

Study on scene classification with unimodal and multimodal deep optimization learning techniques for visual-acoustic data

Chaitanya Nukala *

*Department of Computer Science and Engineering
Jawaharlal Nehru Technological University Kakinada
Kakinada 533003
Andhra Pradesh
India*

V. Raghavendran [†]

*Department of Information Technology
Aditya University
Surampalem 533437
Andhra Pradesh
India*

Abstract

Scene classification is an emerging need in social applications, such as surveillance security and autonomous actions, in which unimodal learning techniques are commonly used to classify scenes from either audio or visual data. Researchers are advancing advances to develop more accurate scene classifications using multimodal techniques. Here, two modalities of data, i.e., both visual and acoustic, are learned for cross-modal fusion scene classification. This paper presents unimodal learning techniques with either acoustic or visual modalities, and the state-of-the-art in multimodal learning techniques for a comparative study of scene classification results. In unimodal, visual-based learning, scene classification is achieved as well as with acoustic-based learning techniques. The experimental study conducted using publicly available benchmarked datasets, TAU datasets, its results demonstrated that recent multimodal learning techniques, Attention graph convolutional network (AGCN), Context-Aware Proposal-Boundary Network (CAPB), Audio-Visual Semantic Modelling (AVSM), gain higher performance values than other existing techniques.

Subject Classification: Primary 93A30, Secondary 49K15.

Keywords: Scene classification, Deep learning, Visual-acoustic data, Unimodal learning, Multimodal learning.

* E-mail: chaitu.anju@gmail.com (Corresponding Author)

[†] E-mail: raghuchv@yahoo.com

1. Introduction

Classification of scenes based on visual frames and acoustic prompts would become a prominent research topic, useful for computer vision applications. Scene feature extraction and learning are the two pivotal steps for deriving scene classification results from large volumes of visual-acoustic data. These two modalities, i.e., visual and acoustic, were primarily targeted in existing works, namely unimodal and multimodal learning techniques. Hershey et al. [1] showed that deep-learning models are able to learn and classify a large amount of acoustic data. The CNN-based models, including deep neural networks (DNNs) and their extensions AlexNet, ResNet, VGG, and Inception, employed deep-feature extraction for efficient audio-based classification. Their feature representations are more robust than acoustic raw features. Of these models, ResNet-50 is a widely adopted model and has been proven to be effective in scene classification task with the Audio Set dataset used in Gemmeke et al. [2]. with nearly 3000 hours of *audio*. Another unimodal, contrastive language-image pre-training” (CLIP) also produces impressive scene classification results based on visual frames data. The CLIP interprets the natural language processing (NLP) task for processing scene information effectively. It encodes the visual and caption (text) data for training and test cases in scene classification. A limitation of the CLIP is that it cannot use audio features or their coefficients, which limits its temporal consistency.

In recent years, several advanced multimodal learning methods have been developed, such as the Attentional Graph Convolutional Network (AGCN) proposed by Zhou et al. [3], the Context-Aware Proposal–Boundary Network (CAPB) proposed by Wang et al. [4] and Video Summary Generation Model Based on Hybrid Attention (VSMHA) proposed by Pang, Gao, and Zhang [5], performing well for scene classification with visual-acoustic learning. The AGCN extracts foreground cues from the context using the structural data of audio-visual graphs. To facilitate efficient scene categorisation, it investigates the connections between auditory and visual cues. It does not, however, explicitly represent the multimodal data at the event level. Event consistency may be achieved by the model CAPB through explicit temporal modelling and accurate temporal boundary detection. To capture the delicate temporal transitions, it encodes both local and global contexts [6]. When defining the event localizations, the CAPB is a powerful tool. Therefore, by minimising the impact of false positives during scene categorisation in noisy situations, it achieves remarkable results through boundary-aware modelling. In order to leverage spatial attention, Zhou et al. [6] added an attention module to their Feature Pyramid Network (FPNet). To emphasise segments with channel attention, it highlights prominent areas in the frames and identifies high-frequency bands in the audio data. The attention-guided pyramidal aggregation in FPANet facilitates the categorisation of complex scenes. Although FPANet’s scene-level recognition performance is outstanding, its temporal flexibility remains limited. Despite its effective utilization of attention processes, FPANets

rely on fixed or heuristic temporal representations. FPNNet gathers features at preset time scales without learning event boundaries.

2. Scene Classification Methods

Visual and acoustic-based scene classification methods ensure semantic consistency in the frame-time alignment between visual and acoustic segments. Its multimodal learning is a significant breakthrough in scene detection from video segments. Tokenization plays vital role in the scene detection. Existing methods were used tokenization for extraction features with identified video segments. Cross-modal learning is used to merge visual and aural representations to enhance cross-modal feature interactions. Results from experiments show increased classification accuracy and robustness, particularly when dealing with sparse or noisy data. It employs hybrid fusion and transfer learning methods to categorize audio-visual situations. Table 1 presents a literature survey of recent methods, including their methodologies, results, and limitations.

Table 1
Literature Study for Scene Classification Techniques

Description of the Paper	Methodology	Results	Limitations	Observations
Hershey et al. [1], "CNN Architectures for Large-Scale Audio Classification"	CNN-based audio classification using spectrograms	Improved large-scale audio classification	Requires large labelled data; unimodal	Establishes baseline; highlights need for multimodal integration
Tan and Le [8], "EfficientNet: Rethinking Model"	Compound scaling of CNNs	High accuracy with fewer parameters	Limited to visual modality	Improves efficiency over traditional CNNs, enabling lightweight backbones for future multimodal models.
Hu et al. [9], "Cross-Task Transfer for Geotagged Audiovisual"	Audio-visual multimodal learning with transfer learning	Improved scene recognition using sound cues	Requires aligned multimodal data	Addresses unimodal limitation. Introduces audio-visual fusion concept
Radford et al. [10], "Learning Transferable Visual Models"	Contrastive learning using image-text pairs	Strong zero-shot performance	Requires massive datasets	Extends multimodal learning. Introduces contrastive alignment, improving cross-modal understanding
Hou, Kang, and Botteldooren [11], "Audio-Visual Scene Classification"	Contrastive event-object alignment with attention fusion	Improved fine-grained classification	Computational complexity	Improves over graph models. Introduces fine-grained alignment using contrastive learning

Mocanu, Tapu, and Zaharia [12], “Multimodal Emotion”	CNN + cross-modal attention+metric learning	High accuracy (89%)	Sensitive to noise	Addresses feature fusion issues. Uses attention to enhance modality interaction
Shaikh et al. [13], “Multimodal Fusion for Audio-Image and Video”	CNN-based multimodal fusion	Good accuracy	Weak temporal modeling	Extends previous fusion models. Highlights temporal dependency modelling
Zhou et al. [6], “Feature Pyramid Attention Network for Audio-Visual”	Multi-scale feature pyramid with attention	High F1-score (95.9%)	High complexity	Addresses feature scale limitations, multi-scale attention mechanisms
Wang et al. [4], “Context-Aware Proposal–Boundary Network”	Temporal context modelling with proposal-boundary learning	Improved localization accuracy	Complex design	Improves temporal understanding. Introduces local-global context modelling
Gong et al. [14], “Complementary and Contrastive Learning for Audio-Visual Segmentation”	Transformer with contrastive multimodal fusion	SOTA performance	High computational cost	Addresses CNN limitations. Leverages global attention for better fusion
Gong et al. [15], “AVS-Mamba: Exploring Temporal and Multi-Modal”	State-space model for temporal modeling	Efficient long-range dependency modeling	Limited validation	Addresses transformer inefficiency, linear-complexity temporal modelling
Wang, Qian, and Zhou [17], “Audio-Visual Segmentation via Transformer and Prompt Learning”	Transformer with prompt-based learning	Improved segmentation	Prompt tuning complexity	Improves transformer adaptability, prompt-guided feature refinement

Audio–visual scene classification has advanced rapidly with the growth of deep learning. Zhou et al. [6] introduced FPANet, which extracts semantic features from visual and acoustic inputs for unimodal and multimodal scene understanding. Although the model performs well, its high computational and memory requirements may limit scalability. Wang et al. [7] explored transfer learning with hybrid feature fusion, while Tan and Le [8] proposed EfficientNet to improve CNN efficiency. Data augmentation in both audio and visual streams can further improve model generalization.

3. Semantic Consistency for Visual-Acoustic Scenes

The consistent alignment of semantic information between visual content (images or video frames) and corresponding auditory signals in visual-acoustic scenarios is known as semantic consistency, ensuring that both modalities depict the same real-world context. Ahmed et al. [18] emphasized that this idea is

essential to multimodal learning, as separate processing of auditory and visual data often results in weak correlations, redundancy, or Leelavathi, Thahniyath, and Prasad [20] discussed the conflicting interpretations. Early methods mostly employed convolutional neural networks for unimodal feature extraction, which worked well for individual modalities but were unable to capture inter-modal interactions [16]. Although auditory and visual information were integrated in later multimodal fusion techniques, there were no means to ensure appropriate semantic alignment. However, cross-modal interaction has greatly improved due to recent developments such as contrastive learning, attention mechanisms, and transformer-based architectures; however, many of these strategies still do not explicitly ensure semantic consistency.

To address this, semantic consistency can be formulated mathematically by projecting both modalities into a shared embedding space. Let V and A denote the visual and audio inputs, respectively. Feature representations are extracted using modality-specific encoders: $f_v = \phi_v(V)$ for visual data and $f_a = \phi_a(A)$ for audio data. These features are then mapped into a common latent space through linear transformations: $z_v = W_v f_v$ and $z_a = W_a f_a$. In this shared space, semantically consistent audio–visual pairs should lie close to each other, while inconsistent pairs should be far apart. This relationship is typically enforced using cosine similarity, defined as $S(v, a) = \frac{z_v \cdot z_a}{\|z_v\| \|z_a\|}$, which measures the alignment between the two modalities.

To further enhance this alignment, contrastive learning is used, similar to CLIP and audio-visual alignment models. The contrastive loss is $\mathcal{L}_{con} = -\log \frac{\exp\left(\frac{S(v, a^+)}{\tau}\right)}{\sum_{a^-} \exp(S(v, a^-)/\tau)}$, where a^+ refers to the positive audio matching the visual input, a^- refers to negative audio, and τ is a temperature. This loss function maximises similarity for the correct pair and minimises it for the negative pairs. To further encourage semantic alignment, we also adopt a semantic consistency loss $\mathcal{L}_{sem} = \|z_v - z_a\|_2^2$, which pushes for embeddings of the matched audio and visual to be close.

4. Comparative Study on Scene Classification Methods

The audio and visual scenario benchmarked dataset TAU, which contains many classes such as “airport”, “shopping mall”, “metro station”, “street pedestrian”, “public square”, “street traffic”, “tram”, “bus” and “park”, is publicly available in Mesaros, Heittola, and Virtanen [19]. These are classes from the 34 hours of high-quality video recordings of the dataset. We used preprocessing in this experiment to extract features from audio and visual data of these types. It is initialized by converting recorded audio data to mel-frequency cepstral coefficients (MFCC) and by performing pre-computations for the VGG19 model to extract visual features. Table 2 compares state-of-the-art methods for scene classification.

The progression from unimodal to more advanced multimodal approaches in visual-acoustic scene processing is evident in the performance results in Table

2. Using unimodal techniques, such as Audio-DCNN and Visual scene classification, the limited performance (0.811 and 0.822 accuracy) is due to the reliance on a single mode of information. While lacking strong semantic interaction, the use of multimodal approaches, such as Early fusion, improves the accuracy to 0.834 by combining auditory and visual cues. The CLIP model, which uses contrastive learning, achieves an accuracy of 0.842 in generalisation. But it has a substantially lower recall (0.736) showing less alignment in specific scene contexts. The advanced techniques exhibit promising improvements. Two recent methods, CEOA and CrossTask, achieve 0.885 and 0.856 accuracy, respectively, due to their semantic alignment between scene events and their physical objects, and their transfer learning approach. Similarly, AGCN employs graph learning to model relationships with an accuracy of 0.877. However, EfficientNet, which uses only deep visual models, has relatively lower accuracy (0.802), suggesting the need for multimodal deep models. AVSM has the highest accuracy (0.892), precision (0.832), recall (0.822) and F1-score (0.827). This demonstrates its ability to fuse multimodal features and align semantic meanings.

Table 2
Performance Analysis of Proposed and Existing Methods

Method Name	Accuracy	Precision	Recall	F1-Score
Audio-DCNN	0.811	0.795	0.798	0.796
Visual scene classification	0.822	0.812	0.801	0.812
Early fusion	0.834	0.822	0.802	0.822
CLIP	0.842	0.785	0.736	0.760
CrossTask	0.856	0.788	0.775	0.781
CEOA	0.885	0.821	0.808	0.815
EfficientNet	0.802	0.782	0.776	0.779
AGCN	0.877	0.802	0.799	0.801
AVSM	0.892	0.832	0.822	0.827

In summary, the results leave no doubt that approaches fusing multimodal information in a structured manner, and considering semantic alignment and attention mechanisms, greatly surpass the unimodal and simple fusion approaches, highlighting the importance of deep interaction between multiple modalities for better scene understanding. The superiority of effective multimodal techniques in visual-acoustic scene understanding is also evident from comparisons of the AUC and mAP across different approaches shown in Figure 1 and Figure 2.

CrossTask has the highest discriminative power, with an AUC of 0.91. AVSM is the next at 0.88 and performs well due to successful semantic fusion. Although unimodal or misaligned models such as Audio model (0.50) and CEOA (0.49) exhibit lower AUC values, due to a lack of effective cross-modal interactions, other methods such as AGCN (0.79), Early fusion (0.80) and Visual model (0.82) show decent performance.

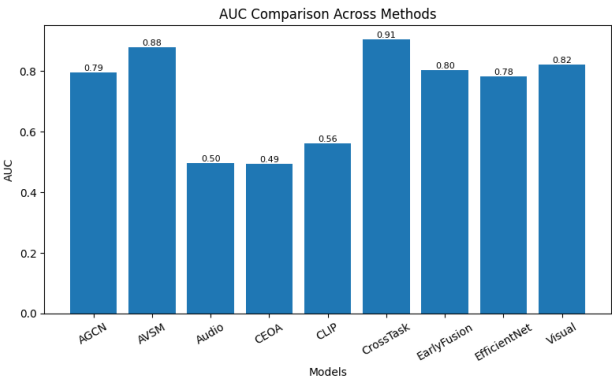


Figure 1
AUC Comparison of Different Audio-Visual Scene Classification Methods

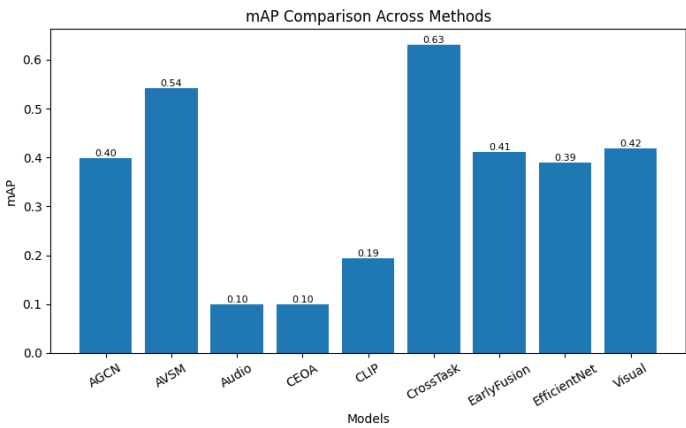


Figure 2
Comparison of mean Average Precision (mAP) across different models

In the mAP ranking, we can see a similar trend, where Cross Task again leads with a score of 0.63, followed by AVSM 0.54. These results confirm the effectiveness of semantic fusion and transfer learning in boosting retrieval and detection. Conversely, the unimodal approaches (Audio model, CEOA) and EfficientNet provide very low mAP values (0.10), indicating poor generalization in retrieving and ranking tasks, while AGCN (0.40), Early fusion (0.41) and EfficientNet (0.39) yield satisfactory results.

5. Conclusion and Scope of the Work

The experimental evaluation suggests that multimodal approaches significantly improve visual-acoustic scene analysis over unimodal and simple fusion approaches by effectively leveraging audio and visual information. The

sophisticated techniques that use contrastive learning, attention and structured representations yield superior performance by capturing richer semantic correlations and promoting strong inter-modal alignment. Methods that emphasise semantic consistency and interaction across contexts yield improved classification, discrimination, and retrieval, as reflected in higher accuracy, area under the curve (AUC), and mean average precision (mAP). But challenges such as computational burden, data requirements and robustness in the presence of spurious or missing features remain. The quest for lightweight, adaptive, and spatially robust models needs to be pursued.

References

- [1] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN Architectures for Large-Scale Audio Classification," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, pp. 131–135 (2017), doi: 10.1109/ICASSP.2017.7952132.
- [2] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, pp. 776–780 (2017), doi: 10.1109/ICASSP.2017.7952261.
- [3] L. Zhou, Y. Zhou, X. Qi, J. Hu, T. L. Lam, and Y. Xu, "Attentional Graph Convolutional Network for Structure-Aware Audiovisual Scene Classification," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15 (2023), doi: 10.1109/TIM.2023.3260282.
- [4] H. Wang, Z.-J. Zha, L. Li, X. Chen, and J. Luo, "Context-Aware Proposal-Boundary Network with Structural Consistency for Audiovisual Event Localization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 15872–15882 (2024), doi: 10.1109/TNNLS.2023.3290083.
- [5] X. Pang, L. Gao, and L. Zhang, "A Video Summary Generation Model Based on Hybrid Attention Using Multimodal Features," *IEEE Access*, vol. 13, pp. 37078–37095 (2025), doi: 10.1109/ACCESS.2025.3545113.
- [6] L. Zhou, Y. Zhou, X. Qi, J. Hu, T. L. Lam, and Y. Xu, "Feature Pyramid Attention Network for Audio-Visual Scene Classification," *CAAI Trans. Intell. Technol.*, vol. 10, no. 2, pp. 359–374 (2025), doi: 10.1049/cit2.12375.

- [7] M. Wang, C. Chen, Y. Xie, H. Chen, Y. Liu, and P. Zhang, "Audio-Visual Scene Classification Using Transfer Learning and Hybrid Fusion Strategy," *DCASE2021 Challenge*, Tech. Rep. (Jun. 2021).
- [8] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, pp. 6105–6114 (2019).
- [9] D. Hu, X. Li, L. Mou, P. Jin, D. Chen, L. Jing, X. Zhu, and D. Dou, "Cross-Task Transfer for Geotagged Audiovisual Aerial Scene Recognition," in *Computer Vision—ECCV 2020*, Lecture Notes in Computer Science, pp. 68–84 (2020), doi: 10.1007/978-3-030-58586-0_5.
- [10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, pp. 8748–8763 (2021).
- [11] Y. Hou, B. Kang, and D. Botteldooren, "Audio-Visual Scene Classification via Contrastive Event-Object Alignment and Semantic-Based Fusion," in *Proc. 2022 IEEE 24th Int. Workshop Multimedia Signal Process. (MMSP)*, Shanghai, China, pp. 1–6 (2022), doi: 10.1109/MMSP55362.2022.9949447.
- [12] B. Mocanu, R. Tapu, and T. Zaharia, "Multimodal Emotion Recognition Using Cross-Modal Audio-Video Fusion with Attention and Deep Metric Learning," *Image Vis. Comput.*, vol. 133, Art. no. 104676 (2023), doi: 10.1016/j.imavis.2023.104676.
- [13] M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "Multimodal Fusion for Audio-Image and Video Action Recognition," *Neural Comput. Appl.*, vol. 36, pp. 5499–5513 (2024), doi: 10.1007/s00521-023-09186-5.
- [14] S. Gong, Y. Zhuge, L. Zhang, P. Zhang, and H. Lu, "Complementary and Contrastive Learning for Audio-Visual Segmentation," *IEEE Trans. Multimedia*, vol. 27, pp. 7407–7418 (2025), doi: 10.1109/TMM.2025.3599048.
- [15] S. Gong, Y. Zhuge, L. Zhang, Y. Wang, P. Zhang, L. Wang, and H. Lu, "AVS-Mamba: Exploring Temporal and Multi-Modal Mamba for Audio-Visual Segmentation," *IEEE Trans. Multimedia*, vol. 27, pp. 5413–5425 (2025), doi: 10.1109/TMM.2025.3542995.
- [16] M. S. Basha, K. R. Prasad, S. K. Mouleeswaran, R. C. Poonia, and S. Sebastian, "Multi-Disease Detection System with X-Ray Images Us-

- ing Deep Learning Techniques," *J. Inf. Optim. Sci.*, vol. 45, no. 5, pp. 1379–1388 (2024).
- [17] Y. Wang, X. Qian, and W. Zhou, "Transformer-Prompted Network: Efficient Audio-Visual Segmentation via Transformer and Prompt Learning," *IEEE Signal Process. Lett.*, vol. 32, pp. 516–520 (2025), doi: 10.1109/LSP.2024.3524120.
- [18] M. R. Ahmed, R. Haque, S. M. A. Rahman, A. W. Reza, N. Siddique, and H. Wang, "Vision-Audio Multimodal Object Recognition Using Hybrid and Tensor Fusion Techniques," *Inf. Fusion*, vol. 126, Art. no. 103667 (2026), doi: 10.1016/j.inffus.2025.103667.
- [19] A. Mesaros, T. Heittola, and T. Virtanen, "TAU Urban Audio-Visual Scenes 2021, Development Dataset," *Zenodo* (Feb. 26, 2021), doi: 10.5281/zenodo.4477542.
- [20] C. Leelavathi, G. Thahniyath, and K. R. Prasad, "Optimizing the Deep Features Using the Spectral Concept for an Efficient Video Classification," *J. Inf. Optim. Sci.*, vol. 46, no. 2, pp. 553–562 (2025), doi: 10.47974/JIOS-1959.

Received March, 2026