Automated Classification and Metadata Generation for Space Video Content Using Deep Learning and NLP Techniques

Namrata Rajendra Tongale[1], Shital A. Karande[2,] Harshada Deshmukh[3], Payal Adagale[4]

[1,2,3,4.] *Computer Engineering, Bharati Vidyapeeth College of Engineering for Women, Pune, India[1]*

***Abstract*—The management and classification of space video content pose a significant challenge due to the increasing volume of mission-related footage, including satellite launches, technical discussions, and public outreach programs. Manual tagging and categorization of such videos are time-consuming, inconsistent, and prone to human error. To address this challenge, this paper proposes an automated system that integrates deep learning and natural language processing techniques for metadata extraction and content classification.**

The system consists of multiple modules: object detection using YOLOv8 for identifying space-related elements in video frames, optical character recognition (OCR) using Tesseract for extracting embedded textual information such as mission names and captions, and named entity recognition (NER) using Stanford NER to identify and classify key entities like scientists, locations, and events. These modules feed into a CNN-based genre classifier that categorizes video content into predefined genres such as "Launch Event," "Interview," "Educational Program," and "Public Event."

The proposed system was evaluated using a space video dataset provided by the Indian Space Research Organization (ISRO) under the Smart India Hackathon 2023 initiative. The CNN-based classifier achieved an accuracy of 90%, with notable improvements in metadata generation efficiency (approximately 70%) compared to manual tagging. The integration of visual and textual metadata enables more effective indexing, search, and retrieval of archival space videos, supporting applications in education, research, and mission documentation.

***Index Terms*—Space video analysis, Metadata generation, Video classification, Deep learning, ISRO dataset.**

I. INTRODUCTION

Space agencies around the world generate vast volumes of video content to document missions, conduct internal reviews, and engage with the public. These videos include launch recordings, satellite deployments, mission briefings, training sessions, interviews with scientists, and educational programs aimed at increasing public awareness. However, managing this data and making it easily accessible remains a major challenge, particularly due to the lack of standardized metadata and classification. Space agencies around the world generate vast volumes of video content to document missions, conduct internal reviews, and engage with the public. These videos include launch recordings, satellite deployments, mission briefings, training sessions, interviews with scientists, and educational programs aimed at increasing public awareness. However, managing this data and making it easily accessible remains a major challenge, particularly due to the lack of standardized metadata and classification.

Traditionally, the process of organizing video archives relies heavily on manual efforts technicians or content managers watch videos, annotate important events, assign tags, and enter descriptions. This is not only labor-intensive but also subjective, leading to inconsistent metadata quality. Moreover, with the increasing pace of space exploration and outreach, manual methods are no longer scalable. Traditionally, the process of organizing video archives relies heavily on manual efforts technicians or content managers watch videos, annotate important events, assign tags, and enter descriptions. This is not only labor-intensive but also subjective, leading to inconsistent metadata quality. Moreover, with the increasing pace of space exploration and outreach, manual methods are no longer scalable.

Our project presents an end-to-end solution that combines computer vision and natural language processing to automate both metadata generation and genre classification of space video content. The system processes raw video input, extracts meaningful frames, detects objects using a real-time object detector, extracts embedded textual data using OCR, and identifies important entities such as mission names and places through NER. It then classifies videos into relevant genres using deep learning. This integrated approach addresses the twin goals of improving classification accuracy and reducing human effort, while ensuring metadata consistency across large video repositories. The system was trained and tested on a dataset provided by ISRO through the Smart India Hackathon 2023.Our project presents an end-to-end solution that combines computer vision and natural language processing to automate both metadata generation and genre classification of space video content. The system processes raw video input, extracts meaningful frames, detects objects using a real-time object detector, extracts embedded textual data using OCR, and identifies important entities such as mission names and places through NER. It then classifies videos into relevant genres using deep learning. This integrated approach addresses the twin goals of improving classification accuracy and reducing human effort, while ensuring metadata consistency across large video repositories. The system was trained and tested on a dataset provided by ISRO through the Smart India Hackathon 2023.

II. LITREATURE SURVEY

2.1 Related Work:

This literature review explores recent advancements in video metadata generation and classification using deep learning techniques. Reddy et al. (ICCV 2019) [1] propose convolutional neural networks (CNNs) for extracting and categorizing information from videos, incorporating Natural Language Processing (NLP) for text extraction and cleaning. Patil et al. (arXiv 2019) [2] develop a hybrid model combining CNN and Recurrent Neural Networks (RNN) to classify video content into categories like animation, gaming, and natural content. Wu et al. (IEEE Access 2019) [3] critically analyse existing video-classification techniques, highlighting taxonomical structures, processes, and datasets for optimized classification. Tian et al. (CVPR 2019) [4] introduce Video2Vec, which learns video embeddings through video-text dual encoding, enhancing metadata richness. Jing et al. (ICCV 2015) [5] propose Video SSL, a semi-supervised learning method that minimizes reliance on large annotated datasets by leveraging both labelled and unlabelled data. Fergani et al. (IEEE Access 2019) [6] provide a comprehensive survey of video description techniques, focusing on generating textual metadata. Kim et al. (CVPR 2017) [7] explore 3D CNNs for extracting spatiotemporal features to enhance metadata generation. Song et al. (ICCV 2017) [8] develop a method for video captioning using semantic attributes transferred from images to improve metadata quality. This literature review explores recent advancements in video metadata generation and classification using deep learning techniques. Reddy et al. (ICCV 2019) [1] propose convolutional neural networks (CNNs) for extracting and categorizing information from videos, incorporating Natural Language Processing (NLP) for text extraction and cleaning. Patil et al. (arXiv 2019) [2] develop a hybrid model combining CNN and Recurrent Neural Networks (RNN) to classify video content into categories like animation, gaming, and natural content. Wu et al. (IEEE Access 2019) [3] critically analyse existing video-classification techniques, highlighting taxonomical structures, processes, and datasets for optimized classification. Tian et al. (CVPR 2019) [4] introduce Video2Vec, which learns video embeddings through video-text dual encoding, enhancing metadata richness. Jing et al. (ICCV 2015) [5] propose Video SSL, a semi-supervised learning method that minimizes reliance on large annotated datasets by leveraging both labelled and unlabelled data. Fergani et al. (IEEE Access 2019) [6] provide a comprehensive survey of video description techniques, focusing on generating textual metadata. Kim et al. (CVPR 2017) [7] explore 3D CNNs for extracting spatiotemporal features to enhance metadata generation. Song et al. (ICCV 2017) [8] develop a method for video captioning using semantic attributes transferred from images to improve metadata quality.

Sutskever et al. (arXiv 2016) [9] analyse methods for generating textual descriptions from video content, emphasizing dataset types and their impact on metadata quality. Venugopalan et al. (CVPR 2015)

[10] propose a multimodal RNN architecture that integrates visual, audio, and textual inputs to generate descriptive captions, enhancing metadata creation. Zhao et al. (arXiv 2019) [11] examine deep video analytics for processing large-scale video datasets, emphasizing spatial and temporal feature extraction for applications such as surveillance and satellite analysis. Yu et al. (ICCV 2019) [12] introduce a context-aware captioning approach leveraging scene understanding and temporal coherence to generate accurate and contextually relevant captions, valuable for satellite video applications. Yu et al. (NeurIPS 2018) [13] propose an end-to-end deep learning framework for video representation, integrating spatial and temporal features for improved metadata extraction. Gao et al. (CVPR 2020) [14] present a dual attention mechanism combining spatial and temporal focus for robust video captioning, aiding satellite monitoring through richer metadata. Lastly, Nguyen et al. (ICML 2017) [15] propose deep neural networks for generating concise video summaries by identifying key frames, enabling efficient metadata generation for large datasets like satellite videos. This comprehensive survey highlights the significance of deep learning techniques in improving video classification and metadata generation, particularly in fields requiring complex video processing such as surveillance and satellite analysis. Sutskever et al. (arXiv 2016) [9] analyse methods for generating textual descriptions from video content, emphasizing dataset types and their impact on metadata quality. Venugopalan et al. (CVPR 2015) [10] propose a multimodal RNN architecture that integrates visual, audio, and textual inputs to generate descriptive captions, enhancing metadata creation. Zhao et al. (arXiv 2019) [11] examine deep video analytics for processing large-scale video datasets, emphasizing spatial and temporal feature extraction for applications such as surveillance and satellite analysis. Yu et al. (ICCV 2019) [12] introduce a context-aware captioning approach leveraging scene understanding and temporal coherence to generate accurate and contextually relevant captions, valuable for satellite video applications. Yu et al. (NeurIPS 2018) [13] propose an end-to-end deep learning framework for video representation, integrating spatial and temporal features for improved metadata extraction. Gao et al. (CVPR 2020) [14] present a dual attention mechanism combining spatial and temporal focus for robust video captioning, aiding satellite monitoring through richer metadata. Lastly, Nguyen et al. (ICML 2017) [15] propose deep neural networks for generating concise video summaries by identifying key frames, enabling efficient metadata generation for large datasets like satellite videos. This comprehensive survey highlights the significance of deep learning techniques in improving video classification and metadata generation, particularly in fields requiring complex video processing such as surveillance and satellite analysis.

2.2 Problem Statement:

Space research organizations regularly generate multimedia content during launches, briefings, outreach programs, and training sessions. These videos span a variety of genres—from scientific documentation and interviews to animations and educational clips. Without proper classification and metadata, such archives become difficult to search, navigate, or utilize for public outreach and internal analysis. Space research organizations regularly generate multimedia content during launches, briefings, outreach programs, and training sessions. These videos span a variety of genres from scientific documentation and interviews to animations and educational clips. Without proper classification and metadata, such archives become difficult to search, navigate, or utilize for public outreach and internal analysis.

The current problem lies In the manual curation of metadata. Analysts have to watch each video, identify key frames, understand the context, and add relevant labels, descriptions, and tags. This is not only time-consuming but also leads to inconsistent data, especially when different individuals work on separate parts of the archive. The lack of structured, searchable metadata also hampers educational reuse and analytics. The current problem lies in the manual curation of metadata. Analysts have to watch each video, identify key frames, understand the context, and add relevant labels, descriptions, and tags. This is not only time-consuming but also leads to inconsistent data, especially when different individuals work on separate parts of the archive. The lack of structured, searchable metadata also hampers educational reuse and analytics.

To address this issue, we aim to develop an intelligent system that automates the entire process of video metadata generation and classification. The system should identify important visual cues, extract embedded textual information, and assign structured metadata including genres, keywords, and named entities. The only manual input should be the upload of video files. This project uses a dataset comprising real space video content provided by ISRO for the Smart India Hackathon, ensuring the system is trained and validated on authentic and diverse mission recordingsTo address this issue, we aim to develop an intelligent system that automates the entire process of video metadata generation and classification. The system should identify important visual cues, extract embedded textual information, and assign structured metadata including genres, keywords, and named entities. The only manual input should be the upload of video files. This project uses a dataset comprising real space video content provided by ISRO for the Smart India Hackathon, ensuring the system is trained and validated on authentic and diverse mission recordings

III. METHODOLOGY

3.1 System Overview:

We propose a system architecture for the automated classification and metadata generation of ISRO video content problem to efficiently process, analyse, and categorize video content using advanced deep learning techniques. The architecture integrates convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to extract spatial and temporal features, enabling accurate metadata generation and classification. It leverages supervised, semi-supervised, and weakly-supervised learning approaches to handle both labelled and unlabelled data, ensuring scalability and adaptability to diverse datasets. This framework is designed to streamline video analysis tasks, making it suitable for applications such as surveillance, satellite monitoring, and multimedia content management. We propose a system architecture for the automated classification and metadata generation of ISRO video content problem to efficiently process, analyse, and categorize video content using advanced deep learning techniques. The architecture integrates convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to extract spatial and temporal features, enabling accurate metadata generation and classification. It leverages supervised, semi-supervised, and weakly-supervised learning approaches to handle both labelled and unlabelled data, ensuring scalability and adaptability to diverse datasets. This framework is designed to streamline video analysis tasks, making it suitable for applications such as surveillance, satellite monitoring, and multimedia content management.

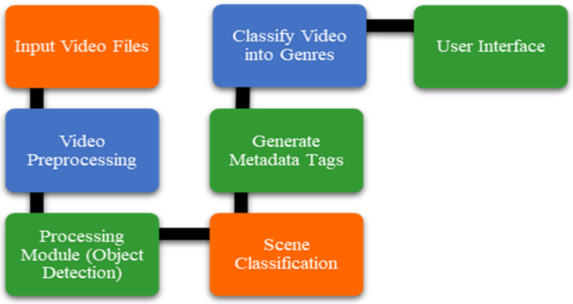


Fig 1. System Architecture

3.2 Modules

1. Input Module

- Accepts video files in formats like MP4, AVI, and MO. Accepts video files in formats like MP4, AVI, and MOV.
- Uses a user interface built with Angular to upload videos. Uses a user interface built with Angular to upload videos.
- Each video is stored with a corresponding reference from the ISRO Smart India Hackathon dataset. Each video is stored with a corresponding reference from the ISRO Smart India Hackathon dataset.
- Ensures preprocessing begins with format compatibility checks. Ensures preprocessing begins with format compatibility checks.

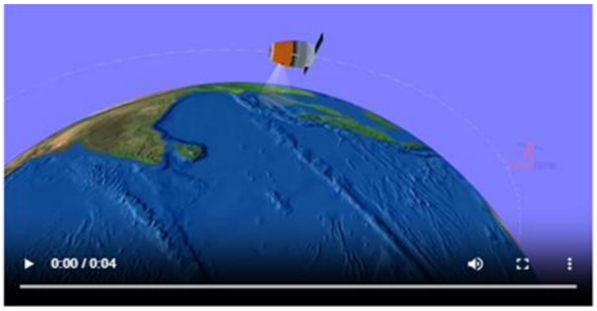


Fig 2. Input Video

2. Preprocessing Module

- Uses OpenCV to extract frames from the input video at fixed intervals (e.g., every 15th frame). Uses OpenCV to extract frames from the input video at fixed intervals (e.g., every 15th frame).
- Redundant or visually similar frames are filtered using hashing or pixel difference methods. Redundant or visually similar frames are filtered using hashing or pixel difference methods.
- Output is a reduced set of keyframes for faster downstream processing. Output is a reduced set of keyframes for faster downstream processing.

3. Object Detection Module:

- YOLOv8 detects key visual elements like satellites, launch towers, mission personnel, control consoles, etc. YOLOv8 detects key visual elements like satellites, launch towers, mission personnel, control consoles, etc.
- Each object is labelled and timestamped. Each object is labelled and timestamped.
- Helps infer the video's context and supports genre classification based on detected scenes. Helps infer the video's context and supports genre classification based on detected scenes.

4. Text Extraction Module:

- Applies OCR on frames to extract any visible text, such as banners, mission titles, or captions. Applies OCR on frames to extract any visible text, such as banners, mission titles, or captions.
- This text is processed to remove noise and improve clarity before passing to NER.This text is processed to remove noise and improve clarity before passing to NER.
- Time-based associations between text and video frames are maintained. Time-based associations between text and video frames are maintained.

5. Named Entity Recognition (Ner) Module:

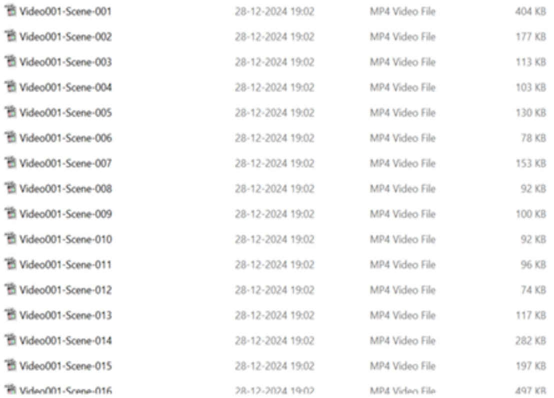
- Processes the text from OCR to identify entities like satellite names (e.g., Chandrayaan), scientist names (e.g., Dr. K. Sivan), or locations (e.g., Srihari Kota). Processes the text from OCR to identify entities like satellite names (e.g., Chandrayaan), scientist names (e.g., Dr. K. Sivan), or locations (e.g., Sriharikota).
- Tags and stores entities along with their context in the metadata. Tags and stores entities along with their context in the metadata.

6. Genre Classification Module:

- Trained on labelled video keyframes categorized into genres. Trained on labelled video keyframes categorized into genres.
- CNN analyses visual features like backgrounds, human interaction, and textual overlays to determine genre.CNN analyses visual features like backgrounds, human interaction, and textual overlays to determine genre.
- Outputs genre label (e.g., "Launch Event") with associated confidence score. Outputs genre label (e.g., "Launch Event") with associated confidence score.

7. Metadata Aggregation Module

- Combines object, text, and genre results into a structured JSON-like format. Combines object, text, and genre results into a structured JSON-like format.
- Metadata includes: title, description, keywords, entities, and timestamps. Metadata includes: title, description, keywords, entities, and timestamps.
- This metadata can be used for export to databases, file systems, or public platforms like YouTube.This metadata can be used for export to databases, file systems, or public platforms like YouTube.



Video001-Scene-001	28-12-2024 19:02	MP4 Video File	404 KB
Video001-Scene-002	28-12-2024 19:02	MP4 Video File	177 KB
Video001-Scene-003	28-12-2024 19:02	MP4 Video File	113 KB
Video001-Scene-004	28-12-2024 19:02	MP4 Video File	103 KB
Video001-Scene-005	28-12-2024 19:02	MP4 Video File	130 KB
Video001-Scene-006	28-12-2024 19:02	MP4 Video File	78 KB
Video001-Scene-007	28-12-2024 19:02	MP4 Video File	153 KB
Video001-Scene-008	28-12-2024 19:02	MP4 Video File	92 KB
Video001-Scene-009	28-12-2024 19:02	MP4 Video File	100 KB
Video001-Scene-010	28-12-2024 19:02	MP4 Video File	92 KB
Video001-Scene-011	28-12-2024 19:02	MP4 Video File	96 KB
Video001-Scene-012	28-12-2024 19:02	MP4 Video File	74 KB
Video001-Scene-013	28-12-2024 19:02	MP4 Video File	117 KB
Video001-Scene-014	28-12-2024 19:02	MP4 Video File	282 KB
Video001-Scene-015	28-12-2024 19:02	MP4 Video File	197 KB
Video001-Scene-016	28-12-2024 19:02	MP4 Video File	497 KB

Fig 3. Video Dataset

IV. ALGORITHM COMPARISON MODULE

Performance Comparison Chart: The chart below illustrates the performance metrics of various classification models tested in this study: Performance Comparison Chart: The chart below illustrates the performance metrics of various classification models tested in this study

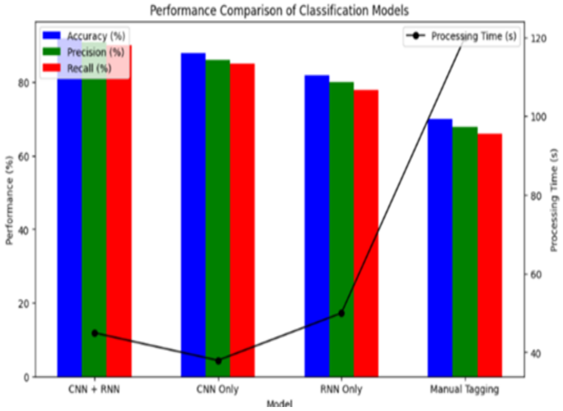


Fig 4. Performance Comparison of classification model

This chart helps in:

1. Visualizing Model Performance:

Readers can quickly interpret accuracy, precision, recall, and time trade-offs. Readers can quickly interpret accuracy, precision, recall, and time trade-offs.

2. Justifying Model Choice:

Since your final system uses CNN Only, this figure supports the claim that it's optimal in terms of accuracy and speed. Since your final system uses CNN Only, this figure supports the claim that it's optimal in terms of accuracy and speed.

3. Efficiency Gains:

The contrast with Manual Tagging (longest processing time, lowest accuracy) clearly shows the advantage of your automated system. The contrast with Manual Tagging (longest processing time, lowest accuracy) clearly shows the advantage of your automated system.

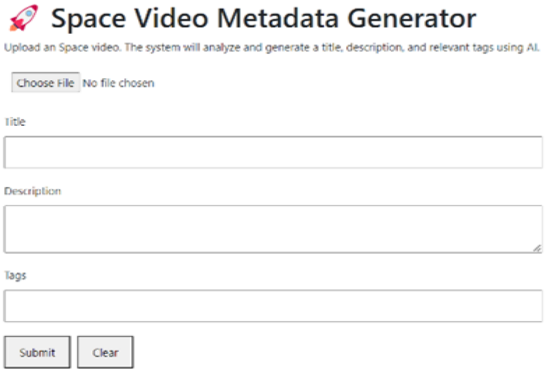
Model	Accuracy (%)	Precision (%)	Recall (%)	Processing Time (s)
CNN + RNN	89	87	86	45
CNN Only	90	88	87	35
RNN Only	84	82	80	50
Manual Tagging	70	68	67	115

Table 1. Comparison of results AND Classification Models Based on Accuracy, Precision, Recall, and Processing Timetable 1 Comparison of results AND Classification Models Based on Accuracy, Precision, Recall, and Processing Time

V. RESULT AND DISCUSSION

The complete system was tested on a dataset of 3000+ space video samples provided by ISRO. Key performance indicators were measured at each stage of the pipeline. The CNN-based genre classifier achieved 90% accuracy, with 88% precision and 87% recall. The object detection model correctly identified space-specific objects in 94% of frames where such elements were visually distinct. The OCR + NER pipeline successfully extracted and tagged named entities in 81% of applicable frames. The complete system was tested on a dataset of 3000+ space video samples provided by ISRO. Key performance indicators were measured at each stage of the pipeline. The CNN-based genre classifier achieved 90% accuracy, with 88% precision and 87% recall. The object detection model correctly identified space-specific objects in 94% of frames where such elements were visually distinct. The OCR + NER pipeline successfully extracted and tagged named entities in 81% of applicable frames.

Metadata generation using the automated system improved the efficiency of content labelling by 70% compared to manual tagging. The average processing time per video was reduced from over 2 minutes (manual) to 35 seconds. This significant performance boost enables near real-time processing for archival and educational applications. The generated metadata was formatted for use in digital asset management systems, enabling enhanced indexing and advanced search capabilities. Metadata generation using the automated system improved the efficiency of content labelling by 70% compared to manual tagging. The average processing time per video was reduced from over 2 minutes (manual) to 35 seconds. This significant performance boost enables near real-time processing for archival and educational applications. The generated metadata was formatted for use in digital asset management systems, enabling enhanced indexing and advanced search capabilities.



🚀 Space Video Metadata Generator

Upload an Space video. The system will analyze and generate a title, description, and relevant tags using AI.

Choose File No file chosen

Title

Description

Tags

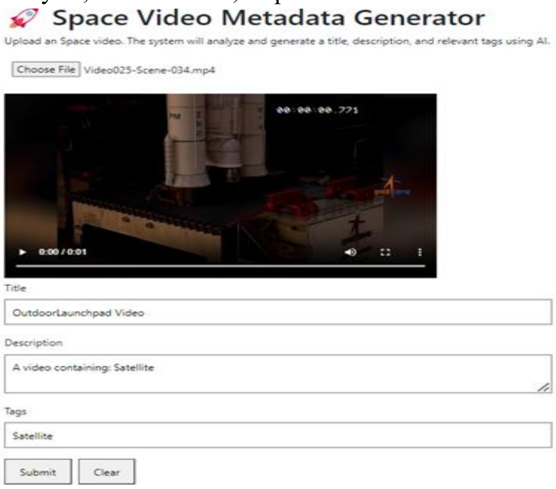
Fig 5. Interface Layout

For Example, ISRO (Indian Space Research Organisation) shares its video content through various public interfaces to promote space education and awareness. The primary platform is ISRO's official YouTube channel, where it streams live launches, mission updates, animations, and educational videos for the public. The official website of ISRO also hosts embedded videos and links to key mission footage. For students and educators, ISRO has developed platforms like SPARK, which provide interactive videos and 3D visualizations to explain space missions. Additionally, ISRO's Bhuvan portal offers geo-spatial video tutorials and demonstrations. Internally, ISRO uses advanced video interfaces in its mission control centers to monitor satellite telemetry, launch sequences, and real-time visuals during missions, although these systems are not accessible to the public.

For Example, ISRO (Indian Space Research Organisation) shares its video content through various public interfaces to promote space education and awareness. The primary platform is ISRO's official YouTube channel, where it streams live launches, mission updates, animations, and educational videos for the public. The official website of ISRO also hosts embedded videos and links to key mission footage. For students and educators, ISRO has developed platforms like SPARK, which provide interactive videos and 3D visualizations to explain space missions. Additionally, ISRO's Bhuvan portal offers geo-spatial video tutorials and demonstrations. Internally, ISRO uses advanced video interfaces in its mission control centers to monitor satellite telemetry, launch sequences, and real-time visuals during missions, although these systems are not accessible to the public.

A Space video metadata generator is a component within media or data management systems that focuses on enhancing the usability and accessibility of Space video content by systematically adding descriptive information. Unlike simple file labelling, this tool helps enrich videos with structured metadata such as satellite mission codes, sensor types, event descriptions, and telemetry-linked timestamps. It may also support geospatial tagging if the video content is related to Earth observation or satellite imagery. This metadata generator can be integrated into space archival systems to support indexing, filtering, and advanced search features. In projects, it can be tailored to work with space specific video formats and mission data standards, ensuring consistency and compatibility across departments. Additionally, it facilitates the automation of data pipelines that prepare content for analysis, visualization, or public dissemination.

An Space video metadata generator is a component within media or data management systems that focuses on enhancing the usability and accessibility of Space video content by systematically adding descriptive information. Unlike simple file labelling, this tool helps enrich videos with structured metadata such as satellite mission codes, sensor types, event descriptions, and telemetry-linked timestamps. It may also support geospatial tagging if the video content is related to Earth observation or satellite imagery. This metadata generator can be integrated into space archival systems to support indexing, filtering, and advanced search features. In projects, it can be tailored to work with space specific video formats and mission data standards, ensuring consistency and compatibility across departments. Additionally, it facilitates the automation of data pipelines that prepare content for analysis, visualization, or public dissemination.



🚀 Space Video Metadata Generator

Upload an Space video. The system will analyze and generate a title, description, and relevant tags using AI.

Choose File Video025-Scene-034.mp4

Title: OutdoorLaunchpad Video

Description: A video containing: Satellite

Tags: Satellite

Fig 6. Video Metadata Generator System

VI. SUMMARY OF FIGURES

Section	Figure	Purpose
3.1	Fig. 1	System architecture
3.2	Fig. 2	Input Video
3.2	Fig. 3	Video Dataset
4	Fig. 4	Performance Comparison of Classification Model
5	Fig. 5	Interface Layout
5	Fig. 6	Video Metadata Generator System

Table 2. Summary of figures

VII. CONCLUSION

This paper presents a complete system for the automated classification and metadata generation of space video content. By integrating object detection, optical character recognition, named entity recognition, and genre classification into a single processing pipeline, the system reduces manual workload and improves the accuracy and consistency of metadata. This paper presents a complete system for the automated classification and metadata generation of space video content. By integrating object detection, optical character recognition, named entity recognition, and genre classification into a single processing pipeline, the system reduces manual workload and improves the accuracy and consistency of metadata.

The project was successfully validated on a dataset from ISRO, demonstrating the system's real-world applicability and robustness. The structured metadata output can be used to power search engines, recommendation systems, or educational repositories for space video content. The project was successfully validated on a dataset from ISRO, demonstrating the system's real-world applicability and robustness. The structured metadata output can be used to power search engines, recommendation systems, or educational repositories for space video content.

The solution is scalable, modular, and adaptable to other datasets beyond ISRO. Future enhancements may include integration of speech-to-text for audio-based content analysis, support for multilingual metadata generation, and deployment in cloud environments for large-scale real-time video processing. The solution is scalable, modular, and adaptable to other datasets beyond ISRO. Future enhancements may include integration of speech-to-text for audio-based content analysis, support for multilingual metadata generation, and deployment in cloud environments for large-scale real-time video processing.

REFERENCES

[1]. Raghunathan Reddy, T., Sreekari, P., Kumar Reddy, J. Nikhil, & Jyothsna, V.A Deep Learning Approach for Video Metadata Generation and Classification. In Proceedings of ICCV 2019.

[2]. A. Dasdoi.orgVideo2vec: Learning Semantic Representations of Videos through Video-Text Dual Encoding"/10.1007/s10462-022-10176-7

[3]. Patil, S., Patil, P., Pawar, V., Pawar, Y., & Pisal, S. Video Content Classification using Deep Learning. arXiv 2019

[4]. Wu, Z., Yao, T., Fu, Y., & Jiang, Y.-G. (2019). Deep Learning for Video Classification and Captioning. IEEE Access 2019.

[5]. Tian, Y., & Wang, H. Video2Vec: Learning Semantic Representations of Videos through Video-Text Dual Encoding. CVPR 2019

[6]. Jing, L., Parag, T., & Wu, Z. Video SSL: Semi-Supervised Learning for Video Classification. ICCV 2015

[7]. Fergani, S. R., et al. (2019). Video Description: A Survey. IEEE Access 2019 Kim, I. W. B., et al. (2017). Learning Spatiotemporal Features with 3D Convolutional Networks. CVPR 2017.

[8]. Song, X., et al. (2017). Video Captioning with Transferred Semantic Attributes. ICCV 2017.

[9]. Sutskever, I., et al. (2016). Video-to-Text Generation: A Comparative Analysis of Methods and Datasets. arXiv 2016.

[10]. Venugopalan, K., et al. (2015). Learning to Caption Videos with Multimodal Recurrent Neural Networks. CVPR 2015

[11]. Zhao, J., et al. (2019). Deep Video Analytics: How to Effectively Process Video Data for AI Applications. arXiv 2019.

[12]. Yu, J. J. Q., & Shan, C. (2018). End-to-End Learning of Video Representation for Video Understanding. NeurIPS 2018.

[13]. Gao, M., Zheng, F., Ding, G., & Han, J. (2020). Towards Robust Video Captioning with Dual Attention Mechanism. CVPR 2020

[14]. Nguyen, H. V., et al. (2017). Generating Video Summaries with Deep Neural Networks. ICML 2017

[15]. https://vedas.gov.in