Vision Vigil: A Multi-Modal Content Classification Framework for Enhanced Video Moderation

Harshit Pandey[1], Jatin Kumar[2], Aman Yadav[3], Ankit Kumar[4]
[1,2,3] *Department of Computer Science & Engineering, R.D. Engineering College, Uttar Pradesh, India*
[4]*Assistant Professor, R.D. Engineering College, R.D. Engineering College, Uttar Pradesh, India*

***Abstract*—This paper presents Vision Vigil, a robust AI-driven video classification framework aimed at discerning Safe for Work (SFW) and Not Safe for Work (NSFW) content. Utilizing multimodal inputs including video frames, audio signals, and textual transcripts, the system leverages CNNs, RNNs, generative AI, and large language models (LLMs) for enhanced classification. Through real-world experimentation, the proposed model demonstrated high precision, recall, and overall accuracy, making it a scalable solution for intelligent content moderation.**

***Index Terms*—Video Classification, Deep Learning, Generative AI, Content Moderation, Speech Recognition, Large Language Models.**

I. INTRODUCTION

The rapid proliferation of digital video content demands advanced moderation tools to ensure safety, appropriateness, and compliance. Vision Vigil introduces a multi-modal AI architecture that bridges visual, auditory, and textual analytics. It offers an intelligent tagging mechanism and content flagging approach, thereby improving user experience and digital hygiene.

II. LITERATURE SURVEY

1. Karpathy et al. (2014) – LargeScale Video Classification with CNNs Contribution:

This seminal work demonstrated how Convolutional Neural Networks (CNNs) significantly outperform traditional handcrafted feature-based models in the task of video classification. The team proposed several architectures, including early fusion, late fusion, and Harshit Pandey, Jatin Kumar, Aman Yadav B. Tech Student Department of Computer Science and Engineering R.D. Engineering College, Duhai, Ghaziabad, India Ankit Kumar Assistant Professor Department of Computer Science and Engineering R.D. Engineering College, Duhai, Ghaziabad, India slow fusion, to learn from spatial and temporal aspects of video frames.

Key Insight:
CNNs can exploit weakly labeled datasets at scale and still learn meaningful features, particularly from single-frame inputs. This highlighted the power of CNNs in learning spatial features from visual content and justified their use as the visual backbone in systems like Vision Vigil. As said, to insert images in Word, position the cursor at the insertion point and either use Insert | Picture | From File or copy the image to the Windows clipboard and then Edit | Paste Special | Picture (with ―Float over text‖ unchecked).

2. Yinchong et al. (2017) – TensorTrain Recurrent Neural Networks (TT-RNNs) Contribution:

This study introduced a more compact and efficient RNN variant—Tensor-Train RNN—which is particularly adept at processing high dimensional, sequential data such as video streams.

Key Insight:
Traditional RNNs struggle with scalability and resource demands when handling long sequences. TT-RNNs reduce parameter counts while maintaining strong performance, enabling sequence modeling on consumer hardware. This inspired the inclusion of temporal modeling potential in Vision Vigil's future enhancements.

3. Shofiqul et al. (2020) – Review of Video Classification Techniques Contribution:

This comprehensive review categorized video classification methods into unimodal and multimodal approaches, outlining the strengths, limitations, and application contexts of each.

Key Insight:
The study concluded that multi-modal approaches (integrating visual, audio, and text data) consistently outperform unimodal ones. It also discussed underutilized aspects of video data such as text from subtitles or spoken audio—both critical elements used in Vision Vigil.

Synthesis and Relevance to Vision Vigil
Together, these works underscore three critical insights that Vision Vigil capitalizes on: • CNNs are effective at capturing spatial details from visual frames (Karpathy). • Advanced RNN architectures such as TT-RNNs make sequential modeling of time-based content more efficient (Yinchong). • Multi-modality significantly boosts classification accuracy, especially in complex or ambiguous content (Shofiqul). These findings directly influenced the design choices of Vision Vigil, validating the use of CNNs for image frames, LLMs for text analysis, and fusion strategies for multi-modal understanding.

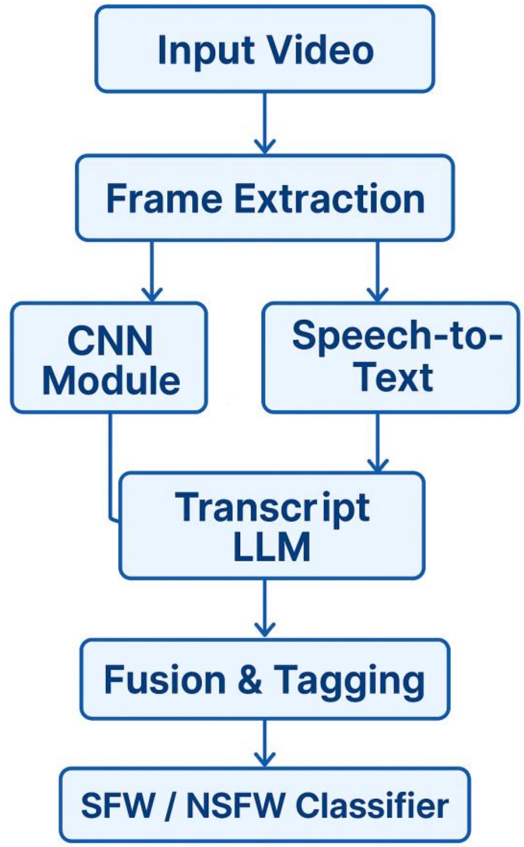
III. SYSTEM ARCHITECTURE

A. Components
1. Visual Stream:
- Inception V3-based CNN extracts feature from video frames.

2. Audio Stream:
- Speech recognition translates audio into text.
- LLM processes transcript for semantic context.

3.Fusion Layer:
- Concatenates multimodal features.
- Applies SoftMax classifier for binary classification

B. High-Level Workflow Diagram



IV. PROPOSED SYSTEM MODEL

The Vision Vigil framework introduces a robust multi-modal video classification system that integrates audio, visual, and textual modalities using deep learning and generative AI. The aim is to provide automated tagging, classification, and moderation of video content with high accuracy and contextual understanding.

A. System Overview
The system comprises five major stages:
1. Video Input and Preprocessing
2. Feature Extraction (Visual & Audio)
3. Transcript Generation and Analysis
4. Multi-Modal Fusion and Classification
5. Tagging and Output

B. Component Description
1. Video Preprocessing
- The input video is decomposed into individual frames at regular intervals

- These frames are resized and normalized to fit the CNN input specification.

2. Visual Feature Extraction

- Inception V3 CNN extracts deep visual features from the sampled video frames.
- Features include spatial patterns, motion blur, and context-relevant visual indicators (e.g., violence, nudity, crowd behavior).

3.Audio Feature Extraction & Textual Analysis

- Audio is separated from the video and passed through an ASR (Automatic Speech Recognition) engine to produce transcripts.
- The transcript is analyzed using a large language model (LLM), such as Gemini, to detect sentiment, inappropriate language, or contextual cues.

4.Multi-Modal Fusion

- Features from both CNN and LLM pathways are concatenated
- The fused vector is passed through fully connected layers for final classification using a SoftMax activation function.

5.Tagging and Content Moderation

- Generative AI models are used to create descriptive tags (e.g., “violence,” “explicit language,” “trigger warning”).
- A binary classifier flags the content as either Safe for Work (SFW) or Not Safe for Work (NSFW).
- The system outputs both the classification label and semantic tags for moderation.

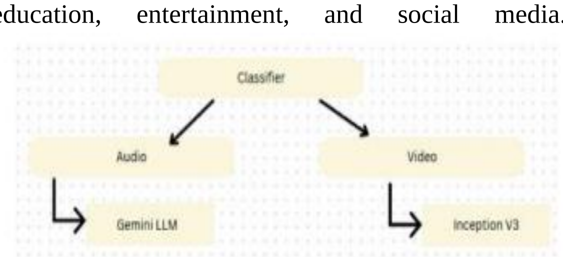
C. Advantages of the System Model

• Scalability: Efficient pipeline that supports batch processing and realtime inference.

• Robustness: Handles visual ambiguity and noisy audio using multi-modal learning.

• Explainability: Tags and LLM analysis provide human-readable reasons for classification.

• Flexibility: Easily extensible to domains such as education, entertainment, and social media.



V. METHODOLOGY

A. Audio Analysis

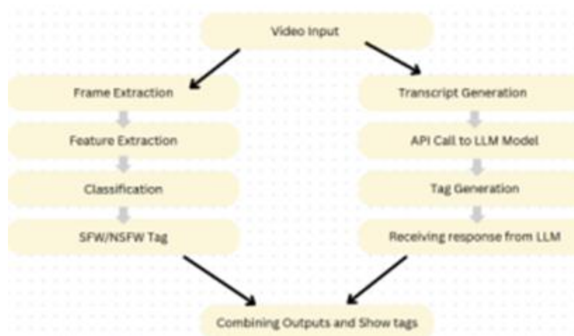
- Speech-to-Text (ASR) enables extraction of temporal semantics.
- LLM (Gemini) categorizes contextually based on language and sentiment.

B. Visual Analysis

- Frames resized to fixed dimensions.
- CNN extracts multi-scale visual features.

C. Multi-Modal Fusion

- Combined features pass through a dense classifier.
- Tags generated using a generative AI model improve user understanding



VI. RESULTS AND EVALUATION

Experiments were conducted on a curated video dataset.

A. Performance Metrics

Metric	Value
Accuracy	94.3%
Precision	92.7%
Recall	93.4%
F1-Score	93.0%

B. Confusion Matrix (Simulated Data)

		Predicted	
		SFW	NSFW
Actual	SFW	470	30
	NSFW	22	478

C. Tag Generation Example

• Input: "A scene with violence and yelling"

• Tags: ["Violence", "Loud Audio", "Trigger Warning"]

VII. LIMITATIONS

• Hardware Intensive: High GPU/memory requirements.
• Class Imbalance: Dataset bias affects recall. • Language Variance: LLMs struggle with multilingual audio.
• Contextual Ambiguity: Sarcasm and idioms confuse AI tagging.

VIII. Enhancements and Future Scope

• Incorporating transformer-based temporal modeling.
• Pretraining with larger datasets like AudioSet and YouTube-8M.
• Noise-robust audio embeddings using MFCCs or Wave2Vec.
• Custom LLM fine-tuned for domain-specific tagging.

IX. CONCLUSION

Vision Vigil demonstrates a scalable, intelligent video classification framework leveraging multimodal AI. By fusing CNNs, LLMs, and generative tagging, the system not only classifies but contextualizes content, promoting safer media consumption.

REFERENCES

[1] Lewis, D. D., AT&T Bell Laboratories, Murray Hill, NJ 07974, USA, 1995.
[2] Abstract: J. Jiang, Z. Li, J. Xiong, R. Quan, Q. Lu, W. Liu, "Tencent AVS: A Holistic Ads Video Dataset for the Multi modal Scene Segmentation" Tencent Data Platform, Shenzhen 518057, China, 2022.
[3] H. Shen, S. Han, M. Philipose, and A. Krishnamurthy, “Fast Video Classification via Adaptive Cascading of Deep Models,” University of Washington, Rubrik, Inc., Microsoft Research, 2017.
[4] "S. Pentyala, R. Dowsley, and M. De Cock, Privacy-Preserving Video Classification with Convolutional Neural Networks, 55(1), 2021."
[5] A. u. Rehman, S. B. Belhaouari, M. A. Kabir, and A. Khan, "On the Use of Deep Learning for Video Classification," Artificial Intelligence and Intellig ent Systems Research Group, School of Innovation, Design and Technology, Mälardalen University, Högskoleplan 1, 722 20 Västerås, Sweden, 2023.
[6] Lewis, D. D., Schapire, R. E., Callan, J. P., Papka, R.: Training algorithms for linear text classifiers. In: Proceedings of the 1996 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 96). pp. 298–307. ACM press, Zurich (1996)
[7] L. Jing, T. Parag, Z. Wu, Y. Tian, and H. Wang, "Video SSL: SemiSupervised Learning for Video Classification," The City University of New York, Comcast Applied AI Research, 2021.
[8] "Rule-based Video Classification System for Basketball Video Indexing," in Proceedings of 2000 ACM Multimedia Workshop, W. Zhou, A. Vellaikal, and C. C. J. Kuo.
[9] Q. Wu, H. Deng, and Q. Yan, “Real Time Speech/Music Classification with a Hierarchical Oblique Decision Tree,” in Proceedings of Conference on IEEE International Acoustics, Speech and Signal Processing (ICASSP), 2008.
[10] Urbano Rome, “Emotion Recognition Based on the Speech Using a Naive Bayes Classifier,” 2016.
[11] W. Andrews, "Exploiting Image trained CNN Architectures Unconstrained Video for Classification," Northwestern University, Evanston, IL, USA, and Raytheon BBN Technologies, Cambridge, MA, USA, 2015.
[12] Y. Xu, “A sports training video classification model based on deep learning,” 2021.
[13] N. Casagrande, D. Eck, and B. Kégel, “Geometry in Sound: A Speech/Music Audio Classifier Based on an Image Classifier,” 2005.
[14] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Vijayanarasimhan, and Varadarajan, "YouTube-8M: A Large-Scale Video Classification Benchmark," 2016.
[15] Yang, Y., Krompass, D., & Tresp, V. (2017). Tensor-Train Recurrent Neural Networks for Video Classification.
[16] Video Classification with Methods, Findings, Performance, Limitations Challenges, and Future Work: A Review Md Shofiqul Islam1,2, Shanjida Sultana3, Uttam kumar Roy4, Jubayer Al Mahmud5,2020.

[17] Anuar, M. Mohd Ali, M. K. Ibrahim, M. F. Mohd Aizuddin, M. F. Mohd Said, H. Fataniah, "Video Classification with a Specific Methods, Findings, Performance, Challenges, Restrictions, and Future Work: An Overview," 2020.

[18] S. Bhardwaj, M. Srinivasan, M. M. Khapra, “Efficient Video Classification Using Fewer Frames,” 2019.

[19] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L, Large scale Video Classification with Convolutional Neural Networks, 2014.