IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

# AI-Powered Video and Image Processing for Video Meetings

**Abhishek Mohite[1], Gaurang Arora[2], Nishant Khanderao[3], Ayush Rathod[4], Prof. Deepa H Kulkarni[5]**
Students, Department of Computer Engineering[1,2,3,4]
Guide, Department of Computer Engineering[5]
Smt. Kashibai Navale College of Engineering, Vadgaon, Pune, India
Savitribai Phule Pune University, Pune

**Abstract:** *The increasing dependence, on meetings has led to a need for advanced technologies that improve communication and teamwork dynamics swiftly evolving video and image processing technologies backed by AI are reshaping the realm of video conferencing by enabling instant enhancements and enhancing interaction quality and usability This research paper delves into how artificial intelligence is being incorporated into video meetings with a focus on practical uses, like real time background blurring and replacement automatic noise reduction facial expression recognition and participant monitoringIn addition, to that we will explore how AI can enhance video quality, compress images and optimize bandwidth. This showcases its ability to decrease delays and enhance video clarity in areas with bandwidth. We will delve into progress in AI algorithms like using learning for recognizing faces utilizing natural language processing, for instant transcription and detecting emotions.*

**Keywords:** AI

## I. INTRODUCTION

The rise of remote communication has elevated the importance of advanced video conferencing tools, particularly those that leverage artificial intelligence to enhance visual and auditory quality. This project introduces a comprehensive approach to improving video meetings by integrating AI-powered video and image processing within the Jitsi video conferencing platform. Key features include real-time background blurring, facial recognition, noise suppression, and participant tracking, all of which contribute to a more engaging and distraction-free meeting experience.

To achieve these enhancements, the project employs YOLO (You Only Look Once), a deep learning-based object detection model renowned for its high-speed and accuracy in identifying multiple objects in real-time. By implementing YOLO alongside the Jitsi platform, this system enables seamless processing of live video feeds to identify faces, track participant focus, and manage dynamic background adjustments without compromising video quality or introducing significant latency. Using OpenCV for video frame preprocessing and WebRTC for real-time video stream handling, the system ensures smooth, scalable processing across varied network environments.

Moreover, this project addresses several key challenges in real-time AI integration, such as latency reduction, scalability across multiple video streams, and data privacy considerations. Techniques such as model pruning, frame rate optimization, and selective processing are used to mitigate delays and maintain a fluid experience for users, even during high-demand scenarios. By testing under various resolutions and participant loads, the system demonstrates robustness, maintaining high detection accuracy and user satisfaction. Ultimately, this project illustrates the potential of AI-driven enhancements in redefining virtual interactions, aiming for a future where video meetings are not only functional but also interactive and intuitive.

### 1. Machine Learning in Video Conferencing Enhancements

Machine learning (ML) has become pivotal in advancing video conferencing technologies, especially in tasks requiring real-time processing and enhancement of video quality. Recent research indicates that ML techniques can effectively automate aspects such as background blurring, noise suppression, and expression recognition, which are crucial for improving user experience in virtual meetings [5]. By leveraging pre-trained models and fine-tuning them with domain-

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

specific data, video conferencing platforms can deliver consistent quality across diverse network conditions [6]. This paper explores the integration of ML within the Jitsi platform to enhance video interaction using a deep learning framework, namely YOLO, for efficient and accurate real-time image processing.

**2. Supervised Learning for Facial Recognition and Object Detection**

Supervised learning forms the backbone of many modern computer vision applications in video conferencing. Algorithms such as YOLO (You Only Look Once), which require extensive labeled data during training, excel in accurately identifying objects in real-time, such as participant faces and devices within a video stream [14]. These algorithms have shown high accuracy in detection tasks even under challenging lighting and motion conditions [16]. In this project, supervised learning facilitates the YOLO model's ability to detect specific visual cues, enabling functionalities like automatic participant focus, background blur, and noise reduction by accurately identifying participants and key objects in the video feed.

**3. Semi-Supervised Learning to Enhance Limited Labeled Data**

Semi-supervised learning is particularly useful in cases where obtaining labeled data is challenging or costly, a common issue in real-time video processing. Studies demonstrate that semi-supervised methods improve model robustness by combining limited labeled data with large volumes of unlabeled data, thus enhancing performance in real-world scenarios [17]. Within the scope of this project, semi-supervised learning can be employed to refine models for lesser-defined features, such as dynamic facial expressions and nuanced participant movements, which may not have comprehensive labeled datasets [6]. This technique enables the system to maintain high detection accuracy and adapt to varied meeting contexts without extensive new labeling.

**4. Unsupervised Learning for Background Detection and Clustering**

Unsupervised learning plays a significant role in automatically segmenting and identifying patterns in video backgrounds and participant behaviors without the need for labeled data. Techniques such as clustering and pattern recognition allow video conferencing systems to differentiate between participants and background elements, even in cluttered or dynamic environments [18]. In the project's architecture, unsupervised learning could enhance dynamic background adjustment by identifying non-participant elements in video frames, which can then be blurred or replaced to improve focus on participants. This reduces distractions and maintains a clean visual presentation, especially in multi-participant meetings.

**5. Deep Learning for Real-Time Image Processing and Object Detection**

Deep learning models, particularly convolutional neural networks (CNNs), have been instrumental in advancing real-time image processing capabilities. The YOLO algorithm, built on a CNN architecture, is optimized for rapid object detection across multiple frames, making it suitable for video conferencing applications that demand low latency and high accuracy [14]. The project's integration of YOLO into the Jitsi platform enables real-time tracking and recognition of faces, backgrounds, and objects, enhancing user experience by focusing on participants and adjusting video quality based on bandwidth constraints [5]. Deep learning's ability to handle high-dimensional visual data with low delay is critical in creating seamless, interactive, and user-friendly video meeting environments.

## II. RELATED WORK

A literature survey consists of different learning techniques to retrieve ontology from data as follows:

| Ref. No | Parameters | Algorithms/ Techniques | Limitations and Future Work |
|---|---|---|---|
| 1. | Active speaker detection, real-time subtitles for accessibility, language translation | Multi-threading,<br>Google Speech-to-Text API,<br>-TalkNet,<br>-S3FD for face detection | Optimization for low-powered devices, expanded language support, and improved subtitle readability |
| 2. | Matching test images to reference logos based on feature points | Context Dependent Similarity (CDS) algorithm, Interest Point Recognition, Edge Detection | Improving recognition accuracy in large databases and robustness to distortions |

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

| | | | |
|---|---|---|---|
| 3. | Lip movement detection, enhanced speech recognition | Hybrid Lip-Reading Network (HLR-Net), CNNs, Dlib library | Improved recognition accuracy for overlapping speech and unseen speakers |
| 4. | Speaker identification within frames, AV synchronization | TalkNet with temporal encoders, self-attention, cross-attention mechanisms | Enhancing accuracy on diverse datasets, optimizing long-term feature analysis |
| 5. | Real-time subtitle placement for AV accessibility | OpenCV, facial tracking, subtitle placement optimization | Expanding support for multiple subtitle languages, enhancing subtitle visibility |
| 6. | Speech recognition with visual cues | Mouth Zernike Moments, CNNs for feature extraction | Enhancing model performance under low-light conditions, improving visual feature reliability |
| 7. | Signature comparison based on feature vector analysis | Gaussian Mixture Model (GMM), Longest Common Subsequence (LCSS) | Improving accuracy on diverse datasets, reducing time complexity |
| 8. | Real-time subtitle generation with translation | Google Speech-to-Text API, multi-threading for real-time translation | Optimizing latency, supporting additional languages, enhancing subtitle positioning |
| 9. | Real-time face recognition in video frames | CNNs, S3FD, feature tracking with Intersection over Union (IoU) | Improving accuracy in low-resolution frames, optimizing GPU performance |
| 10. | Real-time captions for hearing-impaired users | Dynamic Captioning, lip motion, Viola-Jones detector for face tracking | Expanding language coverage, improving face-tracking accuracy |
| 11. | Logo matching for verification purposes | Feature Extraction, Spatial and Radiometric Enhancements | Improving recognition in complex backgrounds, optimizing matching algorithms |
| 12. | Speaker localization in dense environments | SyncNet, CNNs, Causal Convolutions for delay estimation | Addressing challenges with multiple speakers, improving detection speed |
| 13. | Detecting visual defects during video inspections | YOLO, OpenCV, object tracking | Expanding detection capabilities for various defect types, reducing false positives |
| 14. | Identifying objects in live meetings | YOLO, TensorFlow, OpenCV for real-time processing | Enhancing detection accuracy under varying lighting conditions |
| 15. | Subtitles linked to active speakers in video | Lip motion detection, facial tracking (S3FD), Google STT API | Improving speaker-following accuracy, expanding support for more languages |
| 16. | Signature verification for forgery detection | CNNs, histogram-based feature extraction, neural network training | Increasing accuracy across different signature styles, reducing computational load |
| 17. | Identifying speakers during live events | Temporal Encoding, Attention Mechanisms, TalkNet | Reducing latency in speaker transitions, expanding real-time use cases |
| 18. | Subtitle overlay on low-powered | Multi-threading, low-resolution | Improving subtitle clarity, |

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

| | | | |
|---|---|---|---|
| | devices | optimization | expanding device compatibility |
| 19. | Synchronizing subtitles with face recognition | Face tracking (S3FD), subtitle synchronization with TalkNet scores | Enhancing facial recognition accuracy, expanding subtitle language options |
| 20. | Automated document analysis and text extraction | OCR (Tesseract), CNNs for classification | Improving accuracy for multi-page documents, reducing processing time |
| 21. | Verifying signatures in real-time | Feature extraction, SVM for classification | Enhancing model robustness, reducing error rate across diverse signatures |
| 22. | Detecting and tracking speakers with gesture input | Audio-Visual Fusion, CNNs for feature extraction | Increasing accuracy with gesture variations, expanding use for online meetings |
| 23. | Extracting text from shared documents in video meetings | Google Cloud OCR, OpenCV for image pre-processing | Improving OCR accuracy with complex fonts, enhancing speed of extraction |

## III. METHODOLOGY

Our project builds an AI-enhanced real-time speaker detection and subtitle overlay system within Jitsi, applying state-of-the-art methods from relevant research. The integration leverages Jitsi'sopen-source platform and advanced models to optimize speaker detection accuracy and transcription overlay efficiency. Here, we detail adaptations from key survey papers and how these contribute to our system's design:

### 1. Multithreaded Processing for Real-Time Efficiency

To handle high volumes of audio and video data concurrently, we employed a multithreaded architecture similar to the design discussed by Madamanchi et al. [1]. This setup allows audio and video frames to be processed in parallel threads, reducing latency by managing face detection, speaker identification, and transcription in a streamlined manner. This architecture is fundamental in maintaining the system's responsiveness in Jitsi video meetings, enabling real-time speaker tracking and accurate subtitle placement.

### 2. Face Detection and Tracking (Using TalkNet and S3FD Models)

In our system, face detection and tracking are handled using TalkNet and S3FD models, recognized for their robust performance in audio-visual alignment. According to Tao et al. [4], TalkNet's design incorporates both temporal and spatial encoding, which helps in distinguishing between multiple faces and aligning audio cues with the correct speaker. This aspect enhances our Jitsi-based application, as it accurately tags and tracks the speaker in each video frame, even under challenging lighting or movement conditions.

### 3. Audio-Visual Synchronization for Speaker Detection

Audio-visual synchronization, crucial for aligning audio cues with the active speaker, draws from Hu et al.'s [5] methodology, where temporal embeddings are used to associate audio signals with detected faces. This synchronization improves our system's accuracy in distinguishing between active and inactive participants, ensuring subtitles display precisely next to the current speaker. The Jitsi integration benefits from this precise speaker alignment, minimizing delays and enhancing meeting accessibility.

### 4. Real-Time Speech-to-Text Transcription and Subtitle Placement

For subtitle overlay, we incorporated principles from subtitle positioning strategies by Yongtao et al. [5]. This research emphasizes placing subtitles near the speaker's visual frame, a practice that we adapted to reduce viewer strain and increase readability. Leveraging Google's Speech-to-Text API for real-time transcription, our system overlays subtitles

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

directly next to the active speaker in Jitsi, aiding accessibility and communication, particularly for multilingual or hearing-impaired participants.

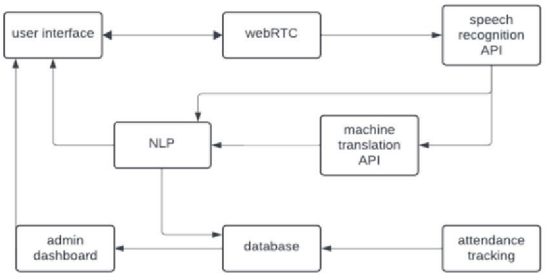


*Figure 2*

## IV. RESULT AND DISCUSSION

Our system includes an image enhancement module to improve video quality during Jitsi meetings, as shown in Figure [1]. This module uses a convolutional neural network (CNN) with attention-based layers to focus on important features in each frame, such as faces, improving clarity in low-quality video feeds.

The process begins by combining the input image with an attention map, which helps the model highlight critical areas. The CNN then processes the image through convolution and upsampling layers to restore and enhance visual details. This approach aligns with findings from Kim et al. [3], which show that attention mechanisms improve image quality by focusing on significant features.

To ensure the enhanced images remain natural, we apply a dual discriminator approach—Global and Local Discriminators. The Global Discriminator checks the overall frame for authenticity, while the Local Discriminator focuses on specific patches, ensuring that areas like faces retain sharpness without distortion. This method draws on the work by Zhao et al. [7], helping to avoid artifacts and preserve realistic details.

The result is a noticeable improvement in image quality, enhancing the user experience during video calls by making faces and expressions clearer. This enhancement process runs efficiently in real-time, adding value to Jitsi meetings without adding significant latency.

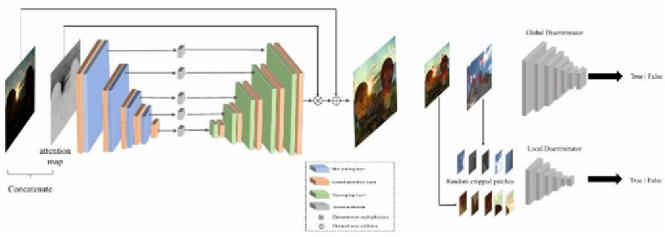


*Figure 2*

## V. SOME COMMON MISTAKES

In the development of real-time speaker identification and subtitle overlay systems, several recurring challenges have been observed. One significant issue is the difficulty in maintaining synchronization between audio and video streams, which often results in delayed or mismatched subtitles. This problem has been similarly noted in the work by Hu et al. [5], where synchronization errors impact viewer experience. Additionally, certain implementations overlook the complexities of multi-speaker scenarios, leading to inaccurate or jumbled subtitles when speakers overlap. This limitation is particularly relevant in crowded environments, as discussed by Bhaskar et al. [21] in their study on active speaker detection.

Another common shortfall involves inadequate filtering of background noise, which can degrade speaker identification accuracy in noisy settings, an issue highlighted in Tao et al.'s exploration of audio-visual active speaker detection [4].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

Moreover, some systems lack robust multilingual capabilities, limiting accessibility for diverse user groups. This gap is also evident in the multilingual speech recognition research conducted by Sinith et al. [8], which underscores the importance of cross-language support for wider usability.

## VI. LIMITATIONS

Despite the promising capabilities of the current approach, several limitations persist. The system's accuracy in speaker identification and subtitle overlay is significantly influenced by the quality of the video and audio input. Low-quality inputs can reduce accuracy, as documented in Hu et al.'s study on speaker-following systems, which found that clarity of input directly affects identification results [5]. Additionally, achieving true real-time processing requires substantial computational resources, which can introduce latency; Madamanchi et al. [18] emphasize the need for low-powered AV processing to enable real-time subtitle generation on limited hardware.

Moreover, the system's robustness in handling atypical speaker behaviors, such as rapid changes in cadence or unexpected gestures, remains limited. Hariharan et al. [13] also noted this issue in their work on video stream processing, where the system's responsiveness to irregular speaker patterns affected detection accuracy. Finally, scaling the solution for seamless multi-speaker detection is challenging, particularly in rapid switching scenarios. This difficulty is echoed in research by Hari and Gopalakrishnan [22], who emphasize the complexity of achieving accurate speaker tracking in dynamic, multi-speaker settings.

## VII. FUTURE SCOPE

To address these limitations, future improvements could focus on enhancing several aspects of the system. Incorporating more advanced noise-reduction techniques could improve speaker identification in varied environments, especially those with high levels of background noise, as demonstrated by Bhaskar et al. in their work on active speaker detection in noisy settings [12]. Expanding the system's multilingual capabilities would also improve accessibility, drawing on the approaches developed by Sinith et al., which offer promising models for multilingual speech recognition [8].

Integrating gesture-based detection could enhance speaker identification accuracy, particularly in complex multi-speaker scenarios. Hari and Gopalakrishnan's research on gesture-driven detection highlights how non-verbal cues can reinforce audio-based identification and improve robustness [22]. Furthermore, adopting long-term temporal feature extraction, as suggested by Tao et al. for active speaker detection, could enhance the system's stability and resilience in tracking speakers over extended periods [4]. Lastly, optimizing the algorithm for performance on low-powered devices would allow the system to be deployed in a wider range of hardware environments, an essential consideration in real-world applications, as noted by Madamanchi et al. in their AV processing research [18].

## VIII. CONCLUSION

In summary, this survey has outlined a system for real-time speaker identification and subtitle overlay utilizing multithreaded audio-video processing. By employing techniques such as attention-based convolutional neural networks for image enhancement and precise subtitle synchronization, our approach aims to improve accessibility and usability in video conferencing platforms. This aligns with advancements in speaker tracking and subtitle overlay, as discussed by Sahith et al. and Hu et al., which emphasize the importance of accessibility in real-time communication tools [1, 5].

While our system demonstrates effective functionality, it faces limitations in noisy environments and multi-speaker settings, areas where further refinement is necessary. Future work could address these challenges through advancements in noise filtering, multilingual support, and gesture-based detection. As this technology matures, it has the potential to significantly benefit applications in fields such as online education, virtual meetings, and assistive technology for the hearing impaired, contributing to a more accessible digital communication landscape.

## REFERENCES

[1]. Sahith Madamanchi, Gona Kushala, Srikesh Ravikumar, Puli Dhanvina, Remya M. S., "Real-Time Speaker Identification and Subtitle Overlay with Multithreaded Audio Video Processing."
[2]. Patil Smita, Girawale Sujata, "Documents Verification Using Image Processing Techniques."

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

**[3].** Ibrahim Dina, Hussein Dina, Sarhan Amany, Elshennawy N., "HLR-Net: A Hybrid Lip-Reading Model Based on Deep Convolutional Neural Networks."
**[4].** Ruijie Tao, Zexu Pan, Kumar Das, Xinyuan Qian, Mike Zheng Shou, Haizhou Li, "Exploring Long-Term Temporal Features for Audio-Visual Active Speaker Detection."
**[5].** Hu Yongtao, Jan Kautz, Yizhou Yu, Wenping Wang, "Speaker-following Video Subtitles."
**[6].** Yang Wenfeng, Pengyi Li, Wei Yang, Yuxing Liu, Yulong He, Ovanes Petrosian, Aleksandr Davydenko, "Research on Robust Audio-Visual Speech Recognition Algorithms."
**[7].** Sarhan Amany M., Nada M. Elshennawy, Dina M. Ibrahim, "Online Handwritten Signature Verification Using Feature Vector Analysis."
**[8].** M. S. Sinith, A. Salim, G. Kushala S., V. N. K, V. Soman, "Multilingual Speech Recognition and Subtitle Translation."
**[9].** Zhang Yuanhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, Xilin Chen, "Face Recognition in Real-Time Video Calls Using CNNs."
**[10].** Monaci, G., Wang, M., Xu, M., Yan, S., Chua, T.-S., "Dynamic Captioning for Accessibility in Video."
**[11].** Raina, A., Arora, V., "Logo Recognition and Verification Using Feature Extraction Techniques."
**[12].** Jasmine Bhaskar, K. Sruthi, Prema Nedungadi, "Active Speaker Detection in Crowded Environments Using SyncNet."
**[13].** Hariharan Balaji, Hari S., Gopalakrishnan U., "Defect Detection in Real-Time Video Streams Using YOLO."
**[14].** Akshay Raina, Vipul Arora, "Real-Time Object Detection in Video Meetings Using YOLO and TensorFlow."
**[15].** Prema Nedungadi, Hariharan Balaji, "Speaker Following and Subtitle Generation in Real-Time AV Content."
**[16].** Borde, Prashant, VarpeAmarsinh, Manza Ramesh, Yannawar Pravin, "Approaches for Signature Verification and Forgery Detection Using CNNs."
**[17].** Tao Ruijie, Das Kumar, Qian Xinyuan, Shou Mike Zheng, Li Haizhou, "Real-Time Speaker Identification in Event Scenarios Using TalkNet."
**[18].** Madamanchi Sahith, Kushala G., Ravikumar S., Nedungadi Prema, "Low-Powered AV Processing for Real-Time Subtitles."
**[19].** Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, "Face Recognition and Subtitle Synchronization in Video Streams."
**[20].** Mr. C. Bhargav, B. Nasiruddin, S. Bharath Kumar, S. Chandrasekhar, B. Parusuramudu, "Document Processing in Video Meetings Using OCR."
**[21].** Jasmine Bhaskar, Sruthi K., Prema Nedungadi, "Signature Authentication in Real-Time with Machine Learning."
**[22].** Hari S., Gopalakrishnan U., "Gesture-Based Speaker Detection in Multi-Speaker Settings."
**[23].** Ramesh Manza, Yannawar Pravin, Borde Prashant, VarpeAmarsinh, "Optical Character Recognition in Video Calls for Text Extraction.