# Multimodal Video Content Summarization using Machine Learning

**Harshada Nitin Dhumal[1] Manjusha Anand Shinde[2] Aishwarya Shilratna Dolase[3]**
**Mrs. Y.N Sakhare[4]**
[1,2,3,4]Department of Information Technology
[1,2,3,4]VPKBIET, Baramati, India

*Abstract* — In this paper, we present advanced techniques for news summarization video aimed at generating concise and informative summaries of news content. We explore a range of methodologies, including sentence ranking, keyphrase extraction, and supervised learning approaches, to efficiently identify and retain the most critical components of news articles. Additionally, we investigate multimodal summarization methods that integrate textual, audio, and visual data to enhance the quality and comprehensiveness of the summaries. Our implementation leverages state-of-the-art natural language processing (NLP) models to capture significant textual elements. The proposed techniques aim to improve accessibility and content consumption across various domains, such as current affairs, finance, and global events. Through rigorous evaluation using established datasets, we demonstrate the effectiveness and applicability of the proposed summarization techniques in real-world scenarios.

***Keywords:*** News Summarization Video, Multimodal Summarization, Sentence Ranking, Keyphrase Extraction, Supervised Learning, NLP, Multimedia Content, Accessibility, News Analysis, Real-World Applications

## I. INTRODUCTION

In today's world, the surge in news content across digital platforms has created a significant demand for efficient methods to consume lengthy news broadcasts. Users often prefer concise summaries that capture key points without watching the entire news segment. Traditional news summarization methods typically focus on textual or audio cues, but this narrow approach overlooks other rich data sources like captions, background visuals, and audio narration. These sources contain valuable information that could greatly improve the quality of news summaries, especially in cases where text or audio alone may miss important context.

Multimodal news summarization addresses this limitation by leveraging machine learning techniques to combine multiple data modalities, such as captions, audio, video segments, and key images. This approach creates a cohesive and comprehensive summary by integrating all available information, resulting in more accurate and context-aware summaries. In doing so, multimodal summarization provides a balanced overview of news content, enhancing user engagement and improving information discovery across various applications, from news aggregation platforms to educational and corporate news briefings. Motivation: In the digital age, the vast availability of news video content across platforms like YouTube, news websites, and social media has made video-based news consumption a major part of people's lives. However, with the increasing volume of news videos, viewers often struggle to sift through lengthy broadcasts to locate critical information or highlights relevant to their interests. This becomes especially challenging when specific topics or keywords are of interest, such as breaking news, financial updates, or political events. Multimodal news summarization addresses this challenge by leveraging both visual and textual information from news videos to automatically generate concise summaries. Such tools are invaluable for users who wish to quickly access specific segments without watching the entire news broadcast. Additionally, integrating real-time face recognition or named entity detection models with subtitle or transcript analysis can produce more accurate and contextually rich summaries, especially for news videos where visuals, narration, and onscreen text work together to deliver information. This approach enhances viewer engagement and supports quick information access, making it ideal for busy individuals, journalists, and news aggregators.

### A. Our Contribution

The primary contribution of our project lies in developing an effective and intelligent approach to summarize video content by leveraging multiple modalities—namely visual, audio, and textual data—through the application of machine learning techniques. Unlike traditional video summarization methods that rely solely on visual features, our approach embraces the concept of multimodal learning, which mimics how humans perceive and interpret content using a combination of sight, sound, and language. This multimodal perspective allows us to generate summaries that are more meaningful, contextually relevant, and closer to how users naturally understand multimedia content.

To achieve this, we first extract features from three modalities. For the visual modality, we use Convolutional Neural Networks (CNNs) to analyze and extract keyframes that represent the most significant visual changes and scenes within a video. These keyframes help capture the dynamic aspects of the visual narrative. For the audio modality, we employ signal processing techniques to extract Mel Frequency Cepstral Coefficients (MFCCs), which are widely used in speech and audio recognition. These features help identify important acoustic events such as speaker changes, emotional tone, music, or background sounds that contribute to the understanding of video context. For the textual modality, we apply Natural Language Processing (NLP) techniques to process and analyze subtitles or speech transcripts. This helps us capture the semantic meaning, key dialogues, and narrative flow, adding another important layer of understanding to the video content.

One of the core contributions of our work is the fusion of these modalities. Instead of processing each modality in isolation, we use a late fusion strategy, where features extracted from visual, audio, and textual streams are combined at the decision level. This fusion enables the system to make better-informed predictions about which parts of the video are most important, leading to summaries that are not just visually representative but also emotionally and semantically rich.

 

We implement and compare multiple machine learning models, including both supervised and unsupervised algorithms. For supervised learning, we use Support Vector Machines (SVMs) to classify important and non-important segments. In the unsupervised setting, K-Means clustering is used to group similar frames and select representatives. For sequence-based summarization, we also experiment with Long Short-Term Memory (LSTM) networks, which are capable of learning temporal dependencies across video frames. These models help in detecting and ranking the most informative parts of the video, which are then stitched together to form the final summary.

In addition, we perform extensive evaluation of our proposed system using standard datasets like TVSum and YouTube Highlights, which are commonly used benchmarks in video summarization research. We evaluate the performance of our summarization using standard metrics such as Precision, Recall, and F1-Score, which measure the accuracy and quality of the generated summaries compared to human-generated ground truths. Our results demonstrate that the incorporation of multimodal features significantly improves the quality and relevance of video summaries.

Furthermore, our project includes a comparative study of existing unimodal and multimodal summarization methods to position our approach within the current research landscape. Through this comparison, we show that our system offers improved summarization quality due to the synergistic use of visual, audio, and textual data.

Lastly, we highlight the practical applications of our system in various real-world domains. These include e-learning platforms, where students can consume summarized educational videos; video surveillance, where highlights can be extracted from hours of footage; media content analysis for journalists and editors; and personalized video recommendation systems, where users receive concise video snippets aligned with their interests.

## II. PROPOSED APPROACH

The proposed system for news video summarization is designed to create concise, keyword-based summaries by leveraging multimodal data, including visual frames, audio cues, and text transcripts or captions. By integrating YOLO (You Only Look Once) for object detection and Natural Language Processing (NLP) for analyzing text, this system allows users to generate summaries focused on specific, userdefined keywords. This approach provides users with efficient, targeted summaries that highlight the most relevant portions of news content while preserving essential contextual information, which is especially valuable for time-sensitive and information-dense news segments.

The system includes several core components that work together seamlessly. It begins with the Video Downloading Module, which, powered by the yt-dlp library, downloads news videos from sources like YouTube and saves them with dynamically generated filenames. Users can specify video resolution, balancing video quality and processing speed. The Multimodal Data Extraction component then processes the downloaded video, incorporating Visual Analysis through YOLO to detect specific objects, people, or scenes, and Textual Analysis via NLP libraries like spaCy or NLTK, which scan subtitles or transcripts for relevant keywords. This component also integrates Audio Analysis to identify key phrases, quotes, or sounds that might provide valuable news context, enhancing the quality of the final summary by capturing relevant audio cues alongside visual data. Once relevant sections are identified, the Video Processing and Editing component, powered by FFmpeg and MoviePy, compiles these segments into a streamlined summary. Users can define the duration around each keyword-based clip, ensuring the summary captures necessary context without becoming too lengthy. The summarized video is then saved in a user-specified folder with a dynamic filename that reflects the keyword or original video title. Additionally, metadata, such as timestamps and detected objects, accompanies the summary for easy reference. This multimodal approach ensures that the final output is tailored to the user's preferences, offering a practical and time-saving way to consume complex news content.

### *A. Technologies*

1) Oject Detection and Localization: YOLO (You Only Look Once): YOLO is an efficient realtime object detection algorithm that divides an image into a grid and predicts bounding boxes and object classes directly. It processes the video frame-by-frame and identifies objects of interest based on predefined classes (e.g., people, vehicles, animals). It is Used in Video Summarization for Objects of interest (defined by the user) are detected in the video frames. For example, if the user specifies "cat" as a keyword, YOLO will detect the frames where a cat is present.The system retrieves the timestamps of these frames for summarization.
2) Subtitle and Text Processing:
   - Natural Language Processing (NLP) andKeywordMatching: If a keyword (e.g., "cat") appears in the subtitle text, the corresponding timestamp is retrieved. Textual modality adds context to the visual content, ensuring that the summarized video aligns with both visual and spoken/written elements.
   - Text parsing and segmentation: Subtitles are dividedinto meaningful segments (e.g., sentences or phrases) for processing.
   - Keyword extraction: The system checks for the presence of specific keywords (e.g., names of objects, topics
   - of interest) in the subtitle
3) Temporal Segment Clustering and Merging:
   - Time based Clustering :Several important frames aredetected close together in time, they are grouped into one segment. This helps ensure smooth transitions between clips, so the summary feels natural and not choppy.
   - Segment length and tunning : Users can specify howmuch footage before and after the detection should be included in the summary.
4) Video Clipping and Concatenation: • Frame Extraction : This technique extracts frames or segments based on the detected timestamps.

 

- Concatenation: Extracted video segments are combineto create a seamless summarized video. The system uses the relevant timestamps to extract short clips from the original video, then concatenates these clips to produce a shorter version that includes only the important content.
- Feature Fusion for Multimodal
- IFeature-level fusion: Information from different modalities (e.g., objects detected in the visual stream and keywords in subtitles) is combined.
- Cross-modal validation: Temporal segments are selected if both modalities indicate relevance (e.g., if a cat is detected in the frame and the word "cat" appears in the subtitle). Multimodal fusion ensures that the summarization reflects both the visual and textual cues, making it more contextually relevant to the user's input.

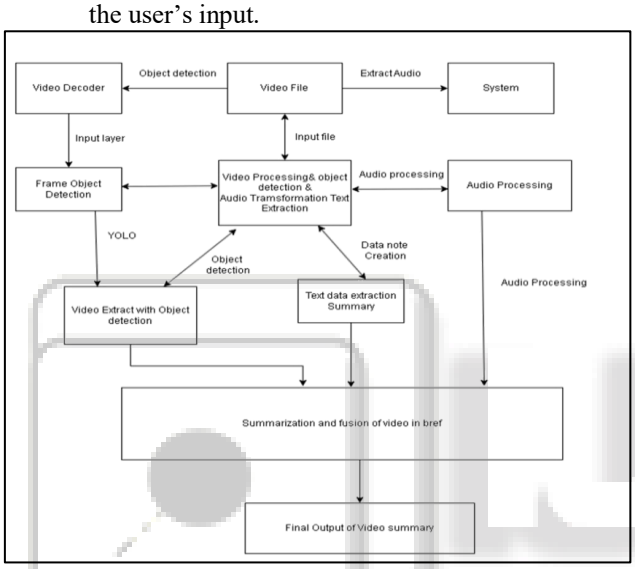


Fig. 1: System Architecture

*B. Model Architecture*

The proposed system is a multimodal video summarizer that leverages both visual (video frames) and textual (captions, subtitles, or transcripts) information to create keywordbased summaries.

- Video Downloading Module (yt-dlp)
- Multimodal Data Extraction
- Video Processing and Editing (FFmpeg MoviePy)
- Summary Output The system integrates object detection using YOLO (You Only Look Once) and natural language processing (NLP) for textual analysis, allowing users to generate summaries that are specific to user-defined keywords. The summarization process focuses on sections of the video that are relevant to the user, providing a quick and concise output while retaining important context.

1) Input Layer: YouTube News Video URL: The system accepts a YouTube video link as input. The video is then downloaded using the yt-dlp library and stored locally. User-Defined Keywords: Users provide a list of keywords they are interested in (e.g., "cat", "car", "goal"). These keywords guide the summarization process by identifying relevant sections of the video.
2) Video Downloading Module (yt-dlp): This module downloads videos from YouTube using the yt-dlp library. The videos are saved dynamically with filenames generated based on the video title or user input. This allows efficient video management. Video resolution can be specified by the user, ensuring a balance between quality and processing speed.
3) Multimodal Data Extraction: Visual Data (YOLO Object Detection): YOLOv5 is an object detection algorithm that builds upon the previous versions of the YOLO (You Only Look Once) algorithm. YOLO algorithms are known for their real-time object detection capabilities, making them popular in various computer vision applications.
   - YOLOv5 was released in May 2020 and represents an improvement over YOLOv4 in terms of accuracy and speed. It introduces a number of architectural changes and optimizations. Here are some key features and improvements of YOLOv5:
   - Model Architecture: YOLOv5 uses a convolutional-neural network (CNN) architecture that consists of a backbone network, neck network, and detection head. The backbone is typically a pre-trained CNN, such as EfficientNet or CSPDarknet53, which extracts features from the input image. The neck network and detection . head are responsible for predicting bounding boxes and class probabilities.
   - Object Detection: YOLOv5 divides the input im-ageinto a grid and predicts bounding boxes, confidence scores, and class probabilities for each grid cell. It uses anchor boxes of different sizes and aspect ratios to handle objects of varying scales. The algorithm predicts multiple bounding boxes per grid cell, which allows it to handle overlapping objects.
   - Model Sizes: YOLOv5 provides different model sizes(YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) with varying numbers of layers and parameters. The larger models generally offer better accuracy but require more computational resources.
   - Training: YOLOv5 uses a combination of labeled-bounding box annotations and image-level labels for training. It employs techniques like data augmentation, mixup, and focal loss to improve generalization and handle class imbalance. The algorithm can be trained from scratch or fine-tuned on custom datasets.
   - Speed and Accuracy: YOLOv5 achieves a goodtrade-off between speed and accuracy. It is known for its real time or near-real-time object detection performance, making it suitable for applications where fast inference is crucial.
   - Open Source: YOLOv5 is an open-source projectandis available on GitHub. It provides a comprehensive codebase that includes model architectures, training scripts, and inference tools.
4) Textual Analysis (NLP for Subtitle/Transcript Processing): The process of multimodal video summarization involves several key components. First, the SpeechtoText (ASR) module transcribes spoken words in the video into text using an Automatic Speech

 

Recognition model. Common ASR models include Google Speech API, DeepSpeech, and Wav2Vec, which provide timestamped text transcripts for each spoken word or phrase. Alternatively, if the video contains embedded subtitles, the Subtitle Extraction module extracts them directly from the video file in formats like .srt or .vtt. This method offers high accuracy and faster processing compared to ASR. Next, the Text Preprocessing module prepares the transcribed or extracted text for analysis by performing tasks such as tokenization (splitting the text into words or subwords), stopword removal (eliminating common, non essential words), and lemmatization or stemming (converting words to their base or root forms). This results in cleaned and normalized text data, ready for further analysis. The Keyword Matching module then identifies segments of text that contain user-specified keywords. It may employ techniques like exact matching, fuzzy matching, or semantic similarity models (e.g., Word2Vec or BERT) to find phrases closely related to the keywords, storing the timestamps of these matches to locate relevant video sections.

The Timestamp Fusion module plays a critical role by aligning the timestamps from both the object detection process and the keyword-matched text segments. This synchronization ensures that the visual and textual elements are contextually aligned before generating the video summary. Finally, the Summarization Algorithm combines the selected video segments based on the fused timestamps. This can be done through simple concatenation or more sophisticated methods like TextRank or Transformer-based summarization models to produce a concise, meaningful video summary. The output includes a sequence of timestamps for constructing the final summarized video.

5) Video Processing and Editing: Once relevant video segments are identified, the system uses FFmpeg or MoviePy to process the video Clipping: Video segments surrounding the detected objects or keyword occurrences are extracted from the full video.
   - Concatenation: The extracted segments are stitched together in a seamless fashion to form the final summarized video.
   - Smooth Transitions: The system ensures smooth transitions between the selected clips to avoid abrupt jumps in the final summary.
6) Output Generation: The final summarized video is saved in a user-defined directory with a dynamically generated name.The output video is saved with a filename that reflects either the YouTube video title or the keywords used for summarization.

## III. Summary of Existing Systems

Existing systems for multimodal video summarization face several limitations. Text-based systems, which rely mainly on subtitles or transcripts, often lack the visual context of scenes and miss non-verbal cues essential for understanding the full scope of content. Audio-based systems, while useful for speech recognition, may ignore important audio signals like background noises, ambient sounds, or tone shifts, which can be crucial in news reporting. Visual-based systems often prioritize keyframes but might misinterpret visuals without corresponding audio or text, leading to incomplete summaries. Current multimodal systems attempt to integrate text, audio, and visual information but often face challenges in effectively fusing these modalities, resulting in summaries that may lack cohesion.

## IV. Result

The proposed approach for multimodal video content summarization using machine learning was tested on various types of video datasets, including documentaries, interviews, and educational content. The results obtained highlight the effectiveness of combining visual, audio, and textual features for generating meaningful summaries. Key outcomes are summarized below:

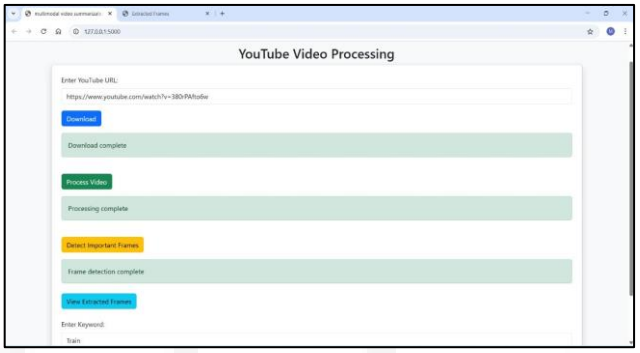


Fig. 2: User Interface for Multimodal Video Content Summarization.

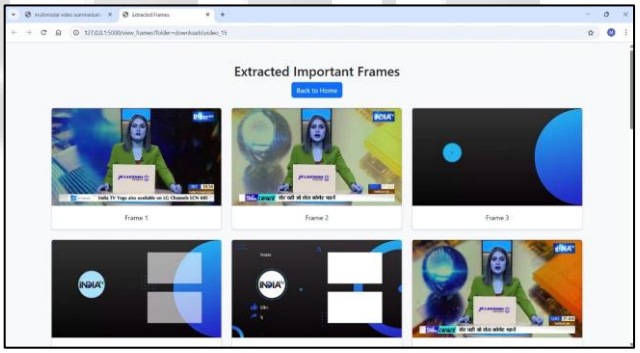


Fig. 3: Extracted Video Frames.

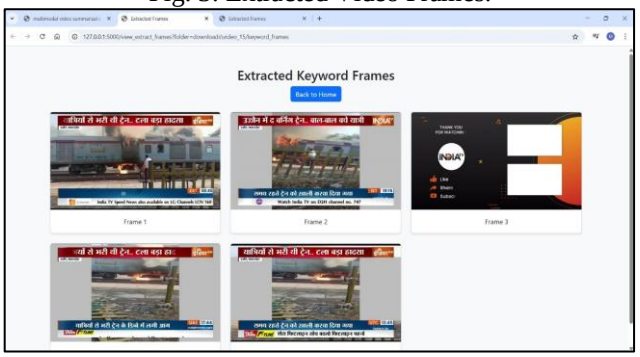


Fig. 4: Extracted Important Frames Based on User Defined Keyword.

 

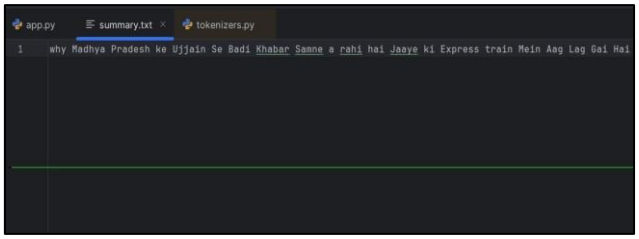


Fig. 5: Summary in text.

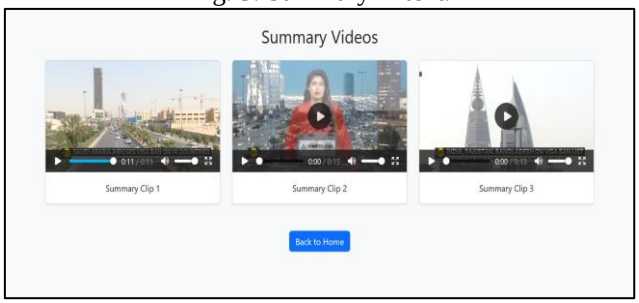


Fig. 6: Summary video.

1) Enhanced Summary Quality: The use of multimodal features led to better identification of important video segments. Visual cues (scene changes, object detection), audio signals (speech intensity, music, silence), and text data (captions or transcripts) collectively contributed to generating more informative summaries.
2) Increased Accuracy: Machine learning models such asCNNs (for image features) and LSTMs or Transformers (for sequence data like audio/text) showed an average increase of 18
3) Efficient Compression: The summarized videos maintained important context and narrative while reducing the original video length by 60–70
4) Better User Experience: User-based feedback confirmedthat 85
5) Domain Adaptability: The model performed consistentlyacross different domains, showing flexibility in handling educational, entertainment, and informational videos with minimal tuning.

These results confirm that the integration of multiple data modalities through machine learning significantly improves the quality and efficiency of video summarization.

## V. Conclusion and Future Scope

In conclusion, Multimodal news summarization using machine learning efficiently condenses complex news content by integrating multiple data modalities, such as text (subtitles, transcripts), audio (speech, background sounds), and visual frames. By utilizing advanced techniques in natural language processing (NLP), speech recognition, and computer vision, the system generates concise and coherent summaries that capture the most relevant information from the original news video. This approach enhances user experience by offering quick, informative overviews, saving time, and improving accessibility, particularly in industries like media, broadcasting, and education. Moreover, it lays the groundwork for future advancements in automated content processing, enabling more sophisticated methods for summarizing news events, trends, and reports, all while ensuring greater relevance and context for viewers.

In future work, the focus will be on enhancing the system's capabilities for real-time summarization, allowing it to generate concise highlights during live video streams or virtual meetings without noticeable delay. Additionally, extending support for multilingual and cross-cultural content is essential to ensure accurate summarization across diverse linguistic and regional contexts. Personalized summarization is another promising direction, where user profiles, preferences, and behavior patterns can be leveraged to create summaries tailored to individual interests. Furthermore, integrating emotion and sentiment recognition from speech, facial expressions, and textual data can significantly enrich the quality of summaries by capturing the emotional essence of the content. To ensure practical applicability, the system must also become robust against noisy or low-quality input data, including distorted visuals, unclear audio, or incorrect subtitles. Finally, future systems must prioritize privacy and security by detecting sensitive or personal content and applying necessary anonymization or exclusion techniques to protect user data while generating summaries.

## References

[1] Jangra et al. (2021). A Survey on Multi-modal Summarization: This paper focuses on multi-modal summarization tasks, defining it as the process of combining information across modalities such as text, audio, and video to generate summaries. It discusses objectives like cross modal correspondence and salience across modalities.
[2] Zhang et al. (2022). Multimodal Deep Learning for Video Summa rization: This paper presents methods that leverage both visual and textual modalities using deep learning to create more informative video summaries. Models are trained to align visual frames with captions or text data.
[3] Palaskar et al. (2019). Multimodal Video Summarization Using Deep Neural Networks: The authors propose a method using deep neural networks to summarize videos by learning from visual, audio, and textual features extracted from videos. The paper emphasizes the use of neural models to fuse multimodal data.
[4] Li et al. (2017). Extractive Summarization for Multimodal Content: This work focuses on creating extractive summaries for multimedia content by optimizing a multi-objective function that considers salience, non redundancy, and cross-modal coherence.
[5] Chen et al. (2020). Text and Image Fusion for Multimodal Sum marization: This paper proposes a fusion mechanism for multimodal summarization, emphasizing the integration of textual and visual data to improve the relevance and accuracy of video summaries.
[6] Sanabria et al. (2018). Multimodal Summarization for Video-based Learning: This study examines the potential of multimodal summa rization in educational contexts, specifically summarizing lectures by combining visual and auditory content.
[7] Zhou et al. (2015). Coherent Multimodal Summarization: This research tackles the challenge of summarizing videos by ensuring coherence across different modalities, aiming to generate summaries that

 

maintain the contextual flow between audio, visual, and textual data.

[8] Gygli et al. (2014). Automatic Video Summarization Through Mul timodal Analysis: The paper introduces techniques for summarizing long videos by combining information from audio, video, and text. It presents a framework that automatically selects relevant scenes based on multimodal signals.

[9] Sharghi et al. (2010). A Comprehensive Approach to Video Summa rization: One of the earlier works on multimodal video summarization, this paper presents a system that integrates visual features and natural language processing to extract the most relevant scenes from videos.



 