Usha Dhankar, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

End-to-End CNN-Based System for Human Detection in Fire Scenes Deployed via a Flask Web Application

Usha Dhankar, Komal Khatak, Dr. Sweety
Computer Science and Engineering Department Puran Murti Vidyapeeth Sonipat, India

Abstract- The exponential growth of video data across domains such as surveillance, aerospace, and digital media has created a significant challenge in efficient content retrieval. Traditional approaches based on manual tagging and low-level visual features fail to capture the contextual semantics of video content. This paper proposes a semantic video discovery framework that integrates deep feature fusion with automated metadata generation. Visual features are extracted using deep learning models such as YOLO and Segment Anything Model (SAM), while textual features are derived using Natural Language Processing (NLP) techniques including FastText and Named Entity Recognition (NER). The fusion of visual and textual embeddings enables context-aware retrieval and improves semantic understanding of video content. Experimental results demonstrate enhanced accuracy, precision, and retrieval efficiency compared to traditional methods.

Keywords— Semantic Video Retrieval, Deep Feature Fusion, Metadata Generation, YOLO, SAM, NLP.

I. INTRODUCTION

The rapid growth of video content across digital platforms, surveillance systems, and research organizations has made efficient video retrieval a critical challenge. Traditional approaches rely heavily on manual tagging and metadata generation, which are time-consuming, inconsistent, and not scalable.

Recent advancements in deep learning, computer vision, and natural language processing have enabled intelligent systems capable of extracting semantic meaning from video data. Techniques such as object detection, segmentation, and multi-modal learning have significantly improved video understanding [1]–[3].

In particular, models like YOLO enable real-time object detection [4], while SAM provides high-precision segmentation capabilities [5]. Additionally, NLP models such as FastText and transformer-based approaches improve semantic interpretation of textual data [6], [7].

This paper proposes a semantic video discovery framework that integrates deep visual features with textual metadata to enable efficient and context-aware retrieval.

II. EASE OF USE

A. User-Friendly Query Interface

The proposed semantic video discovery system is designed to provide a simple and intuitive interface for users. Unlike traditional video retrieval methods that depend on manual browsing or predefined metadata, the system allows users to input queries in natural language. This eliminates the need for technical expertise or structured query formats.

The integration of Natural Language Processing (NLP) techniques enables the system to interpret user intent effectively and extract meaningful information from queries. As a result, users can interact with the system in a more natural and efficient manner, improving accessibility across diverse user groups.

B. Automated Processing and Retrieval Efficiency

The system incorporates a fully automated processing pipeline that handles object detection, segmentation, and metadata generation without requiring manual intervention. Deep learning models such as YOLO and SAM are used for visual feature extraction, while NLP-based techniques are applied for generating contextual metadata.

This automation significantly reduces human effort and ensures consistency in video analysis. Additionally, the system provides precise retrieval by mapping user queries directly to relevant video segments. Users can quickly access the required content without navigating through entire videos, thereby reducing search time and computational complexity.

Overall, the system enhances ease of use by improving efficiency, reducing manual workload, and delivering accurate and context-aware results.

III. RELATED WORK

Recent advancements in video understanding and retrieval have been largely driven by deep learning, computer vision, and natural language processing techniques. Researchers have explored multiple approaches to improve semantic

understanding of video data by leveraging visual, textual, and multi-modal features.

Deep learning-based object detection models such as YOLO have demonstrated high efficiency and real-time performance in detecting objects within video frames [4]. These models are widely used due to their speed and accuracy in identifying multiple objects simultaneously. However, object detection alone is insufficient for capturing detailed contextual information within a scene.

To address this limitation, segmentation techniques such as the Segment Anything Model (SAM) have been introduced, which provide pixel-level accuracy and enable precise localization of objects [5]. Segmentation enhances visual representation by focusing on meaningful regions within frames, thereby improving feature extraction quality.

In parallel, Natural Language Processing (NLP) techniques have been applied for metadata generation and video classification. Transcript-based models and embedding techniques such as FastText and transformer-based architectures improve semantic interpretation and enable structured metadata extraction [6], [8]. These approaches facilitate better indexing and retrieval by converting unstructured textual data into meaningful representations.

Furthermore, multi-modal learning frameworks have been proposed to combine visual and textual information for improved semantic understanding. Models such as CLIP and VideoCLIP align image/video data with textual descriptions in a shared embedding space, enabling more accurate retrieval and zero-shot capabilities [9], [10]. Despite these advancements, most existing systems either focus on visual features or textual metadata independently or do not effectively integrate both modalities.

Additionally, many approaches lack efficient and scalable metadata generation mechanisms for large-scale video datasets, particularly in domains such as surveillance and aerospace, where videos are long and complex.

To overcome these limitations, the proposed work introduces a unified framework that integrates deep feature fusion with automated metadata generation. By combining object detection, segmentation, and NLP-based semantic analysis, the system enables context-aware video retrieval with improved accuracy and efficiency.

Table I Comparative Analysis Of Existing Approaches

Model / Approach	Technique Used	Strengths	Limitations	Relevance to Proposed Work
YOLO [4]	Object Detection (CNN-based)	Real-time detection, high speed	No semantic understanding	Used for visual feature extraction
SAM [5]	Image Segmentation	Pixel-level accuracy, precise localization	No textual/contextual understanding	Enhances region- based feature extraction
FastText / NLP Models [6], [8]	Text Processing & Embedding	Efficient metadata generation	No visual integration	Used for textual feature extraction
CLIP / VideoCLIP [9], [10]	Multi-modal Embedding	Strong visual-text alignment	Limited fine-grained retrieval	Basis for multi-modal fusion
Traditional Metadata Systems	Manual Tagging	Simple implementation	Time-consuming, inconsistent	Motivation for automation

IV. PROPOSED METHODOLOGY

A. System Overview

The proposed system in figure 1, integrates computer vision and NLP techniques to perform semantic video discovery. The workflow includes video input, feature extraction, metadata generation, feature fusion, and retrieval.

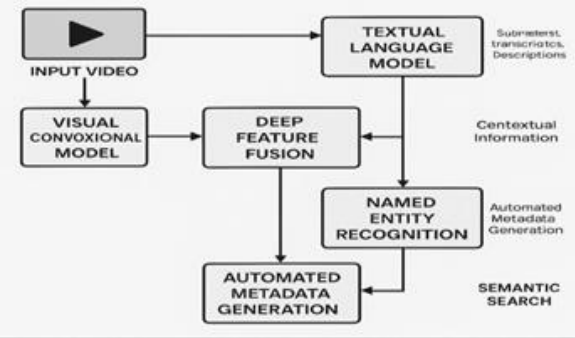

Fig 1. Semantic Video Discovery Framework integrating deep visual feature extraction, contextual text processing, multi-modal feature fusion, and metadata generation for video retrieval

B. Visual Feature Extraction

Visual features are extracted using:

- YOLO for object detection [4]
- SAM for segmentation [5]

These models identify objects and regions of interest within video frames, improving visual understanding.

C. Textual Feature Extraction

Textual data such as subtitles and transcripts are processed using NLP techniques. FastText embeddings [6] and transformer-based models [7] are used to generate contextual representations.

D. Metadata Generation

Named Entity Recognition (NER) is applied to extract entities such as objects, locations, and events from textual

data. This enables automatic metadata generation for efficient indexing and retrieval.

E. Multi-Modal Feature Fusion

The visual and textual embeddings are fused to create a unified representation of video content. This multi-modal representation enables semantic search and improves retrieval accuracy.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

The proposed semantic video discovery system was evaluated using a custom dataset obtained from ISRO, comprising more than 3500 images organized into multiple semantic categories. The dataset reflects real-world complexity, including diverse objects, scenes, and contextual variations, making it suitable for validating the robustness and scalability of the proposed framework.

During experimentation, video data was processed through multiple stages, including frame extraction, object detection using YOLO, segmentation using SAM, and metadata generation using NLP techniques. The integration of these components enabled the creation of a unified multi-modal representation for efficient retrieval.

B. Performance Metrics

To evaluate the effectiveness of the proposed system, the following standard performance metrics were used [3]:

- Accuracy: Measures the overall correctness of classification and retrieval results.
- Precision: Evaluates the proportion of relevant results among retrieved outputs. $\text{Precision} = \frac{TP}{TP+FP}$
- Recall: Measures the system's ability to retrieve all relevant instances. $\text{Recall} = \frac{TP}{TP+FN}$
- Mean Average Precision (mAP): Provides a comprehensive evaluation of detection and retrieval performance across all classes.

These metrics collectively assess both the correctness and completeness of the retrieval process.

C. Results and Analysis

The experimental results demonstrate that the proposed system outperforms traditional video retrieval approaches that rely on manual tagging or single-modal analysis. The integration of deep feature fusion significantly enhances semantic understanding by combining visual and textual information into a unified representation.

The training and validation graphs presented in Figure 2 show a consistent increase in accuracy and a gradual reduction in loss values, indicating effective model convergence and learning . Additionally, the Mean Average Precision (mAP) values confirm that the system achieves reliable object detection and classification performance across multiple categories.

Table II Results Metrics

Parameters	MaP	Precision	Recall
Results	87.2%	78.3%	81.6%

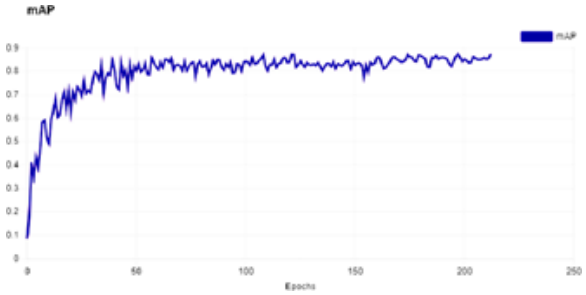

Fig. 2 Mean Accuracy Precision - Training Graph

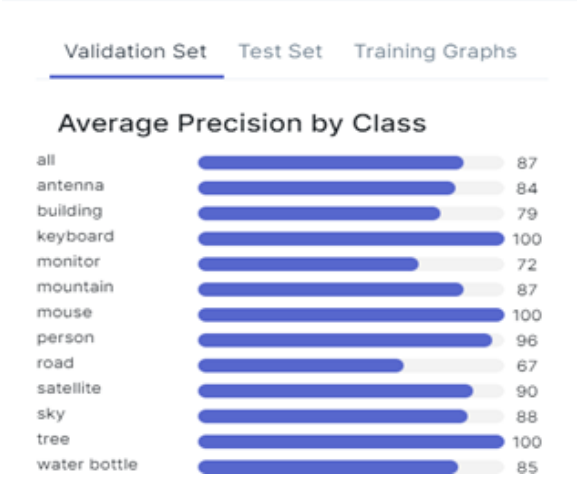

Fig. 3 The average precision per class in the validation dataset is presented, delineating specific performance metrics for individual classes.

Object Detection Losses: Box, Class, and Object

In object detection, three primary losses drive the optimization process:

Box Loss:

- Measures how well the predicted bounding box overlaps the ground truth bounding box.
- Commonly uses Intersection over Union (IoU) as the metric [16].
- Higher IoU indicates better localization and lower box loss.
- Represented mathematically by functions like Smooth L1 or GIoU loss.

Class Loss:

- Measures how accurately the model predicts the class of the object within the bounding box.
- Typically uses cross-entropy loss [16][17].
- Lower the cross-entropy, the more confident the model is in its class prediction.

Object Loss:

- Determines whether an object exists within a specific region of interest (anchor box) [16][17].
- Helps the model differentiate between background and foreground regions.
- Often uses binary cross-entropy loss.
- Lower object loss indicates the model is confident when an object is present or absent.
- These losses are combined with weights to form the total loss function used to train the object detector. The weights are adjusted to prioritize specific aspects of the detection task, such as accurate localization or precise classification.

Fig 4. Displays training Graphs loss(s) depicting a perfect rectangular hyperbolic shape, emphasizing the precise accuracy achieved in the dataset generation process

Fig. 4 illustrates the training behavior of the proposed object detection model through Object Loss, Box Loss, and Class Loss across epochs. It can be observed that all three loss components show a consistent decreasing trend, indicating effective learning and convergence of the model. The Box Loss decreases steadily, reflecting improved localization of bounding boxes. The Class Loss also reduces significantly, demonstrating accurate classification of detected objects. Similarly, the Object Loss shows stabilization, indicating that the model effectively distinguishes between object and background regions.
Overall, the smooth decline and stabilization of these loss curves confirm that the model has been successfully trained and achieves reliable detection performance.

OUTPUT :- (Object Detection)

output_from_yolo_cleaned_: ultralytics python-3.10.12 2.1.0+cu118 cuda:0 tesla t4 15102mib yolov8s summary fused 168 layers 11156544 parameters 0 gradients 28.6 gflops image 1/1 /root/jpg.jpg 800x800 ***3***

persons 3 bottles 2 chairs 1 dining table 2 mouses 3 cell phones 2 books 25.0ms

OUTPUT :- (Object Detection & NLP Integration)

Detected Objects Information:

persons bottles chairs dining table mouses cell phones books

FastText Results:-

(('label indoorLab', 'label indoorGeneric"), ' label crowd"),
array([0.96146, 0.73579, 0.026461]))

Thus the model has the highest confidence that the video source provided belongs to the "Indoor Lab" category, that is, ~96%. Also, the model is confident that it belong to ~73% to "Indoor Generic" and ~26% in "Crowd"

Fig. 5. Before training the frame

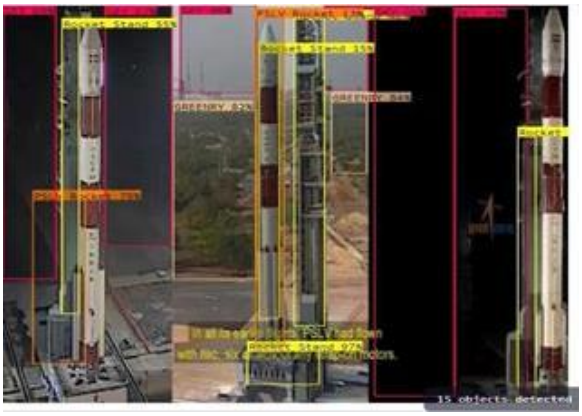

Fig. 6. After training the frame

Fig. 5 illustrates the initial state of the system before model training. In this stage, the model is unable to accurately detect or classify objects within the video frames. The absence of bounding boxes and labels indicates that the system has not yet learned meaningful visual features. This highlights the need for training using deep learning models to enable effective object recognition and scene understanding and Fig. 6 presents the output after model training, where the system successfully detects and localizes multiple objects within the video frames. The presence of bounding boxes and class labels (e.g., rocket components, structures) demonstrates that the model has effectively learned visual patterns from the dataset. This confirms improved detection accuracy and validates the

effectiveness of the proposed deep learning-based approach.

D. Key Findings

The proposed system improves retrieval accuracy, reduces manual effort through automated metadata generation, and enables efficient multi-modal video understanding, making it suitable for large-scale real-world applications.

VII. CONCLUSION

This paper presents a semantic video discovery framework that integrates deep feature fusion and automated metadata generation.

The system successfully improves video retrieval by combining visual and textual features, enabling context- aware search and efficient indexing. The proposed approach is scalable and suitable for real-world applications such as surveillance, aerospace analytics, and digital content management.

REFERENCES

1. A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision (CLIP)," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 6789– 6802, 2023
2. G. Bertasius et al., "Is Space-Time Attention All You Need for Video Understanding?" ICML, 2021.
3. X. Wang, L. Zhu and Y. Yang, "Multi-Modal Video Retrieval via Deep Semantic Alignment," in Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 2022, pp. 1–6.
4. Jana, A. P., Biswas, A., & Mohana. (2018). YOLO based Detection and Classification of Objects in video records. In 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 2448- 2452). IEEE.
5. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," arXiv:2304.02643, 2023.
6. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Transactions of the Association for Computational Linguistics (TACL), vol. 5, pp. 135–146, 2017.
7. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre- training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.
8. R. Reddy, S. Kumar, and A. Sharma, "Automated Video Metadata Generation Using NLP," 2023.
9. A. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval," in Proc. ICCV, 2021.
10. R. Zellers, J. Lu, X. Lu, Y. Yu, and Y. Choi, "MERLOT: Multimodal Neural Script Knowledge Models," in Proc. NeurIPS, 2021.
11. J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," arXiv:2301.12597, 2023.
12. S. Kamath, M. Singh, I. Leal-Taixé, M. Rohrbach, and S. Tulsiani, "MDETR: Modulated Detection for End-to-End Multi-Modal Understanding," in Proc. ICCV, 2021
13. Y. Wang, Z. Li, H. Zhang, and X. Liu, "Multimodal Named Entity Recognition Using Deep Learning Techniques," Applied Sciences, vol. 15, no. 3, 2025.
14. Y. Huang, X. Chen, Z. Zhang, and L. Wang, "Multi-Image Multimodal Named Entity Recognition," 2024.
15. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, 2016.
16. M. Mahendru and S. K. Dubey, "Real Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 734- 740, doi: 10.1109/Confluence51648.2021.9377064.

17. Y. -S. Poon, C. -C. Lin, Y. -H. Liu and C. -P. Fan, "YOLO-Based Deep Learning Design for In-Cabin Monitoring System with Fisheye-Lens Camera," 2022 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2022, pp. 1-4, doi: 10.1109/ICCE53296.2022.9730235.