windows in batches for efficiency). The visualization module displays each audience member's current attention level (e.g. as a bar indicator) and plots the history of attention over time for the group. This real-time dashboard allows an instructor or presenter to see when audience engagement is high or low and adjust accordingly.

4 Experiments and results analysis

4.1 Experimental setup

4.1.1 Data split

We split the combined DAiSEE and EmotiW data into training, validation, and test sets in a 60:20:20 ratio. The self-recorded presentation dataset was reserved entirely for testing model generalization. To mitigate class imbalance, we oversampled under-represented attention classes in the training set so that all four classes had roughly equal samples. In total, the training set contained ~270, 000 samples, and the validation and test sets ~90, 000 samples each, with an additional ~50,000 samples in the separate presentation test set.

4.1.2 Evaluation metrics

We report Accuracy, macro-average F1 score (F1_macro), Precision, and Recall to evaluate classification performance on four attention levels. To assess real-time capability, we also measured the average inference time per 2-second sample (including feature extraction and classification) and the frame processing rate (frames per second) for different numbers of audience members processed in parallel.

4.1.3 Experimental environment

All experiments ran on a laptop with an AMD Ryzen 7 5800H CPU and an NVIDIA RTX 3060 GPU. Our model was implemented in PyTorch 1.12, and we used MediaPipe and OpenPose Python libraries for feature extraction. To ensure real-time throughput, we employed asynchronous threading to parallelize the keypoint detection and model inference stages.

4.2 Baseline vs fusion model performance

Baseline vs. Fusion: On the DAiSEE test set, our fusion model achieves 86.3% accuracy (vs. ~76.3% for the XGBoost baseline), with a macro F1 of 0.84 (vs. ~0.72). On the TED-style presentation test set, it reaches 80.5% accuracy (vs. 70.1% for baseline), indicating strong generalization to new data. These performance gains reflect the fusion model's ability to capture richer temporal and cross-modal patterns, especially for distinguishing high vs. low attention states.

4.3 Ablation studies

To quantify the contribution of each component, we conducted ablation experiments:

· Single modality only: Training the fusion model using only facial features yielded 79.4% accuracy, and using only posture features gave 74.6% – both much lower than the full model's 86.3%. This confirms that facial and postural cues provide complementary information about attention.

· No Transformer: Replacing the Transformer encoder with a simpler fusion (concatenating CNN features followed by dense layers) reduced accuracy to 82.1%. This indicates that the Transformer's self-attention mechanism is effective at learning temporal dependencies and cross-modal interactions, significantly improving classification performance.

4.4 Error analysis

We analyzed common failure cases of the fusion model. First, the model sometimes confuses very subtle differences in attention (e. g. "High" vs. "Very High"), which is unsurprising since even human raters often disagree on such fine

distinctions. Second, individual differences can cause misclassifications: some people display very few outward signs even when fully attentive, while others might appear to face the speaker and maintain good posture despite losing focus. These factors limit the reliability of any one-size-fits-all behavior model and suggest that extremely fine-grained labels and a lack of personalization can impact accuracy. In the future, incorporating a short calibration phase per user or adaptive model components may help accommodate different behavioral patterns.

4.5 Real-time performance

We evaluated our system's speed under different audience sizes. With 1–5 people being tracked, the system processed approximately 30–40 frames per second (FPS); with 10–15 people, it ran at ~18–25 FPS (ensuring at least ~1 FPS per person). In a CPU-only mode, throughput dropped to ~8–12 FPS, so we recommend using a GPU for deployment. Overall, the system meets real-time requirements for small-to-medium audience sizes. (Our fusion model is also significantly faster than heavier video-based models, which makes it suitable for use on devices with limited computing resources).

4.6 Attention visualization

To illustrate the model's output, we analyzed a representative 30-second segment from our presentation dataset involving 6 audience members. The attention level curves showed that when the speaker asked an interactive question (~10 s into the talk), several audience members' attention levels spiked, whereas during a dense technical explanation (~20 s), some members' attention declined. Such real-time feedback could allow a presenter to adjust pacing or add interaction at key moments. In our interface, we also provided a playback visualization with each audience member's pose and facial landmarks overlaid, so that presenters in training can review precisely which parts of their talk succeeded in engaging the audience and where attention was lost.

4.7 Extended experiments and analysis

We further validated the model's reliability and generalization. In K-fold cross-validation, it showed stable performance (≈85.7% ± 0.5% accuracy across folds). The model also retained most of its accuracy under moderate video noise and partial occlusions, outperforming the baseline model in those stressed conditions. Finally, we observed that the model's accuracy was slightly lower for older individuals and for audiences from cultural backgrounds not seen in training. However, by fine-tuning the model on a small amount of target-domain data, these accuracy gaps were reduced significantly – indicating that the model can adapt across different populations with minor retraining.

5 Application exploration and discussion

5.1 Smart education and online classrooms

In educational settings, an automated attention monitor could alert instructors when student engagement drops so they can intervene (for instance by posing a question, showing a quick video, or initiating a discussion). Such a system should be used in compliance with student privacy laws and strictly as a teaching aid rather than for evaluating or penalizing students.

5.2 Presentation training and conference feedback

Public speaking trainees can use our system to get objective, real-time feedback during practice sessions. The audience's attention curve can reveal which parts of a presentation were engaging and where attention dropped, allowing the speaker to adjust their delivery or content in future sessions. Conference organizers could likewise monitor aggregate audience attention during events and schedule breaks or Q&A sessions when overall engagement wanes, thereby improving participation.

5.4 Ethical and privacy considerations

Continuous video-based attention tracking raises important privacy issues. The system should follow data minimization principles – only collecting information necessary for the task and processing it locally without storing raw video. Outputs should be anonymized (e. g. group attention statistics rather than individual data) and used only with informed user consent. In classroom deployments (especially with minors), strict privacy protections and regulations must be observed. Developers should also ensure transparency and fairness to avoid bias, keeping the technology's use

ethical.

5.5 Future directions

Future research should focus on collecting more diverse attention data (covering different presentation styles, languages, and cultures), integrating additional modalities such as speech audio and textual content to complement visual cues, and enabling quick personalization of models to individual behavior through calibration or fine-tuning. These efforts will make vision-based attention detection more robust, generalizable, and effective in real-world applications.

6 Conclusion and outlook

We proposed a vision-based model that fuses facial expression and body posture cues to detect audience attention during presentations. The model achieved high accuracy on both a public benchmark (86.3% on DAiSEE) and a new TED-talk dataset (~80.5%), significantly outperforming a traditional baseline and demonstrating strong generalization. Ablation experiments confirmed that combining facial and postural features substantially improves accuracy, and further analysis identified remaining challenges such as individual differences and occlusions. Overall, our model effectively recognizes audience attentiveness and could serve as a useful tool to enhance communication in online and offline presentations. In the future, collecting more diverse data and incorporating other modalities (e. g. speech) could make attention detection even more robust and applicable across domains, as long as fairness, transparency, and privacy are ensured.

References

[1] Gupta A., D'Cunha A., Awasthi K., et al. DAiSEE: Towards User Engagement Recognition in the Wild. arXiv:1609.01885, 2016.

[2] Kang H., Yang R., Song R., et al. An Approach of Query Audience's Attention in Virtual Speech. Sensors, 2024, 24(16): 5363.

[3] Mehrabian A., Ferris S. R. Inference of Attitudes from Nonverbal Communication in Two Channels. J. Consulting Psychology, 1967, 31(3): 248–252.

[4] Kartynnik Y., Ablavatski A., Grishchenko I., et al. Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs. In Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[5] Cao Z., Simon T., Wei S. E., et al. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017: 1302–1310.

[6] Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. In Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), 2016: 785–794.