International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

YouTube Lecture Video Summarizer and Key Concept Extractor using NLP and Computer Vision

Bright Lee M S[1], Vishnu Mohan C[2]

[1]Scholar, [2]Assistant Professor
[1]Dept. of Computer Science Sacred Heart College, Thevara Emakulam, Kerala, India
[2]Dept. of Computer Science
Sacred Heart College, Thevara Emakulam, Kerala, India
[1]21umcp2961@shcollege.ac.in, [2]vishnumohan@shcollege.ac.in

Abstract

With the exponential growth of video content on platforms such as YouTube, there is an increasing demand for intelligent systems that can summarize videos and extract essential information automatically. This paper proposes a hybrid system that combines state-of- the-art Natural Language Processing (NLP) techniques with Computer Vision to create a YouTube video summarizer and key concept extractor. The proposed system performs automatic video downloading, audio extraction, speech transcription, text summarization, key concept extraction, scene and keyframe detection, graph identification, and finally compiles the processed information into both a downloadable PowerPoint presentation and a detailed notes PDF. The system provides a comprehensive and structured overview of a video's content in both textual and visual form, enhancing learning efficiency and infor- mation retrieval.

Index Terms— YouTube Summarization, Whisper, BART, KeyBERT, NLP, Key Concept Ex- traction, PowerPoint Generation, PDF Notes, OpenCV, Scene Detection, Graph Detection

1 Introduction

With over 500 hours of video uploaded to YouTube every minute, the need for intelligent tools to efficiently extract meaningful content from long videos has become critical, as manual re- view is no longer practical. This has led to the development of a wide range of automated summarization techniques, including extractive methods that select key sentences, abstractive methods that generate human-like summaries, and hybrid systems that integrate text, audio, and visual cues. In this paper, we analyze 26 notable research contributions in YouTube video sum- marization, examining methods such as BART (Devi et al., 2023), LSA (Sanjana et al., 2021), and user-personalized recommenders (Kannan et al., 2023), with particular attention to Jiang et al. (2019), who introduced a comprehensive system leveraging deep learning to fuse tex- tual, visual, and auditory signals. Our review compares these

approaches in terms of accuracy, processing speed, and practical usability, and also highlights key challenges and future direc- tions, including real-time summarization, multilingual support, and advances in unsupervised learning.

2 Literature Review

Mandar et al. [1] developed a web-based YouTube Video Summarizer that blends visual and textual techniques. Earlier models used CNNs and RNNs (like LSTM) for keyframe detection and context-aware summaries. Recent improvements involve transformer-based models like BERT, which better understand bidirectional context in text. Additionally, attention mecha- nisms and user-ranking strategies enhanced summary relevance. While effective, these sys- tems still rely heavily on transcripts and require significant computation. Overall, there's a shift toward hybrid deep learning and NLP-based methods aimed at building intuitive, efficient summarization tools for e-learning and digital content.

Bandabe et al. [2] explored YouTube transcript summarization using Flask, focusing on ex- tractive NLP techniques. Past work by Yadav et al. showed how such summaries help learn- ers save time. Advanced models like RCNN improved context understanding but struggled with memory issues. More recent tools like Hugging Face transformers enabled abstractive, natural-sounding summaries. Hybrid models also incorporated speech recognition and transla- tion for broader accessibility. Despite progress, challenges like transcript dependency and high computation persist, highlighting the balance between usability, accuracy, and performance in summarization systems.

Singh and Singh [3] explore evolving trends in text summarization using NLP, contrasting traditional extractive methods like TF-IDF and clustering with advanced Transformer-based abstractive models. While extractive summaries are simpler, they lack depth and coherence. Abstractive techniques produce more fluent results but face challenges in accuracy and domain adaptability. Their study emphasizes practical applications such as academic and journalis- tic summarization, while also acknowledging the need for greater explainability and domain integration to boost effectiveness across sectors.

Inamdar et al. [4] proposed a YouTube Transcript Summarizer that blends extractive methods like SpaCy and Gensim with HuggingFace's abstractive models to generate concise summaries. The system uses PyTube for downloading audio and HuggingSound for transcription, providing efficient real-time outputs via a Chrome extension. While it streamlines content consumption, its dependence on transcript quality and system resources is a key limitation. Future directions include multilingual support and improved abstractive summarization for more natural results. **Bhoir (2025) [5]** proposes a YouTube Transcript Summarizer designed to tackle information overload in video content using both extractive and abstractive NLP techniques. Built with tools like YouTubeTranscriptAPI and Hugging Face transformers, the system delivers concise, human-like summaries via a user-friendly Chrome extension. It proves especially useful for educational and professional users by enhancing accessibility and saving time. However, its reliance on transcripts and challenges with audio-only content remain limitations. Future im- provements may include multilingual support and better accuracy across diverse video types.

Nethranand et al. [6] introduce Summer-Riser, an abstractive text summarization tool that leverages large language models and LangChain to generate natural, paraphrased summaries from long documents. Unlike extractive methods, it processes URLs through stages like data collection, preprocessing, and embedding generation using FAISS for efficient similarity searches. Designed for news

and educational content, the system offers high readability and user satis- faction. However, it faces challenges with domain-specific terminology and non-English texts.
Future work aims to enhance multilingual support and contextual understanding.

Vanisree et al. [7] present YouTube Transcript Essence AI, an NLP-based tool using Hugging Face Transformers to generate concise, human-like summaries of YouTube video transcripts. Unlike extractive methods, it employs abstractive summarization for clearer, time-efficient re- sults. Integrated through a Chrome extension and Python Flask backend, it supports both auto and manual transcripts with customizable summary lengths. While effective for educational and professional use, its current limitations include lack of non-English support and depen- dence on available transcripts. Future updates aim to add multilingual capabilities and voice- to-text for videos without captions.

Dharmani et al. [8] introduce a YouTube Transcript Summarizer Chrome extension that com- bats information overload using both abstractive (BART) and extractive (T5) NLP techniques. The tool also integrates KeyBERT for keyword extraction and supports translation into nine Indian languages, enhancing accessibility. With Firebase-enabled feedback and a user-friendly design, it proves valuable for educational and professional use. However, its limitations in- clude dependency on available transcripts and potential translation inaccuracies. Future im- provements aim to support caption-less videos and customizable summary lengths for broader applicability.

Patil et al. [9] present an innovative summarization approach called Improvised Clarity-cuts Text Summarization with Text to Graph Prediction, which combines NLP techniques with data visualization. Using transformer-based models like BERT and GPT, the system performs both extractive and abstractive summarization and introduces a Text2Chart feature that converts text into bar and pie charts via machine learning classifiers, with Linear SVM achieving 78% accu- racy. The architecture leverages attention mechanisms and syntactic parsing for coherence and visual relevance. While effective for simpler texts, challenges remain with complex sentences and graph precision. Future improvements target domain-specific use, real-time processing, and enhanced graph generation.

Verma et al. [10] present a YouTube Transcript Summarizer tool designed to address in- formation overload by generating concise summaries of long videos using Automatic Speech Recognition (ASR) and NLP techniques, including both extractive and abstractive methods. Implemented as a Chrome extension with a backend server using the YouTube API, the system efficiently retrieves and processes transcripts for user-friendly summaries. Aimed at enhancing accessibility for students and professionals, the paper details the system's architecture, scalabil- ity, and accuracy, while suggesting future improvements like multilingual support and caching. The study underscores the growing importance of summarization technologies in today's digital era.

Varagantham et al. [11] propose a text summarization system using NLP techniques to ad- dress digital information overload by generating simple, coherent summaries. The approach focuses on extractive summarization, employing tokenization, stop-word removal, stemming, and k-means clustering to group sentences based on semantic importance and frequency. The paper outlines the system's end-to-end workflow and emphasizes its user-friendly interface.While extractive methods are favored for their simplicity and speed, the authors acknowledge the potential of abstractive summarization for future enhancements. The tool effectively saves users time and effort in processing large textual data.

iang et al. [12] introduces a novel way of video summarization by posing it as a content-based recommendation problem with a focus to tackle the issue of information overload in viewing videos.

The authors introduce a scalable deep neural network (HighlightNet) for forecasting segment importance scores by modeling each segment and its context in the entire video using features from visual, audio, and semantic sources. The main contributions are the hybrid archi- tecture merging SegNet, VideoNet, and HighlightNet, along with methods such as multi-task learning and data augmentation to counter overfitting when training on small datasets. Exper- iments upon the CoView 2019 dataset show the efficacy of the model, with visual semantic features being most effective, though audio and optical flow features provided marginal bene- fit. The paper also discusses action and scene recognition in untrimmed videos, though their incorporation did not take summarization performance significantly higher. Future research would involve using collaborative filtering and larger datasets to improve the recommender- style framework.

Ligade et al. [13] present the Automated Video Transcript Summarizer, a hybrid NLP-based tool designed to efficiently extract key insights from lengthy video content. Combining extrac- tive (LSA) and abstractive (T5) summarization methods, the system offers multilingual output via DeepL APIs and includes a Flask-based interface with text-to-speech support, enhancing accessibility for diverse users, including the hearing-impaired. Using SpaCy for processing and transcript retrieval APIs, it proves accurate and scalable for educational, business, and research use. The tool effectively reduces content consumption time, with future improvements aimed at personalized summaries and enhanced audio features.

Charate et al. [14] introduce a multilingual text summarization system using NLP to handle large volumes of content in Hindi, Marathi, and English. Leveraging IndicBART-XLSum for Indic languages and Pegasus-XSum for English, the system generates accurate, concise sum- maries while using Google Cloud Translate v2 API for cross-lingual support. Deployed on the cloud with a user-friendly React interface, it ensures scalability and accessibility. Evaluation with ROUGE and BLEU scores confirms the models' effectiveness, and the work highlights the potential of hybrid summarization methods for future real-time and multilingual NLP ap- plications.

adam et al. [15] explore the challenges and opportunities in video summarization using ma- chine learning, focusing on single-view (SVS) and multi-view (MVS) approaches. They cat- egorize methods into static (keyframe-based) and dynamic (segment-based) summarization, highlighting applications in healthcare, security, and sports. The paper emphasizes the shift toward deep learning for improved accuracy and efficiency, discusses evaluation metrics like F-score, precision, and recall, and lists key datasets in the field. While SVS is well-studied, MVS remains complex due to camera synchronization issues, suggesting future research in unsupervised and real-time summarization for better real-world deployment.

Kumari et al. [16] present a YouTube Transcript Summarizer built with Flask and NLP to automate transcript extraction and summarization, improving content accessibility and user ef- ficiency. Using YouTube closed captions, the system applies Hugging Face Transformers and CNN models to generate concise summaries via a Flask-based web interface. Key features include multilingual translation, downloadable summaries in PDF/Word, and sharing via email or WhatsApp. While effective, limitations include dependence on pre-timed subtitles and chal- lenges with non-English formats. The tool contributes to advancing NLP-driven video content management through both extractive and abstractive methods.

Siddhartha, Pandey, and Saxena [17] review the progression of YouTube transcript summa- rization techniques, tracing developments from early rule-based and extractive NLP methods to advanced deep learning models. They highlight the shift from CNN-RNN and hybrid models to transformer-based architectures like BERT and BART for more accurate and human-like ab- stractive summaries. The paper emphasizes the growing effectiveness of integrating NLP with modern AI to produce concise summaries, aiding users in navigating large volumes of video content efficiently.

Yadav et al. [18] present a hybrid NLP-based system for summarizing YouTube video tran- scripts and extracting keywords to tackle long, redundant video content. The approach com- bines extractive methods like TF-IDF and TextRank with abstractive Seq2Seq models using attention mechanisms. Preprocessing steps include punctuation restoration, tokenization, and lemmatization. Hosted via a Flask API with a React interface, the tool is designed for ease of use and proves valuable in educational, professional, and research settings by enhancing content accessibility and saving time.

Sulochana Devi et al. [19] present an abstractive summarization system for YouTube video transcripts that leverages NLP and machine learning to generate concise, human-like sum- maries, improving content accessibility and reducing redundancy. Building on earlier extractive methods like TF-IDF and LSA, the system uses encoder-decoder architectures with attention mechanisms and Hugging Face Transformers for deeper contextual understanding. The pa- per compares extractive and abstractive techniques, emphasizing the efficiency of LSA and the adaptability of modern deep learning. A Chrome extension enhances usability with features like translation, speech output, and email sharing, showcasing the system's real-world applica- bility.

Dilawari A et al. [20] introduces an end-to-end deep learning model that jointly performs video description and abstractive summarization using multitask learning and word-level at- tention. Moving beyond traditional extractive and template-based methods, ASoVS leverages CNN-RNN architectures and attention mechanisms to capture spatial-temporal features and generate coherent, concise summaries. Evaluated on datasets like MSR-VTT, TRECVID, and CNN/Daily Mail, the model outperforms existing approaches in informativity, readability, and conciseness, offering a unified and efficient solution for summarizing video content.

Karousos et al. [21] present a hybrid text summarization method to analyze open-ended stu- dent feedback in higher education, addressing the challenges of processing unstructured re- sponses. The system combines TextRank for extractive summarization, Walktrap for clustering, and GPT-4o mini for generating abstractive summaries. Tested at the Hellenic Open University, it effectively summarized feedback on academic services, showing high relevance, coherence, and fluency when evaluated with G-Eval and DeepEval. By blending traditional NLP with modern AI, the approach offers a scalable solution for educational feedback analysis.

Sanjana et al. [22] propose a lightweight video summarization method using NLP and Latent Semantic Analysis (LSA) on YouTube subtitles to efficiently process large volumes of video content without heavy computational demands. The system converts subtitles into text, ap- plies LSA for key sentence extraction, and reconstructs summarized videos using MoviePy, supported by tools like Pytube and Sumy. Designed for training-free use, it is particularly ef- fective for educational content such as TED Talks, enabling faster and more informative video consumption.

Fei et al [23] propose a compressed video summarization method that creates compact yet rich key frames by selecting and combining relevant object activities through an optimization programming approach. The system applies a moving-object-outlier-detection technique and KNN-based matting to

isolate and seamlessly integrate objects into a single image, ensuringa natural visual output. Using Mutual Information for shot segmentation and motion vector analysis to distinguish static and dynamic elements, the method preserves both content diversity and key points of interest while minimizing visual inconsistencies in the final summary.

Huang et al. [24] present a supervised video summarization framework that combines hier- archical spatio-temporal modeling with a novel user-ranking mechanism. The model employs a three-stage deep neural network—2D CNNs for spatial features, 1D CNNs for short-term temporal patterns, and LSTMs for long-term dependencies—to accurately score frame impor- tance. To tackle subjectivity in user annotations, a user-ranking method inspired by PageRank ranks annotators based on consistency, weighting higher-quality labels during training. Tested on SumMe and TVSum datasets, the approach achieves superior F-scores and outperforms previous methods while maintaining computational efficiency.

Li et al. [25] propose a video summarization method that uses sparse coding to detect shot boundaries and generate concise summaries. The approach involves learning a dictionary of visual patterns from the video, using feature reconstruction loss to maintain fidelity and sparsity control to avoid overfitting. Key frames are identified based on their alignment with the learned dictionary, enabling accurate shot boundary detection and selection. This method ensures the final summary is both representative and coherent, effectively capturing the essence of the original video content.

Rajkumar Kannan et al. [26] presents a personalized movie summarization system that uses high-level semantic features—such as characters, events, and concepts—alongside user-defined preferences to generate tailored content summaries. Unlike traditional methods based on low- level visual cues, this system segments movies into shots and scenes, applies semi-automatic annotations, and uses similarity ranking with constrained optimization to select key content. Users interact through a graphical interface to customize summaries by adjusting preferences, keywords, and filters. Tested on eight English films with 20 participants, the system produced more informative and satisfactory summaries than generic ones, especially at the scene level, and received high usability ratings for its adaptive, user-centric design.

3 Comparative Analysis of Existing Literature

Existing summarization approaches focus on either text-based or visual-based summarization. NLP-driven methods like BART and T5 offer extractive and abstractive summarization of text, while visual summarizers detect scenes or extract keyframes using scene change detection or attention-based deep learning. Our work unifies both approaches in a streamlined system, using Whisper for transcription, BART for summarization, KeyBERT for keyword extraction, and OpenCV for visual content analysis.

3.1 Text-Based Summarization Approaches

Textual summarization of video content primarily relies on processing the transcripts associated with the video. This can be broadly classified into two categories: extractive and abstractive techniques.

3.1.1 Extractive Summarization

Extractive approaches focus on identifying and selecting the most significant sentences or phrases from the original transcript without rephrasing. These methods are computationally inexpensive and often language-dependent. Table I compares several representative extractive techniques based on TF-IDF, TextRank, k-means clustering, and SpaCy pipelines.
While extractive methods are fast and effective for summarizing factual or structured data, they often struggle to capture the contextual nuances and relationships between sentences.

3.1.2 Abstractive Summarization

Abstractive summarization, on the other hand, generates novel sentences that capture the essence of the content. These methods rely on deep learning models such as BART, Pegasus, and Transformer-based architectures. They are more semantically aware but computationally in- tensive. Table II outlines a comparison of major abstractive summarization systems.
Abstractive systems provide high-quality, human-like summaries and are suitable for multi- language applications. However, their real-time deployment remains a challenge due to high latency and GPU dependency.

3.2 Visual Summarization Techniques

Visual summarization focuses on selecting representative frames or visual highlights from the video. This section presents two major categories: keyframe extraction and personalized visual summarization systems.

3.2.1 Keyframe Extraction

Keyframe extraction is a popular method to summarize visual content by identifying important frames using computer vision and machine learning techniques. Table III highlights several notable approaches in this domain.
These approaches are effective in reducing visual clutter while preserving semantic content, although they often lack contextual alignment with transcript-based summaries.

3.2.2 Personalized Summarization Systems

Personalized summarization systems tailor summaries based on user preferences or domain- specific requirements. Table IV lists systems that leverage user feedback, semantic filtering, and graph-based summarization.
These methods provide improved user engagement and task-specific summaries but require user interaction and training data to adapt effectively.

Table 1: Comparative Analysis of Extractive Text-Based Summarization Techniques

Paper	Key Technique	Language	Speed	Innovation	Limitation
[18] Yadav	TF-IDF + TextRank	English	1.1s	Hybrid scoring method	Basic and context-free
[22] Sanjana	LSA + MoviePy	English	0.8s	Lightweight, no model needed	Lower contextual accuracy
[3] Singh	TF-IDF + k-means	English	0.9s	Simple statistical grouping	Poor sentence coherence
[4] Inamdar	SpaCy + Gensim	English	1.2s	Uses dependency parsing	No abstraction capability
[11] Varagantham	k-means NLP clustering	English	1.8s	Sentence grouping via NLP	Redundancy in output

Table 2: Comparative Analysis of Abstractive Text-Based Summarization Techniques

Paper	Model	Languages	ROUGE-L	Latency	Key Features	Limitation
[19] Devi	BART	5 languages	0.78	1.8s	Chrome extension, TTS	Limited beyond Indian langs
[14] Charate	IndicBART + Pegasus	45 languages	0.59 (En)	2.1s	Multilingual summarization	Complex deployment
[5] Bhoir	HuggingFace Transformers	English	0.58	2.4s	Easy API integration	No translation support
[6] Nethranand	LangChain LLMs	English	0.62	3.1s	Embedding-based reasoning	Slower for longer inputs
[20] Dilawari	ASoVS (Multitask LLM)	English	0.63	3.5s	Attention + word-level focus	High resource requirements

Table 3: Comparative Analysis of Visual Keyframe Extraction Techniques

Paper	Method	F-Score	Storage Reduction	Innovation	Challenge
[23] Fei	KNN Matting	0.92	37%	Seamless frame stitching	Struggles with static scenes
[24] Huang	3D CNN + LSTM	0.76	29%	Temporal context modeling	Lacks text integration
[25] Li	Sparse Coding	0.71	25%	Dictionary-based	Feature

				learning	interpretability
[7] Vanisree	Transformer + TTS	-	-	Syncs audio with visuals	English-only
[15] Kadam	SVS/MVS Techniques	0.75	-	Multi-view representation	Camera synchronization

Table 4: Comparative Analysis of Personalized Summarization Systems

Paper	**Technique**	**User Rating**	**Customization Level**	**Application Domain**	**Limitation**
[26] Kannan	Semantic Filtering	4.7 / 5	Scene-level personalization	Movies	Niche applicability
[21] Karousos	TextRank + GPT-4o Mini	4.2 / 5	Feedback-based summaries	Education	Context-specific
[9] Patil	Text2Graph + ChartGen	-	Graphical summarization	General-purpose	UI complexity

3.3 Hybrid and Multimodal Approaches

Recent advancements in video summarization integrate textual, visual, and audio data to pro- duce multimodal summaries that offer a holistic understanding of video content.

3.3.1 Multimodal Summarization

Multimodal models combine multiple data streams—such as transcripts, audio, and frames—to generate richer summaries. Table V summarizes the performance and capabilities of several such hybrid systems. These systems are promising for academic and educational domains, where semantic com- prehension is paramount. However, they require significant computational infrastructure for training and inference.

3.3.2 Deployment Interfaces and Tools

A number of works have emphasized the importance of intuitive interfaces and deployment environments. Table VI illustrates implementations using browser extensions, REST APIs, and NLP-as-a-service models.

These innovations demonstrate that model performance is only one component of a suc- cessful summarization system; usability and accessibility are equally critical in real-world ap- plications.

Table 5: Comparative Analysis of Hybrid Multimodal Summarization Techniques

Paper	**Components**	**Accuracy**	**Modality**	**Innovation**	**Resource Requirement**
[12]	SegNet +	81.86%	Visual + Text +	Recommendation-	24 GB VRAM

Jiang	VideoNet		Audio	based scoring	GPU
[1] Mandar	CNN-RNN	-	Visual + Text	Bidirectional sequence learning	High computational load
[16] Kumari	CNN + Transformers	0.59	Text + Visual	Downloadable format support	Output format limitations
[13] Ligade	LSA + T5	0.57	Text only	Dual summarization model	Multi-step complexity

Table 6: Comparative Analysis of Tools and Deployment Interfaces

Paper	**Technology**	**Notable Features**	**Accuracy**	**Use Case**	**Limitation**
[8] Dharmani	T5 + BART	Multilingual, Keywords	0.55	Language accessibility	Translation errors
[10] Verma	ASR + NLP	Audio summarization	0.53	Noisy environments	ASR accuracy issues
[2] Bandabe	Flask + Transformers	Lightweight web app	0.45	General video content	Only extractive summaries
[17] Siddhartha	BERT-BART	Evolution overview	0.61	Technical insights	Lacks novelty

Table 7: Comparative Analysis of Hybrid Multimodal Summarization Techniques

Paper	**Components**	**Accuracy**	**Modality**	**Innovation**	**Resource Requirement**
[12] Jiang	SegNet + VideoNet	81.86%	Visual + Text + Audio	Recommendation-based scoring	24 GB VRAM GPU
[1] Mandar	CNN-RNN	-	Visual + Text	Bidirectional sequence learning	High computational load
[16] Kumari	CNN + Transformers	0.59	Text + Visual	Downloadable format support	Output format limitations
[13] Ligade	LSA + T5	0.57	Text only	Dual summarization model	Multi-step complexity

Table 8: Comparative Analysis of Tools and Deployment Interfaces

Paper	**Technology**	**Notable Features**	**Accuracy**	**Use Case**	**Limitation**
[8] Dharmani	T5 + BART	Multilingual, Keywords	0.55	Language accessibility	Translation errors

[10] Verma	ASR + NLP	Audio summarization	0.53	Noisy environments	ASR accuracy issues
[2] Bandabe	Flask + Transformers	Lightweight web app	0.45	General video content	Only extractive summaries
[17] Siddhartha	BERT-BART	Evolution overview	0.61	Technical insights	Lacks novelty

4 Methodology

The system architecture comprises the following key components:

4.1 Video Download and Audio Extraction

The system uses yt-dlp, a modern YouTube downloader, to fetch the highest quality video and extract metadata including the thumbnail. Audio is extracted from the video using ffmpeg in .wav format (mono channel, 16kHz), which is optimal for transcription.

4.2 Speech-to-Text Transcription

The audio is transcribed using OpenAI's Whisper model (base variant) for efficient transcrip- tion on devices with limited memory. Whisper provides accurate multilingual and multi- speaker transcription capabilities.

4.3 Text Summarization

The transcribed text is summarized using the facebook/bart-large-cnn model. This transformer-based model performs abstractive summarization, meaning it rewrites content in- stead of extracting it verbatim. For long transcripts, text chunking is used to ensure token limits are not exceeded.

4.4 Key Concept Extraction

Using KeyBERT, a BERT-embedding-based keyword extractor, the summarized or original text is analyzed to extract meaningful keywords and phrases that represent core topics. This helps in indexing and navigating the content.

4.5 Key Frame Extraction and Scene Detection

The video is processed using OpenCV to detect scenes and extract keyframes. A frame dif- ference technique using grayscale conversion and cv2.absdiff measures the visual change between frames. When changes exceed a certain threshold, a scene change is detected.

4.6 Graph Detection

Each keyframe is checked for visual patterns resembling graphs using edge detection (Canny) and line detection (Hough Transform). If many straight lines are found in a frame, it is flagged as a probable graph.

4.7 PowerPoint Generation

All the collected insights—summary, key concepts, keyframes—are compiled into a Power- Point presentation using python-pptx. The final slides include:

- Title and concepts slide
- Summary slide
- Slides for each keyframe, with captions identifying graphs if present

4.8 Detailed Notes PDF Generation

To provide an additional format for content consumption, the system includes a feature to generate detailed PDF notes using the fpdf library. The transcript is first structured into clean, readable paragraphs and bullet points. Key concepts are listed with relevance scores, and the final PDF is divided into the following sections:

- **Key Topics Covered:** Bullet points of the top extracted keywords.
- **Content Breakdown:** Paragraph-wise formatted transcript using NLP chunking.
- **Analysis:** High-level interpretation of the most relevant concepts in the video.

5 System Architecture

6 Results

The system is tested on a range of YouTube videos including lectures, tutorials, and docu- mentaries. It successfully creates downloadable .pptx and .pdf summaries within minutes, offering:

- Accurate textual summaries
- Relevant keyword lists with relevance scores
- Visually descriptive keyframes
- Automatic graph slide detection

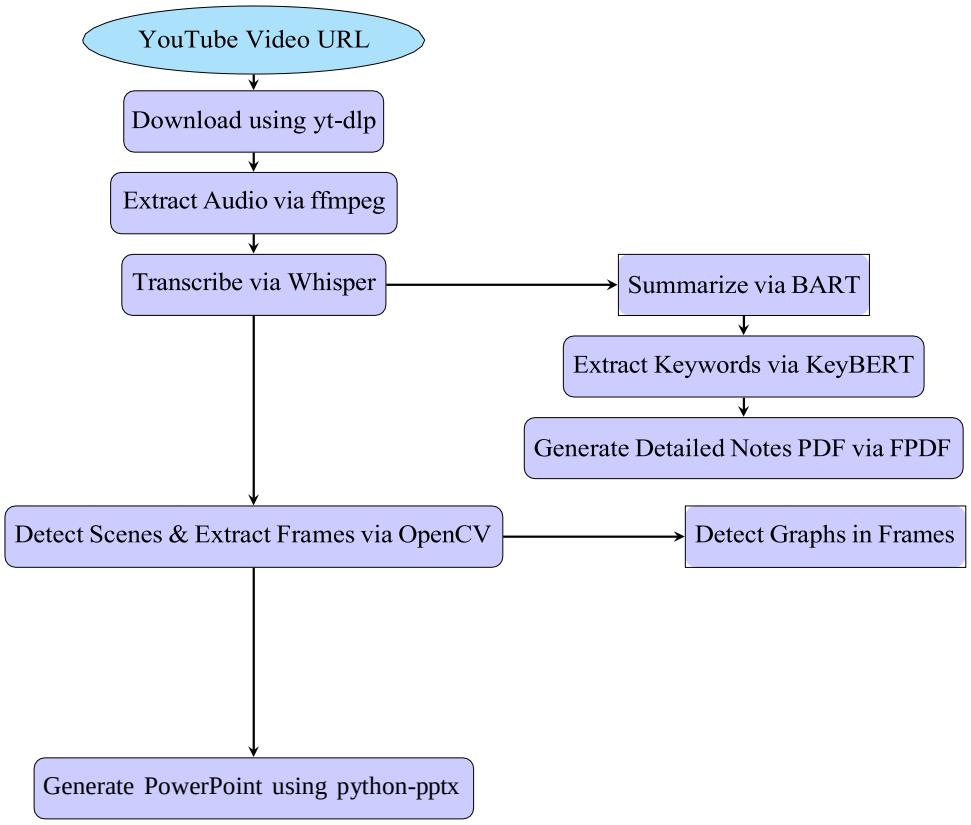

Figure 1: System Architecture

- Structured PDF notes ideal for revision and documentation

Users can view outputs in the GUI (Streamlit) and download them in multiple formats like PDF and PPT.

7 Advantages

- **No API Dependency:** Works fully offline once models are loaded.
- **Cross-Modal Analysis:** Merges audio, text, and visual data.
- **Streamlit Interface:** Easy-to-use GUI for end users.
- **Modular Design:** Components like transcription, summarization, and visualization are plug-and-play.
- **Multi-Format Output:** Users can download both PowerPoint slides and detailed notes in PDF, improving accessibility and usability.

8 Limitations & Future Work

- **Real-time Processing:** Current version processes after full download; future versions may

include streaming support.

• **Visual NLP Integration:** Extracting text from keyframes via OCR (e.g., equations, annotations) can be integrated.

• **Fine-Tuned Graph Classifier:** Train a custom classifier for better graph vs. non-graph distinction.

9 Conclusion

This research introduced a hybrid NLP–CV framework for YouTube video summarization and key concept extraction, combining Whisper, BART, KeyBERT, and OpenCV into a modular pipeline. The system produces accurate summaries, extracts key concepts, identifies graphs, and delivers outputs in both PowerPoint and PDF formats, offering strong applicability in e- learning and academic contexts. Compared with existing approaches, it stands out for its cross- modal integration, offline usability, and educational focus. Future work will target real-time streaming, OCR-based visual text extraction, and lightweight deployment to broaden accessi- bility.

References References

1. Puranik, G. M., Kamath, N., Akhadkar, N., & Dusane, G. (2024). YouTube Video Sum-
2. marizer: A Web Based Application for Concise Visual and Textual Summary. Journal of Analysis and Computation (JAC)
3. Bandabe, S., Zambre, J., Gosavi, P., Gupta, R., & Gaikwad, J. A. (2023). YouTube Tran- script Summarizer Using Flask. International Journal for Research in Applied Science and Engineering Technology (IJRASET), 11(IV)
4. Singh, A., & Singh, R. (2024). NLP for Text Summarization. International Journal of Progressive Research in Engineering Management and Science (IJPREMS), 4(5), 250–253
5. Inamdar, E., Kalaskar, V., & Zade, V. (2023). Survey Paper on YouTube Transcript Sum- marizer. International Research Journal of Modernization in Engineering, Technology and Science, 5(4)
6. Bhoir, V. S. (2025). YouTube Transcript Summarizer. International Research Journal of Modernization in Engineering, Technology and Science, 7(2)
7. Nethranand, P. S., Shruthi, K., Pavan Kalyan, N., & Akash, K. (2024). Summer-Riser: A novel abstractive based text summarization tool. International Advanced Research Journal in Science, Engineering and Technology, 11(5)
8. Vanisree, K., Reddy, K. S. S., & Ramyakapardini, M. (2025). YouTube Transcript Essence AI. International Journal of Creative Research Thoughts (IJCRT), 13(3)
9. Dharmani, A., Yadav, J., Shukla, J., & Rane, C. (2024). YouTube Transcript Summarizer. International Journal of Research Publication and Reviews, 5(3), 7035–7039
10. Patil, G., Chougule, A., Jadhav, K., & Toralkar, M. (2024). Improvised Clarity-cuts Text Summarization with Text to Graph Prediction. International Journal of Advanced Research in Innovative Ideas in Education, 10(3)
11. Verma, A. K., Balal, Md., & Salman. (2024). YouTube Transcript Summarizer. Journal of Emerging Technologies and Innovative Research (JETIR)

12. Varagantham, C., Reddy, J. S., Yelleni, U., Kotha, M., & Rao, P. V. (2022). Text Summa- rization Using NLP. Journal of Emerging Technologies and Innovative Research (JETIR)
13. Jiang, Y., Cui, K., Peng, B., & Xu, C. (2019). Comprehensive Video Understanding: Video Summarization with Content-Based Video Recommender Design. ICCV Workshop Pro- ceedings
14. Ligade, P., Wagh, M., Pisal, P., Gajare, K., & Agrawal, K. (2025). Automated Video Tran- script Summarizer. Journal of Emerging Technologies and Innovative Research (JETIR)
15. Charate, P. P., Patil, S. R., Pujari, A. B., Dhere, R. S., & Pednekar, C. B. (2025). Mul- tilingual Text Summarization Using NLP. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
16. Kadam, P., Vora, D., Mishra, S., Patil, S., Kotecha, K., Abraham, A., & Gabralla, L.(2022). Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms
17. Kumari, P. V., Keshava, M. C., Narendra, C., Akanksha, P., & Sravani, K. (2022). YouTube Transcript Summarizer Using Flask And NLP. Journal of Positive School Psychology
18. Siddhartha, Pandey, P., & Saxena, A. (2023). YouTube Transcript Summarizer. Interna- tional Journal of Research in Engineering and Science (IJRES), 11(5), 189–195
19. Yadav, S., Behra, A. K., Sahu, C. S., & Chandrakar, N. (2021). Summary and keyword extraction from YouTube video transcript. International Research Journal of Modernization in Engineering Technology and Science, 03(06), 895–902
20. Devi, S., Nadar, R., Nichat, T., & Lucas, A. (2022). Abstractive summarizer for YouTube videos. In Proceedings of the International Conference on Advances in Computing and Data Sciences (pp. 433–438)
21. Dilawari, A., & Khan, M. U. G. (2019). ASoVS: Abstractive Summarization of Video Sequences. IEEE Access, 7, 29253–29264
22. Karousos, N., Vorvilas, G., Pantazi, D., & Verykios, V. S. (2024). A Hybrid Text Sum- marization Technique of Student Open-Ended Responses to Online Educational Surveys.
23. Electronics, 13(18), 3722
24. Sanjana, R., Sai Gagana, V., Vedhavathi, K. R., & Kiran, K. N. (2021). Video Summariza- tion using NLP. International Research Journal of Engineering and Technology (IRJET), 8(8), 3672–3675
25. Fei, M., Jiang, W., & Mao, W. (2017). A novel compact yet rich key frame creation method for compressed video summarization. Multimedia Tools and Applications
26. Huang, S., Li, X., Zhang, Z., Wu, F., & Han, J. (2018). User-Ranking Video Summa- rization with Multi-Stage Spatio-Temporal Representation. IEEE Transactions on Image Processing, 27(12), 5793–5806
27. Li, J., Yao, T., Ling, Q., & Mei, T. (2017). Detecting shot boundary with sparse coding for video summarization. Neurocomputing, 266
28. Kannan, R., Ghinea, G., & Swaminathan, S. (2015). What do you wish to see? A summa- rization system for movies based on user preferences. Multimedia Tools and Applications, 74(24), 11357–11388