engineering proceedings

MDPI

Proceeding Paper

Audiovisual Fusion Technique for Detecting Sensitive Content in Videos †

Daniel Povedano Álvarez *, Ana Lucila Sandoval Orozco and Luis Javier García Villalba

Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Facultad de Informática, Universidad Complutense de Madrid (UCM), 28040 Madrid, Spain; asandov@ucm.es (A.L.S.O.); javierv@ucm.es (L.J.G.V.)

* Correspondence: dpovedano@ucm.es

† Presented at the First Summer School on Artificial Intelligence in Cybersecurity, Cancun, Mexico, 3–7 November 2025.

Abstract

The detection of sensitive content in online videos is a key challenge for ensuring digital safety and effective content moderation. This work proposes the Multimodal Audiovisual Attention (MAV-Att), a multimodal deep learning framework that jointly exploits audio and visual cues to improve detection accuracy. The model was evaluated on the LSPD dataset, comprising 52,427 video segments of 20 s each, with optimized keyframe extraction. MAV-Att consists of dual audio and image branches enhanced by attention mechanisms to capture both temporal and cross-modal dependencies. Trained using a joint optimisation loss, the system achieved F1-scores of 94.9% on segments and 94.5% on entire videos, surpassing previous state-of-the-art models by 6.75%.

Keywords: deep learning; audio-visual analysis; Mel spectrograms; video classification; attention mechanisms

1. Introduction

In recent years, the widespread use of multimedia devices and the rapid online dissemination of content has increased the presence of illegal and sensitive material, particularly affecting children and adolescents [1].

Artificial intelligence (AI) has shown potential in supporting crime prevention and child protection [2]. Automated tools help detect explicit content, restrict access in sensitive environments, and assist forensic investigations [3]. However, strict legal constraints limit access to original material, requiring systems capable of operating without direct exposure.

Multimodal approaches combining visual and auditory cues are essential, as up to 35% of CSA samples lack visible nudity [4]. Integrating audio features via techniques like Mel spectrograms or speech-to-text (e.g., *Whisper*) can reveal abuse indicators invisible to visual models. Prior studies demonstrate the effectiveness of audiovisual fusion using attention mechanisms or stacking, achieving accuracies above 92% [5].

This work introduces MAV-Att (Multimodal Audiovisual Attention), a lightweight deep learning architecture designed to improve sensitive content detection through hierarchical fusion and attention mechanisms. The proposed framework advances the development of effective multimodal detection systems that reduce false negatives and improve overall accuracy in detecting CSA and related material.

Check for updates

Academic Editors: Héctor Manuel Pérez Meana and Gabriel Sánchez Pérez

Published: 2 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

2. Related Work

Recent advances in deep learning (DL) have significantly improved the detection of sensitive and sexual content, achieving unprecedented accuracy and efficiency. These improvements rely on multimodal features, each contributing to the identification of online sexual exploitation and abuse material.

Multimodal Features

Combining audio and visual information enables a more comprehensive and robust understanding of content, improving contextual interpretation and classification accuracy.

Fu et al. [5] introduced PornNet, a multimodal CNN framework integrating visual and auditory cues with attention, achieving 93.4% accuracy. Zhao et al. [6] developed Porn2Vec, leveraging multimodal embeddings for pornographic website detection with an F1-score of 99.25%. Recent studies [7] explored cross-attention mechanisms for fusing heterogeneous modalities, enhancing performance in emotion recognition and sensitive content detection.

3. Sensitive Content Detection in Audio and Video

Transformer-based architectures and attention mechanisms have proven particularly effective in modeling spatial and temporal dependencies across modalities, reaching accuracy above 99% on benchmark datasets such as Pornography-2k. However, most approaches treat each modality independently, limiting contextual understanding. The Large-Scale Sexual Private Dataset (LSPD) [8] was selected due to its unique multimodal structure, including both class-level labels and polygonal masks identifying sexual organs in images and videos. With 500,000 images and 4000 videos (2000 safe and 2000 sexual), it provides a robust foundation for multimodal learning and evaluation.

3.1. Methodology

To address this limitation, we propose MAV-Att, a novel end-to-end multimodal architecture that jointly optimizes auditory and visual branches using attention-based feature fusion. The workflow consists of the following stages:

1. Preprocessing: Videos are segmented into 20 s clips following best practices [9]. Keyframes temporally aligned with audio segments are extracted using perceptual hashing (pHash) to remove redundant visuals.
2. Feature Extraction: Log-Mel spectrograms encode temporal and spectral properties of audio, while CNNs extract semantic and spatial visual representations.
3. Unimodal Attention: Each branch applies self-attention to highlight the most discriminative features within its modality.
4. Fusion and Classification: Outputs from both modalities are fused through a weighted softmax function to predict safe vs. sensitive content.

This attention-guided fusion allows the model to focus on the most relevant patterns and interactions between auditory and visual cues, enhancing interpretability and robustness. Figure 1 shows an overview of the MAV-Att architecture.

https://doi.org/10.3390/engproc2026123011

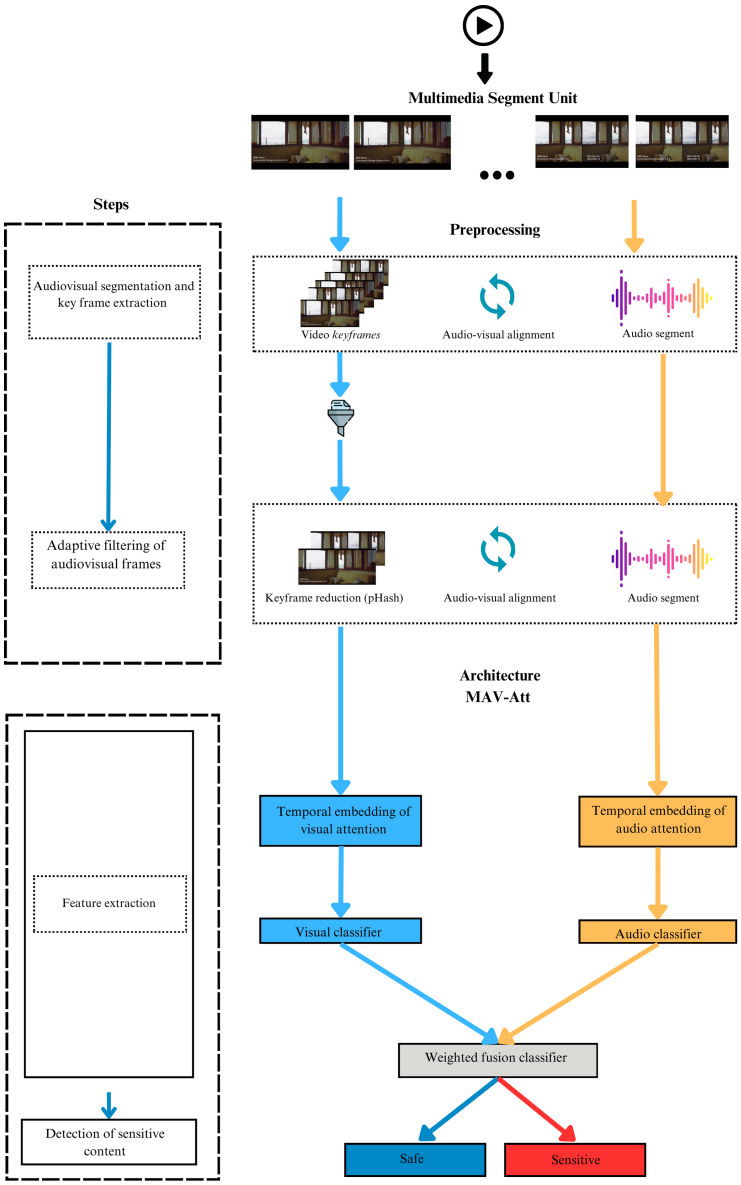


Figure 1. Overview of the MAV-Att architecture and workflow.

3.2. Data Preprocessing and Experimental Setup

A structured preprocessing pipeline was implemented to enhance the extraction of audiovisual features. The process included data cleaning, noise reduction, normalization, and temporal alignment between visual frames and their corresponding audio segments. Videos were segmented into 20 s clips, while redundant frames were removed via a perceptual hashing (pHash) algorithm with a Hamming distance threshold of 10, retaining between 2 and 10 keyframes per segment. Data augmentation—such as random cropping, flipping, and pitch shifting—was applied to improve model generalization and robustness under varying input conditions.

After preprocessing, the dataset comprised 52,427 audio segments and 277,133 aligned images across both classes. This process ensured temporal coherence and increased data diversity without additional manual labeling.

Experiments were conducted using the LSPD (Large-Scale Pornography Dataset), divided into 80% for training and 20% for testing, ensuring that segments from the same video did not appear in both sets to prevent data leakage. A five-fold cross-validation strategy was applied to validate model consistency. Performance was measured using

precision, recall, F1-score, and confusion matrices. Predictions were obtained through majority voting across aligned audiovisual segments.

3.3. Implementation Details and 20-s Segment Experiments

AdamW optimizer ($lr = 1 \times 10^{-4}$, weight decay 1×10^{-3}) was used with batch size 32 for 50 epochs and early stopping (patience 5). Preprocessing included resizing images to 224×224, normalizing using ImageNet statistics, and converting audio to log-Mel spectrograms ($N_{\text{FFT}} = 1024$, 16 kHz, $N_{\text{MELS}} = 224$, hop length 512). Data augmentation was applied in both modalities to increase robustness. Audio and visual segments were packed into tensors preserving temporal alignment.

Experiments were conducted using 20-s audiovisual segments. Baseline tests without attention or augmentation showed that combining audio and visual features improves F1 score from 80.8% (audio only) and 91.0% (visual only) to 92.3% (fusion with combined loss). Fine-tuning the MAV-Att model with trainable layers and data augmentation increased F1 to 94.2% in cross-validation. Experiments on 20 s segments with selected key frames achieved F1 scores of 94.1% for safe content, 95.6% for sensitive content (see Table 1) and a global F1 of 94.8% in videos, demonstrating that attention-enhanced segments provide valuable contextual information.

Table 1. Classification report with precision, recall, F1-score, and support for the test set.

Label	Precision (%)	Recall (%)	F1-Score (%)	Support
Safe	94.7	93.58	94.1	4627
Sensitive	95.1	95.98	95.6	6048
Overall Accuracy			94.9	10,675
Macro Avg	94.9	94.8	94.8	10,675
Weighted Avg	94.9	94.9	94.9	10,675

4. Conclusions and Future Work

This study presents a video classification model integrating visual and audiovisual attention mechanisms to enhance sensitive content detection. Overall, combining modalities with attention significantly improves audiovisual classification, offering a scalable solution for automated digital content moderation. Future work will explore alternative key frame selection strategies and investigate more sophisticated methods to effectively fuse visual and audio modalities.

Author Contributions: D.P.Á., A.L.S.O. and L.J.G.V. contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out with funding from the Recovery, Transformation and Resilience Plan, financed by the European Union (Next Generation EU), through the Chair "Cybersecurity for Innovation and Digital Protection" INCIBE-UCM. In addition, this work has been supported by Comunidad Autonoma de Madrid, CIRMA-CM Project (TEC-2024/COM-404).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are publicly available. The datasets and relevant details are described in the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ramaswamy, S.; Seshadri, S. Children on the Brink: Risks for Child Protection, Sexual Abuse, and Related Mental Health Problems in the COVID-19 Pandemic. *Indian J. Psychiatry* **2020**, *62*, S404. [CrossRef] [PubMed]
2. The Times. AI in Crime Fighting: Transforming Child Protection Efforts, 2025. Available online: https://www.thetimes.co.uk/article/ai-there-youre-nicked-tech-is-reshaping-how-we-fight-crime-jb9bv7qh3 (accessed on 26 January 2025).
3. Cifuentes, J.; Sandoval Orozco, A.L.; García Villalba, L.J. A Survey of Artificial Intelligence Strategies for Automatic Detection of Sexually Explicit Videos. *Multimed. Tools Appl.* **2022**, *81*, 3205–3222. [CrossRef]
4. Laranjeira, C.; Macedo, J.; Avila, S.; dos Santos, J.A. Seeing without Looking: Analysis Pipeline for Child Sexual Abuse Datasets. *arXiv* **2022**, arXiv:2204.14110. [CrossRef]
5. Fu, Z.; Li, J.; Chen, G.; Yu, T.; Deng, T. PornNet: A Unified Deep Architecture for Pornographic Video Recognition. *Appl. Sci.* **2021**, *11*, 3066. [CrossRef]
6. Zhao, J.; Shao, M.; Peng, H.; Wang, H.; Li, B.; Liu, X. Porn2Vec: A Robust Framework for Detecting Pornographic Websites based on Contrastive Learning. *Knowl.-Based Syst.* **2021**, *228*, 107296. [CrossRef]
7. Khan, M.; Gueaieb, W.; El Saddik, A.; Kwon, S. MSER: Multimodal Speech Emotion Recognition using Cross-attention with Deep Fusion. *Expert Syst. Appl.* **2024**, *245*, 122946. [CrossRef]
8. Phan, D.D.; Nguyen, T.T.; Nguyen, K.N.K.; Nguyen, Q.H.; Vu, D.L.; Tran, H.L. LSPD: A Large-Scale Pornographic Dataset for Detection and Classification. *Int. J. Intell. Eng. Syst.* **2022**, *15*, 198–213. [CrossRef]
9. Lovenia, H.; Lestari, D.P.; Frieske, R. What Did i Just Hear? Detecting Pornographic Sounds in Adult Videos Using Neural Networks. In Proceedings of the 17th International Audio Mostly Conference, St. Pölten, Austria, 6–9 September 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 92–95. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.