

A Survey of AI Techniques for Meeting Summarization and Engagement Analysis in Video Conferencing

Assignment Report

Submitted by Name: Mohammed Omar Abdullah Alqathmi

Lecturer: Dr. Mohammed Anbar

Assignment Report: A Survey of AI Techniques for Meeting Summarization and Engagement Analysis in Video Conferencing

Course Title: Research Methodology in Computer Science (AI800)

Lecturer: Dr. Mohammed Anbar

Weighting: 20%

Hand-out Date: 25 October 2025

Hand-in Date: 15 November 2025

Hand-in Method: Online Submission via E-Learning

Submitted by Name: Mohammed Omar Abdullah Alqathmi

Email: 253001437@st.midocean.edu.km

Student Number: 253001437

CONTENTS

1. Introduction.....	4
2. Summary of Research Papers	4
2.1 Paper 1: Action-Item-Driven Summarization of Long Meeting Transcripts (Golia & Kalita, 2023)	4
2.2 Paper 2: Investigating Automatic Metrics on Meeting Summarization (Kirstein et al., 2024)	5
2.3 Paper 3: The Application of Real-time Emotion Recognition in Video Conferencing (Lin, 2024)	5
2.4 Paper 4: Automated Minutes of Meeting Using a Multimodal Approach (Singh et al., 2023)	6
2.5 Paper 5: Facial Expression Recognition using Visible and IR by Early Fusion (Multiple Authors, 2025)	6
2.6 Paper 6: Automatic Student Engagement Recognition in Online Classrooms (Multiple Authors, 2025)	7
3. Critical Analysis	7
4. Connections Between the Studies	8
5. Conflicts and Contradictions.....	8
6. Synthesis Matrix for Literature Review	9
7. References	11

1. INTRODUCTION

As a result of the rapid growth of online collaboration tools, meetings have increasingly moved into virtual spaces. This creates both opportunities and challenges. Platforms like Zoom, Google Meet, and Microsoft Teams have become standard. The volume of recorded meetings, discussions, and decisions is growing exponentially. Extracting meaningful information such as key points, action items, or engagement levels has become a critical task for research and enterprise applications alike.

This document reviews the recent papers on AI for meeting summarization, and engagement analysis. In this review of AI-driven Meeting Room over Video Conferencing, the goal is to comprehend the current methodologies, recognize limitations and summarize what will be the most promising paths for future work.

2. SUMMARY OF RESEARCH PAPERS

2.1 PAPER 1: ACTION-ITEM-DRIVEN SUMMARIZATION OF LONG MEETING TRANSCRIPTS (GOLIA & KALITA, 2023)

This paper focuses on extracting actionable content from condensed speech. The paper highlights that meetings are not dialogues, they are structured discussions with goals, decisions and responsibilities.

They propose a recursive summarization approach combined with an action-item extraction algorithm using BART transformers. By segmenting meeting transcripts into topic-based sections, the model reduces memory loss common in long-context models. The system was evaluated on the AMI corpus and achieved a **BERTScore of 64.98**, improving by nearly **5%** over existing baselines.

While effective, the model's reliance on synthetic meeting data limits generalization to real corporate scenarios. Moreover, processing multiple transcript sections increases computational cost, making large-scale deployment challenging.

🔗 <https://arxiv.org/abs/2312.17581>

2.2 PAPER 2: INVESTIGATING AUTOMATIC METRICS ON MEETING SUMMARIZATION (KIRSTEIN ET AL., 2024)

This research evaluates the reliability of metrics used to measure meeting summarization quality. The team applied PRISMA review methodology, developing an eight-category taxonomy of possible summarization errors—ranging from redundancy to hallucination. Using 35 annotated QMSum samples, they assessed five leading models (LED, DialogLED, PEGASUS-X, GPT-3.5 turbo, and Zephyr-7B-a) across nine metrics.

Results revealed that **one-third of metric-error combinations either ignored errors or rewarded them**. Metrics like Perplexity often favored grammatically fluent but semantically incorrect summaries. (Krippendorff's $\alpha \approx 0.79$), meaning Human correlation remained weak suggests evaluation benchmarks may not align with actual human judgment.

In spite of limited dataset and English-only scope, the study underscores a critical issue: the field may be optimizing for the wrong targets, mistaking surface-level coherence for comprehension.

 <https://arxiv.org/pdf/2404.11124.pdf>

2.3 PAPER 3: THE APPLICATION OF REAL-TIME EMOTION RECOGNITION IN VIDEO CONFERENCING (LIN, 2024)

Lin's paper explores how AI can improve the emotional quality of virtual meetings. The proposed framework is integrated with CNN-based facial feature extraction based on LSTM and TCN architectures to analyze emotion sequences over a period of time.

Edge computing handles almost all processing locally leading to significantly reduced latency and assured real-time responsiveness. Multi-modal fusion (combining facial, vocal, and physiological cues) enables enhanced emotion detection accuracy while Security is strengthened through homomorphic encryption and federated learning.

However, continuous emotion tracking raises significant privacy concerns. Differences in environmental conditions such as lighting, camera angle and even cultural expression differences reduce reliability. Despite these challenges, the study shows that emotion-aware conferencing can significantly improve engagement and communication quality.

 <https://www.suaspress.org/ojs/index.php/JCTAM/article/view/v1n4a10>

2.4 PAPER 4: AUTOMATED MINUTES OF MEETING USING A MULTIMODAL APPROACH (SINGH ET AL., 2023)

This work targets a practical business problem—automating meeting minutes generation. Using Whisper AI for speech-to-text transcription, the pipeline integrates NLP techniques for text normalization, topic extraction, and speaker diarization.

Results were impressive: **92% of key points and 87% of action items** were identified correctly, with a **Word Error Rate under 10%** for clear audio. The system produced minutes five times faster than humans.

This solution does not work well in noisy environments or when transcription contains domain-specific jargon and overlapping conversations. It however works well in structured, well-recorded meetings, limiting its flexibility in diverse conditions.



<https://www.ijraset.com/research-paper/automated-minutes-of-meeting-using-a-multimodal-approach>

2.5 PAPER 5: FACIAL EXPRESSION RECOGNITION USING VISIBLE AND IR BY EARLY FUSION (MULTIPLE AUTHORS, 2025)

This paper explores dual-modal facial expression recognition using both visible and infrared (IR) imaging. The idea is that visible light captures surface-level expressions, while infrared captures subtle temperature shifts linked to emotional states.

The authors enhance ResNet-18 with an attention mechanism and apply early fusion to combine both modalities. Tested on VIRI and NVIE datasets, the model achieved **84.44% and 85.20% accuracy**, outperforming prior methods by over 10%.

However, small datasets and hardware dependency (IR cameras) limit scalability. Performance drops result in occlusions (masks, glasses). poor lighting remains unresolved. In spite of these issues, the study points toward richer, more reliable emotion detection in hybrid environments.



<https://peerj.com/articles/cs-2676/>

2.6 PAPER 6: AUTOMATIC STUDENT ENGAGEMENT RECOGNITION IN ONLINE CLASSROOMS (MULTIPLE AUTHORS, 2025)

Focusing on virtual learning, this paper presents a model for detecting student engagement using multi-scale CNNs and attention fusion modules. Spatial interaction modeling captures relationships between the student, environment, and device.

The model achieved **88% accuracy** for highly engaged students and maintained 20–25 FPS in real time. The study demonstrates that engagement recognition can work effectively with quality video data.

Obstacles include a requirement for higher resolution videos ($\geq 480p$), limited scalability with larger classes, and concerns related to privacy because of continuous monitoring. Despite these drawbacks, this research shows how computer vision can bridge the observation gap in remote education.

 <https://www.iieta.org/journals/ts/paper/10.18280/ts.420428>

3. CRITICAL ANALYSIS

All the 6 papers highlight how AI is transforming meeting understanding. Starting from text-based summarization to emotional and behavioral analysis. Each paper addresses at least one real-world challenge but faces limitations related to dataset size, generalization, and computational demand.

Paper 1 and Paper 4 stand out for their applied focus—directly addressing business needs such as meeting documentation and action tracking. Paper 2 makes an essential contribution by questioning the validity of current evaluation metrics. Papers 3, 5, and 6 pushes the boundaries for what is possible in emotional intelligence and engagement detection but reveal major privacy and scalability trade-offs.

Across the board, transformers for text and CNNs with attention for vision dominate. Yet, data and ethical limitations remain the key bottlenecks to practical adoption.

4. CONNECTIONS BETWEEN THE STUDIES

A few strong themes connect these works:

- **Attention mechanisms** play an important role across modalities i.e transformers for summarization (Papers 1 & 2) and CNN attention modules for vision (Papers 5 & 6).
 - **Multimodal models** consistently outperforms single-source models. These models are able to benefit from a rich blended context including audio, text, and visual signals improving understanding.
 - **Temporal modeling** in Papers 1 and 3 shows the need to capture how conversations evolve over time.
 - **Application maturity** from prototype-level research to deployable systems shows the field is moving toward real-world usability.
-

5. CONFLICTS AND CONTRADICTIONS

While most studies align on core goals, several disagreements emerge:

- **Model design:** Paper 1 favors encoder–decoder setups, while Paper 2 suggests decoder-only models manage errors better.
 - **Metric reliability:** Papers 1 and 4 rely heavily on accuracy metrics, yet Paper 2 shows these metrics may not reflect human evaluation accurately.
 - **Real-time feasibility:** Paper 3's low-latency claims contrast with Paper 6's scalability issues, suggesting different assumptions about system load.
 - **Privacy approaches:** Paper 3 proposes encryption-based privacy, but others ignore it, highlighting inconsistent attention to ethics.
 - **Fusion strategy:** Paper 4 applies late fusion, whereas Paper 5 promotes early fusion, both claiming superiority without standardized comparison.
-

6. SYNTHESIS MATRIX FOR LITERATURE REVIEW

Paper	Year	Author(s)	Methodology	Key Findings	Weaknesses
Action-Item-Driven Summarization	2023 - 2024	Golia & Kalita	Recursive summarization; topic segmentation; BART on AMI corpus	BERTScore 64.98 (4.98% improvement); action-item focus effective; topic segmentation resolves dependencies	AMI corpus only; high computational cost; needs clear discussions
Investigating Automatic Metrics	2024	Kirstein et al.	PRISMA review; 8-category error taxonomy; 35 QMSum samples; 5 models; 9 metrics	~33% metrics mask errors; weak correlations; encoder-decoder vs decoder-only patterns; $\alpha=0.79-0.82$	English only; 35 samples; subjective annotations; limited diversity
Real-time Emotion Recognition	2024	Lin	CNN features; LSTM/TCN temporal; edge computing; encryption	Multi-modal beats single-modal; temporal tracking; low latency; detects disengagement	Privacy concerns; lighting/angle sensitivity; cultural differences; high computation

Automated Meeting Minutes	2023	Singh et al.	Whisper ASR; NLP preprocessing; ML topic extraction; speaker diarization	92% key points; 87% actions; <10% WER; 5x faster; 85% diarization	Poor audio degrades; needs domain training; struggles with context; high resources
Facial Expression Recognition	2025	Multiple authors	ResNet-18 attention; visible+IR early fusion; VIRI/NVIE	84.44% VIRI (12.14% up); 85.20% NVIE; attention crucial (+3-6%)	Small datasets; overfitting; occlusions/lighting issues; IR cost; speed needs optimization
Student Engagement Recognition	2025	Multiple authors	Variable kernel (3×3,5×5,7×7); attention fusion; spatial interaction; segmentation	88% highly engaged; 82% moderate; 85% disengaged; 20-25 FPS; spatial +6%	Needs ≥480p/15FPS; off-camera fails; cultural variation; large-scale prohibitive; privacy issues

7. REFERENCES

1. Golia, L., & Kalita, J. (2023). Action-Item-Driven Summarization of Long Meeting Transcripts. *arXiv preprint arXiv:2312.17581*.
<https://arxiv.org/abs/2312.17581>
 2. Kirstein, F., Wahle, J. P., Ruas, T., & Gipp, B. (2024). Investigating Automatic Metrics on Meeting Summarization. *arXiv preprint arXiv:2404.11124*.
<https://arxiv.org/pdf/2404.11124.pdf>
 3. Lin, W. (2024). The Application of Real-time Emotion Recognition in Video Conferencing. *Journal of Computer Technology and Applied Mathematics*, 1(4), 79-88. <https://www.suaspress.org/ojs/index.php/JCTAM/article/view/v1n4a10>
 4. Singh, A., et al. (2023). Automated Minutes of Meeting Using a Multimodal Approach. *International Journal for Research in Applied Science and Engineering Technology*, 11(12).
<https://www.ijraset.com/research-paper/automated-minutes-of-meeting-using-a-multimodal-approach>
 5. Multiple authors (2025). Facial Expression Recognition using Visible and IR by Early Fusion of Deep Learning with Attention Mechanism. *PeerJ Computer Science*, e-2676. <https://peerj.com/articles/cs-2676/>
 6. Multiple authors (2025). Automatic Student Engagement Recognition in Online Classrooms Based on Deep Image Segmentation and Attention Mechanisms. *Traitement du Signal*, 42(4).
<https://www.iieta.org/journals/ts/paper/10.18280/ts.420428>
-