IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131104

Context-Based Audio-Video Analysis

Prof. S. P. Gade[1], Dnyaneshwari Mahadev Sawant[2], Neha Sanjay Patil[3], Sujal Balasaheb Mane[4], Payal Yogiraj Sable[5]

Associate Professor, PDEA's College of Engineering, Manjari (BK), Pune, Maharashtra, India[1]

Department of Computer Engineering, PDEA's College of Engineering Manjari (BK), Pune, Maharashtra, India[2-5]

Abstract: A video carries two streams of evidence, and each one misses things the other catches. The picture can show a person moving but cannot tell us whether we are hearing footsteps, a siren or rain on a roof. The sound can tell us something happened but often not what produced it. This paper proposes a context-based audio-video analysis system that takes a video as input, separates its audio and visual content, analyzes both, and fuses the results to work out what is going on. The output is a natural-language caption. If the frames show rainfall and the soundtrack contains falling raindrops, for example, the system could produce "It is raining outdoors." The design accounts for background noise, overlapping activities, audio-video synchronization and uneven video quality. Possible uses include surveillance, accessibility, activity recognition and video summarization. The system has not been built yet, so its accuracy, reliability and processing speed still have to be measured.

Keywords: audio-video analysis, multimodal learning, context understanding, activity recognition, audio-visual fusion, event detection, automatic caption generation.

I. INTRODUCTION

People record video to document daily routines, public events, conversations and changes in the weather. Making sense of such a recording takes more than naming the objects in the frames or the sounds on the track. The two have to be read against each other. Suppose a clip shows someone moving quickly while the audio contains a steady beat of footsteps. Taken together, the likely reading is that the person is running. Rain works the same way: frames showing rainfall plus the sound of drops hitting the ground give a firmer conclusion than either source gives alone.

Work on multimodal learning, audio-visual speech recognition, activity recognition, event localization and video-language understanding gives us a base to build on [1]–[10]. The system proposed here draws on that work to identify activities and events in a video and to describe the overall situation in a caption.

A. Problem Statement

Many existing approaches handle audio and video as separate problems. Others merge them without really modeling how a sound relates to what is happening on screen. Either way, the description can end up incomplete or wrong, especially when several things happen at once or when noise covers the sound that matters. The problem addressed here is to build a multimodal system that analyzes both streams, combines the relevant information, identifies the activities and events, and writes a caption that reflects the context.

B. Objectives

- Accept a video as input and extract its audio and visual content.
- Identify activities, environmental sounds and events from what is extracted.
- Combine audio and visual features so the system understands context better than either stream allows.
- Link detected sounds to the visual activities or events that likely caused them.
- Generate clear captions that describe the context of the video.
- Evaluate caption relevance, event recognition, processing time and behavior under difficult conditions.

II. LITERATURE SURVEY

Several studies have looked at how to combine information from different sources to interpret events in video. Zhang proposed a dynamic sparse attention mechanism for multimodal video understanding, aimed at picking out relevant features and keeping processing efficient [1]. Liu et al. studied human-inspired computing for audio-visual speech recognition and showed what is gained by processing sound and lip movement together [2]. Fabian et al. compared classical machine learning algorithms for multimodal human activity recognition, which shows that conventional

 This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131104

methods still have a place when several data sources are available [3]. Li et al. used video tubelets for ultrasound video analysis, paying attention to spatial context and to how frames relate over time [4].

Lin and Harte asked how much the visual channel actually contributes to audio-visual speech recognition, a question that matters for any system that must decide how heavily to weight each modality [5]. Zhuo et al. presented SurgRAW, a multi-agent workflow using chain-of-thought reasoning for robotic surgical video analysis [6]. Xu et al. proposed a cross-modal feature enhancement network for audio-visual event localization, which tries to strengthen the link between what is heard and what is seen [7]. Taken together, these papers offer methods for feature extraction, temporal analysis and audio-visual combination that apply directly to event recognition.

Yuan et al. examined surveillance video-and-language understanding with small and large multimodal models, where language is the means of interpreting what a video shows [8]. Shaikh et al. introduced MAiVAR-T, a transformer-based recognizer that uses audio-image and video inputs for action recognition [9]. Chen et al. released a dataset and baseline for surgical video action recognition, targeting the difficulty of recognizing actions across long, complicated sequences [10]. Across all of these, the common threads are multimodal learning, feature fusion, reasoning over time and the use of language to describe video.
These studies each handle one part of the problem. The present project joins audio processing, visual analysis, event recognition and caption generation into a single workflow. A video goes in, the two modalities are analyzed, their relationship is examined, the event is identified, and a caption comes out. Whether this produces better descriptions than single-modality methods can only be settled by building the system and testing it.

A. Research Gap

Prior work covers recognition within one modality, feature fusion, temporal action analysis, event localization and video-language interpretation. What is harder to find is a practical pipeline that connects all of these steps from raw video to finished caption, and that is an engineering problem more than a theoretical one. Audio-only analysis may not reveal what is making a sound, and video-only analysis can miss acoustic evidence entirely. Noise, simultaneous activities, timing mismatches and poor video quality further reduce reliability. This project targets the integration gap by treating sound recognition, action recognition and captioning as connected steps rather than isolated tasks.

B. Proposed Contribution

We propose a Context-Based Audio-Video Analysis system. It accepts a video, processes the audio and visual components, and identifies the activities and events in them. Audio features and video-frame information are extracted separately, analyzed, and then combined to give a fuller picture of the scene. The resulting context drives a natural-language caption. Frames of rainfall paired with the sound of raindrops, for instance, should lead to a caption about a rainy outdoor scene. The workflow is meant to show whether using both modalities makes captions more informative than using one. Evaluation will cover event-recognition quality, caption relevance, processing time and performance under varied audio and video conditions.

III. SYSTEM OVERVIEW AND ARCHITECTURE

The system takes a video file and analyzes its audio and visual parts in parallel. It merges the extracted features, decides what event is taking place, and shows the user a caption. Five stages make up the architecture: video input and preprocessing, audio analysis, visual analysis, audio-video fusion with event recognition, and caption generation with display.

A. System Modules

Module	Responsibility
Video Input and Preprocessing	Accepts the video, extracts the audio, samples frames and prepares both for analysis.
Audio analysis	Identifies speech, footsteps, rainfall and other environmental sounds.
Visual analysis	Detects visible objects, actions and scene information in the frames.
Audio-video feature fusion	Combines audio and visual features and aligns related events in time.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131104

Context recognition	Picks the most likely activity or event from the combined evidence.
Caption generation	Turns the recognized context into a natural-language caption and shows it to the user.

B. Workflow

1. Video upload. The user selects a video file.
2. Preprocessing. The system extracts the audio, samples frames and prepares the data.
3. Audio analysis. An audio model identifies the environmental sounds or speech it supports.
4. Visual analysis. A video or image model identifies visible objects, actions and scene information.
5. Feature fusion. Audio and visual features are combined, with their timing relationship taken into account.
6. Context identification. The fused information is used to choose the most likely event.
7. Caption generation. The caption module converts that event into a sentence.
8. Output. The caption is displayed to the user.

As an example, a video with visible rainfall and matching rain sounds might yield "Rain is falling outdoors." This is an illustration, not a measured result. When the audio and video disagree, the system should not make a confident claim that the evidence does not support.

IV. METHODOLOGY

A. Video Input and Preprocessing

The system reads a video file and pulls out what the later stages need. Frames can be sampled at intervals to avoid unnecessary computation, and the audio track is extracted and converted to a suitable format. Preprocessing may involve resizing frames, normalizing inputs, handling videos with no audio, and checking duration and format. The exact sampling and preprocessing settings will be fixed during implementation.

B. Audio Feature Extraction and Analysis

The audio branch analyzes the sound in the video. Depending on the model chosen, it may recognize environmental sounds, speech or other acoustic events. The audio can be represented as a spectrogram or as learned embeddings. Because background noise can degrade recognition, the chosen model needs to be tested on noisy recordings.

C. Visual Feature Extraction and Analysis

The visual branch looks at sampled frames or short segments to find visible objects, activities and scene characteristics. A suitable image or video model may recognize walking, running or dancing, provided those activities appear in its training data. A single frame often says too little to identify an activity, so temporal information has to be used.

D. Audio-Video Feature Fusion

The audio and visual outputs are merged into one representation of the event. Two approaches are practical, and the choice depends on the models used and the form of their outputs.

1. Feature-level fusion combines audio and visual representations before the event is classified.
2. Decision-level fusion combines the predictions or confidence scores of separate audio and video models.

Decision-level fusion is probably easier for a first prototype, since separate model predictions are usually available. Feature-level fusion becomes an option if compatible embeddings can be obtained. Whichever is used, the fusion stage must allow for timing differences and must cope with a modality that is missing or gives only weak evidence.

E. Context Recognition and Caption Generation

The fused information determines the likely event and its context, and a caption component turns that into language. Example outputs include "A person is dancing to music," "It is raining outdoors" and "A person is running along the path." These are illustrations, not measured results. Each caption should rest on actual model predictions and should not add details the evidence does not support.

F. Testing and Evaluation

The finished system should be tested on videos that vary in activity, environmental sound and recording conditions. Table II lists what will be assessed.

Evaluation aspect	**Purpose**
Event recognition	Measure whether the detected event matches the reference label.
Caption relevance	Check whether the sentence describes the actual video.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131104

Caption completeness	Determine whether the important audio and visual details are included.
Audio-video fusion	Compare fused predictions with audio-only and video-only results.
Behavior under difficult conditions	Examine performance with noise, poor lighting and simultaneous events.
Processing time	Measure how long it takes to analyze a video and produce a caption.

Caption quality can be judged by human review together with standard text-based metrics. No numerical performance claims will be made until testing is done.

V. EXPECTED BENEFITS AND DISCUSSION

The system is expected to improve video understanding by pairing what is heard with what is seen. Instead of describing individual frames, it tries to consider the whole event and how the two modalities relate.

A. Expected Benefits

- Combining audio and video may help identify events that are hard to interpret from one modality alone.
- The captions should be readable descriptions of activities and environmental events.
- Audio cues may separate visually similar activities, and visual cues may resolve ambiguous sounds.
- Automatic captions would reduce the effort of describing video by hand.
- With suitable models and data, the architecture can be adapted to different kinds of video.
- Contextual captions may help people who cannot access all of the audio or visual information in a recording.

These are expected benefits and have not been confirmed by experiment.

B. Limitations

- Background noise can cause wrong audio predictions.
- Several simultaneous activities may be hard to tell apart.
- Audio and visual events may not line up exactly in time.
- Running both modalities can need a lot of memory and processing power.
- A caption may include unsupported details even when the event prediction is reasonable.
- Poor lighting, low resolution, camera movement and occlusion can hurt visual analysis.

To reduce these problems, the system should use sensible preprocessing, confidence-based decisions and temporal alignment, and should be evaluated on varied samples. If the evidence is insufficient, it should not present an uncertain event as a confirmed fact.

VI. FUTURE SCOPE

- Real-time analysis: extend the system to live camera feeds so captions appear as events happen.
- Advanced fusion: try cross-modal attention or transformer-based methods, following the work in [1], [7] and [9].
- Temporal event localization: find the start and end of each activity and caption specific segments.
- Long-video summarization: produce a series of captions or a short summary for longer recordings.
- Multilingual captions: generate captions in more than one language.
- Uncertainty handling: use confidence scores and disagreement detection to avoid misleading captions.
- Domain-specific use: adapt the system for surveillance, education, sports, environmental monitoring and accessibility.
- Performance optimization: look at lightweight models, frame sampling and efficient feature extraction to cut processing time.

Which of these is feasible depends on the models, datasets and computing resources available, and on the results of the first evaluation.

VII. CONCLUSION

This paper describes a proposed Context-Based Audio-Video Analysis system that uses sound and images together to recognize activities and write captions. The surveyed research supports using multimodal learning, feature fusion and reasoning over time, and the system is designed around those ideas. Its performance will depend on model quality, audio-

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131104

video synchronization and caption accuracy. The next step is a working prototype, followed by tests on a varied set of videos to see how well it holds up.

REFERENCES

[1] J. Zhang, "Multimodal Video Context Understanding Based on Dynamic Sparse Attention Mechanism," in Proc. IEEE International Conference on Machine Learning and Computer Application (ICMLCA), 2025.

[2] Q. Liu et al., "Human-Inspired Computing for Robust and Efficient Audio-Visual Speech Recognition," IEEE, 2025.

[3] C. L. Fabian et al., "Evaluating Classical Machine Learning Algorithms for Multimodal Human Activity Recognition," in Proc. IEEE AMLDS, 2025.

[4] G. Y. Li et al., "Spatiotemporal Learning with Context-Aware Video Tubelets for Ultrasound Video Analysis," in Proc. IEEE International Symposium on Biomedical Imaging (ISBI), 2025.

[5] Z. Lin and N. Harte, "Uncovering the Visual Contribution in Audio-Visual Speech Recognition," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025.

[6] Z. Zhuo, Z. Zeng, E. B. Mazomenos, and Y. Jin, "SurgRAW: Multi-Agent Workflow with Chain of Thought Reasoning for Robotic Surgical Video Analysis," IEEE Transactions on Image Processing, 2026.

[7] Q. Xu et al., "Cross-Modal Feature Enhancement Networks for Audio-Visual Event Localization," IEEE Sensors Journal, 2025.

[8] T. Yuan et al., "Surveillance Video-and-Language Understanding: From Small to Large Multimodal Models," IEEE Transactions on Circuits and Systems for Video Technology, 2025.

[9] M. B. Shaikh et al., "MAiVAR-T: Multimodal Audio-image and Video Action Recognizer using Transformers," in Proc. IEEE European Workshop on Visual Information Processing (EUVIP), 2023.

[10] T. Chen et al., "Neural Wave Propagation for Surgical Video Action Recognition: A New Dataset and Baseline," in Proc. IEEE International Conference on Information Technology and Engineering (ITOEC), 2021.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License