Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 47 (2026), No. 10, pp. 3575–3584
DOI : 10.47974/JIOS-2485

Adaptive semantic fusion based scene aligned segmentation for optimizing the efficacy of multimodal video classification

Karre Rathi *
T. Sasi Rooba †
Department of Computer Science & Engineering
Annamalai University
Annamalainagar 608002
Tamil Nadu
India

K. Subba Reddy §
Department of Computer Science and Engineering
Rajeev Gandhi Memorial College of Engineering & Technology
Nandyal 518501
Andhra Pradesh
India

Abstract

Multimodal video segmentation and classifications are an emerging need for the detection of video-based fraudulent applications. Existing fusion methods, such as the Multimodal Embedding Alignment System (MEAS), Transformer models, and Visual Semantic Multi-Head Attention (VSMHA), have produced prominent multimodal video classification results. However, they rely on static segmentation and uniform fusion strategies, which weakens cross-modal interaction and leads to the loss of contextual information. Nevertheless, they rely on uniform fusion and static segmentation, which weaken cross-modal interactions and lead to the loss of relevant context. To address these limitations, this paper presents a Dynamic Adaptive Semantic Fusion Framework (DASFF) that integrates an adaptive video segmentation system that synchronises semantic scene changes with segment boundaries in real time. The proposed method enables more coherent and discriminative feature representations by capturing both temporal dynamics and contextual semantics. To demonstrate the efficacy of the proposed DASFF method, an extensive experimental study is conducted using the TAU urban audio-visual dataset. The proposed DASFF achieved an accuracy of 98.70% and demonstrated robust improvements in statistical performance.

* *E-mail:* **karrerathi8@gmail.com (Corresponding Author)**
† *E-mail:* **sasiruba@gmail.com**
§ *E-mail:* **mrsubbareddy@yahoo.com**

© Taru Publications

The results indicate the usefulness of semantic-aware fusion and adaptive segmentation for improved multimodal video classifications.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Multimodal video classification, Adaptive segmentation, Semantic fusion, Attention mechanisms, Deep learning.*

1. Introduction

The rapid growth of video data from surveillance systems, social networks, and smart monitoring applications has made multimodal video understanding an important research topic in computer vision and artificial intelligence. By combining visual information with changes over time and contextual meaning, multimodal approaches provide a more comprehensive understanding of video content and can improve the accuracy and reliability of its interpretation. Moreover, the emergence of large-scale datasets and the advancement of deep learning architectures further propelled its adoption in various applications such as anomaly detection, smart surveillance, human activity recognition, and content-based video retrieval. However, although all this has happened, it is hard to effectively integrate multimodal information.

Existing multimodal video classification approaches typically rely on a prior defined segmentation and fusion strategy. The traditional early and late fusion methods fuse the modality features prior to, or after processing. However, they often do not represent complex relations between modalities, and they are often simply assuming that all contexts have the same contribution from all modalities, which is seldom the case in reality as noted by Vaswani et al. [1]. To overcome this drawback, attention-based approaches have come into vogue. Transformer-based architectures can capture long-range dependencies in video data using self-attention, as described by Liu et al. [2]. Similarly, MEAS uses language models, like RoBERTa by Baltrušaitis, Ahuja, and Morency [3] in this work, to obtain semantic embeddings which are then used to enhance the cross-modal representation, and VSMHA is formed to encode jointly both the visual and semantic information by using multi-head attention present in Zadeh et al. [4]. While these approaches advance the field of representation learning, they also have drawbacks in terms of computational complexity, scalability of the system, and adaptive fusion of context-aware features. Previously, multimodal video classification approaches have been difficult to develop due to the problem of determining which modality is the most informative and prioritizing it, thus restricting their discriminative capability.

Existing approaches have a major drawback: They assume a fixed structure of the video, ignoring semantic transitions and also making classification less effective as noted by Wu et al. [5]. To resolve this issue, DASFF fuses adaptive segmentation and dynamic semantic fusion. It uses the semantic cues and temporal dynamics to find meaningful transitions. It learns the modality importance from the context and achieves more robust cross-modal learning,

more accurate performance, better generalization and more efficient performance than existing approaches.

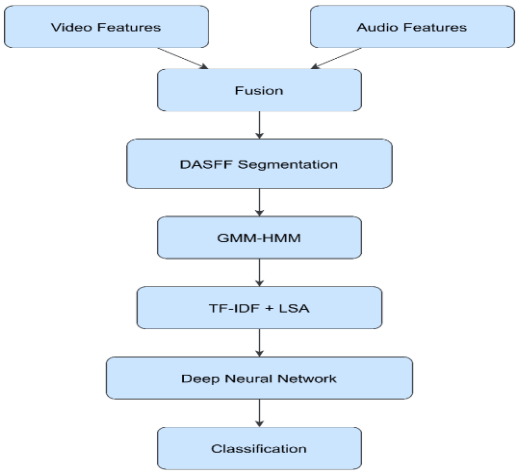


Figure 1
Key Steps for the Proposed DASFF frame work flow

The DASFF framework is composed of two parts: video and audio features, both of which are represented in a single unit as shown in Figure 1. The framework uses adaptive segmentation to find meaningful scene changes, GMM-HMMs to model the temporal aspect, and TF-IDF and LSA to obtain semantic features. Lastly, using a deep neural network, the scene is classified. DASFF combines multimodal fusion, adaptive segmentation and semantic learning to achieve more accurate and robust classification than the current methods.

The main contributions of this work are: (1) an adaptive video segmentation approach that aligns segment boundaries with semantic scene transitions, (2) a dynamic semantic fusion framework that assigns context-aware importance to multimodal features, (3) extensive experiments demonstrating superior performance over existing methods, and (4) a scalable and efficient architecture suitable for real-world multimodal video understanding applications.

2. Literature Study

Recent advances in multimodal video understanding combine audio-visual information using deep learning and attention mechanisms presented in Mocanu, Tapu, and Zaharia [14]. Feature pyramid attention models improve feature extraction through multi-scale and spatial-channel attention, but do not effectively handle temporal segmentation or semantic scene transitions as discussed by Zhou et al. [6]. The boundary methods with context are provided

by temporal encoding and consistency modelling to improve event localization, but they assume the existence of segments and cannot guarantee semantic continuity at boundaries proposed by Wang et al. [7] and Gong et al. [13]. Likewise, multimodal fusion methods enhance classification accuracy, but they typically adopt a static fusion method, which is not flexible for different scene conditions as discussed by Shaikh et al. [8]. Recent multimodals model methods are good on feature learning and efficient computational, but still there are difficulties of adaptive fusion, temporal consistency and segmentation accuracy, as examined [9]. Several methods have problems to capture semantic relations between frames, and often require separate modality-specific encoders, which leads to less effective multimodal integration mentioned in Gong et al. [10]. Multimodal emotion recognition and spatio-temporal attention methods also need a lot of pre-processing and segmentation, with fixed length, thus limiting their adaptability to the dynamic scenes, as highlighted by Wei et al. [11].

In the video understanding task, there are hybrid attention-based video summarization models and attention-based token-based models like in Iqbal et al. [17] that enhance video understanding, but are still based on predetermined video segmentation strategies. These methods are not able to accurately represent real semantic scene transitions, deteriorate temporal continuity and do not include adaptive multimodal fusion, which hinders the ability to select the best modality to highlight. In general, these methods have limitations such as fixed segmentation, static fusion and weak temporal-semantic modelling, which require the development of an adaptive segmentation and context-aware multimodal fusion for more effective video understanding.

3. Proposed Dynamic Adaptive Semantic Fusion Framework

DASFF combines adaptive segmentation with dynamic semantic fusion to better capture and integrate information from multiple modalities in videos. It detects meaningful scene transitions, assigns context-aware importance to different modalities, and combines them into a unified representation. By capturing both temporal patterns and cross-modal relationships, DASFF achieves more accurate and robust classification, especially in dynamic environments where modality importance changes over time.

Algorithm 1: Dynamic Adaptive Semantic Fusion Framework (DASFF)

Input: Multimodal video $V = \{(x_t^v, x_t^a, x_t^s)\}_{t=1}^{T}$

Output: Predicted class label Y

Methodology:

1. Extract visual features $f_t^v = \phi_v(x_t^v)$, audio features $f_t^a = \phi_a(x_t^a)$, and semantic features $f_t^s = \phi_s(x_t^s)$for all frames
2. Concatenate multimodal features $f_t = [f_t^v, f_t^a, f_t^s]$
3. Compute temporal variation $\Delta_t = \| f_t - f_{t-1} \|$
4. Detect segment boundaries using threshold τbased on Δ_t
5. Partition video into segments $S = \{S_1, S_2, \ldots, S_K\}$

6. Compute segment-level features F_k^v, F_k^a, F_k^s using average pooling
7. Calculate adaptive modality weights α_k^m using softmax normalization
8. Perform semantic-aware fusion $F_k = \sum_{m\in\{v,a,s\}} \alpha_k^m F_k^m$
9. Aggregate global representation $F = \frac{1}{K}\sum_{k=1}^{K} F_k$
10. Apply classification layer $Y = \sigma(W_c F + b)$
11. Compute loss using cross-entropy and update model parameters

The Dynamic Adaptive Semantic Fusion Framework (DASFF), proposed in this study, represents multimodal video data by jointly modelling visual, auditory, and semantic features across temporal sequences. A video can be represented by

$$V = \{(x_t^v, x_t^a, x_t^s)\}_{t=1}^{T} \tag{1}$$

Where x_t^v, x_t^a, and x_t^s are the visual, audio, and semantic input respectively at time t, and T denotes the total number of frames. The above equation describes how the multimodal signal evolves over time and serves as the basis for subsequent feature modelling. To extract discriminative features, modality-specific feature extractors are used. The visual, audio, and semantic features are extracted as

$$f_t^v = \phi_v(x_t^v),\ f_t^a = \phi_a(x_t^a),\ f_t^s = \phi_s(x_t^s) \tag{2}$$

where ϕ_v, ϕ_a, and ϕ_s represent deep feature extraction networks. The resultant features from these modalities are combined into a single multimodal feature vector.

$$f_t = [f_t^v, f_t^a, f_t^s] \tag{3}$$

The extracted modalities provide complementary information for scene understanding. To overcome the limitations of fixed segmentation, DASFF employs adaptive segmentation based on semantic scene transitions. The temporal difference between consecutive frames is then calculated to identify meaningful changes in the video sequence.

$$\Delta_t = \| f_t - f_{t-1} \| \tag{4}$$

which determines the magnitude of change within multimodal content. An increase in Δ_t signals a possible semantic transition. A boundary detection function based on thresholds can be expressed as

$$b_t = f(x) = \begin{cases} 1, b_t < t \\ 0, Otherwise \end{cases} \tag{5}$$

where $\tau$is a predefined threshold. When $b_t = 1$, a new segment boundary is created. This process yields a set of adaptively segmented video portions

$$S = \{S_1, S_2, \ldots, S_K\} \tag{6}$$

where K is the number of segments. This adaptive strategy matches segment boundaries to some coherent, meaningful transition in the scene, maintaining temporal continuity and important events. The segment-level representations are

produced by average pooling and multimodal features are fused for each segment (S_k) to extract the local temporal information.

$$F_k^v = \frac{1}{|S_k|}\sum_{t\in S_k} f_t^v \tag{7}$$

$$F_k^a = \frac{1}{|S_k|}\sum_{t\in S_k} f_t^a \tag{8}$$

$$F_k^s = \frac{1}{|S_k|}\sum_{t\in S_k} f_t^s \tag{9}$$

where $|\,S_k\,|$ is the number of frames in S_k. These representations reflect the important information that exists in each segment, but also preserve the distinctiveness of each modality. The relevance of the different modalities could be different in different contexts, so DASFF has a dynamic adaptive weighting mechanism. Each modality's weight is determined with the softmax method

$$\alpha_k^m = \frac{\exp(W_m F_k^m)}{\sum_{m\in\{v,a,s\}} \exp(W_m F_k^m)} \tag{10}$$

where $m \in \{v, a, s\}$, and W_m represents learnable parameters. This formulation allows the weights to be properly normalized and it allows the contribution of each modality to be adapted for a segment. The weighted features for each of the modalities are then fused to create semantic-aware representation for each segment.

$$F_k = \sum_{m\in\{v,a,s\}} \alpha_k^m \cdot F_k^m \tag{11}$$

This approach considers how relevant each type of information is to the context. By combining informative features and reducing the influence of less useful ones, it helps distinguish between different video patterns. The features extracted from each segment are combined to represent the video as a whole

$$F = \frac{1}{K}\sum_{k=1}^{K} F_k \tag{12}$$

This representation combines the temporal relationships and shared meaning across video segments to describe the video as a whole. It is then passed to a fully connected layer, followed by a softmax function that converts the output scores into class probabilities. The class with the highest probability is selected as the final prediction.

$$Y = \sigma(W_c F + b) \tag{13}$$

where W_c and b are learnable parameters, and σdenotes the softmax function. The output Yrepresents the predicted class probabilities. The model is optimized using the cross-entropy loss function defined as

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{14}$$

DASFF combines adaptive segmentation, dynamic modality weighting, and semantic fusion in a unified framework. It separates videos into segments at scene changes and adjusts the importance of each modality based on the context. By integrating this information, DASFF builds a video representation that supports more accurate and reliable classification across diverse scenarios.

4. Experimental Results and Results

The proposed framework is evaluated on the TAU Audio-Visual dataset introduced by Mesaros, Heittola, and Virtanen [12], a benchmark for multimodal scene understanding. The experiments use 12,291 preprocessed video sequences: 6,145 for model training and 6,146 for evaluating classification performance.

Table 1

Experimental Analysis with Performance Measures

Method	Accuracy	Precision	Recall	F1-Score
Base [17]	0.9769	0.9763	0.9760	0.9760
MEAS [3]	0.9614	0.9661	0.9560	0.9602
Transformer [13]	0.9229	0.9217	0.9176	0.9184
VSMHA [4]	0.9681	0.9666	0.9653	0.9653
DASFF	0.9870	0.9865	0.9858	0.9861

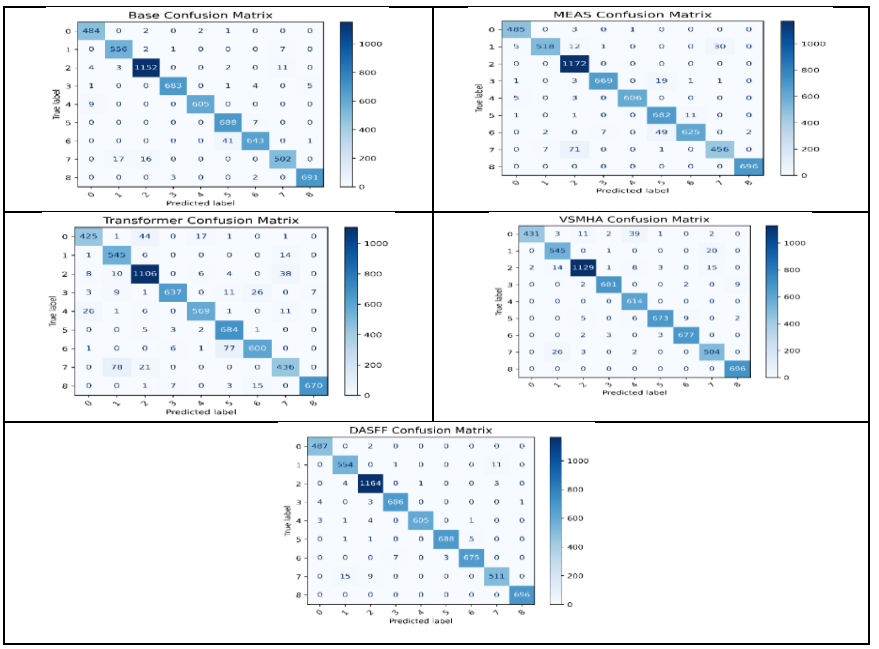


Figure 2

Confusion Matrices for the Existing and Proposed Methods

Table 1 compares the performance of DASFF and existing methods in terms of accuracy, precision, recall, and F1-score. These metrics are also used by Nalini, Raghavendra, and Prasad [15] and Leelavathi, Thahniyath, and Prasad [16]. Although all models perform well, DASFF achieves the highest score on each metric. These findings suggest that integrating adaptive segmentation,

dynamic modality weighting, and semantic-aware fusion enables DASFF to capture multimodal information more effectively.

The confusion matrices in Figure 2 indicate that DASFF distinguishes video scenes more accurately than the other models. Its predictions are more concentrated along the main diagonal, with fewer off-diagonal entries, indicating fewer misclassifications. In contrast, Base, MEAS, Transformer, and VSMHA show greater confusion between scene categories.

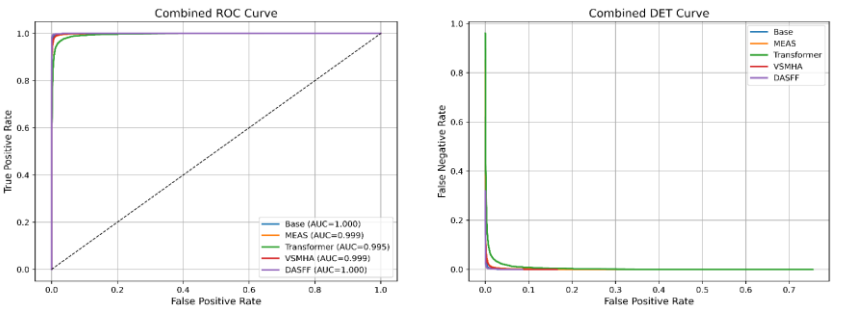


Figure 3

Performance Analysis for the Existing Methods and Proposed DASFF with ROC and DET Curves

Figures 2 and 3 show that all models perform well in classification, with DASFF achieving the highest AUC and the lowest error rates in the DET analysis. These findings suggest that its adaptive segmentation and semantic fusion approach help distinguish between classes more effectively than the compared methods.

Table 2

Statistical Significance of the Proposed DASFF

Comparison	t-value	p-value
DASFF vs Base	4.6549	0.000003
DASFF vs MEAS	10.1549	0.000000
DASFF vs Transformer	18.2357	0.000000
DASFF vs VSMHA	7.8956	0.000000

Table 2 shows that DASFF performs better than the compared methods, with the greatest improvement over the Transformer model. The p-values below 0.001 indicate that these differences are statistically significant. Together, these findings support the effectiveness of DASFF for multimodal scene classification under the evaluated conditions.

5. Conclusion and Scope of the Work

This paper presented a successful multimodal algorithm DASFF for video and audio classification. The adaptive segmentation and temporal modelling,

coupled with the semantic feature extraction and multimodal fusion, enables DASFF to better extract contextual and temporal information than current approaches. Experimental results are shown, showing its superior classification performance, and statistical significance tests are performed, as well as confusion matrix, ROC curve and DET curve. The framework also has good robustness and can be extended for large-scale real-time multi-modal applications.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, pp. 5998–6008 (2017).

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692* (2019), doi: 10.48550/arXiv.1907.11692.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443 (2019), doi: 10.1109/TPAMI.2018.2798607.

[4] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, pp. 1103–1114 (2017).

[5] C.-Y. Wu, S. Zaheer, H. Hu, R. Manmatha, M. Smolić, and P. Krähenbühl, "Compressed video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6026–6035 (2018).

[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] 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.

[8] 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.

[9] B. Gayathri and R. Gopi, "Some properties of k-even mean graph," *J. Discrete Math. Sci. Cryptogr.*, vol. 29, no. 4, pp. 1579–1583 (2025), doi: 10.47974/JDMSC-2244.

[10] 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.

[11] Q. Wei, Y. Zhou, S. Xiang, L. Xiao, and Y. Zhang, "MEAS: Multimodal emotion analysis system for short videos on social media platforms," *IEEE Trans. Comput. Social Syst.*, vol. 12, no. 5, pp. 2398–2410 (2025), doi: 10.1109/TCSS.2024.3490846.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "TAU Urban Audio-Visual Scenes 2021 dataset," in *Proc. Detection Classification Acoust. Scenes Events (DCASE)*, Tampere Univ. (2021).

[13] 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. 7407–7420 (2025).

[14] 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.

[15] C. Nalini, C. Raghavendra, and K. Rajendra Prasad, "Comparative observation and performance analysis of multiple algorithms on Iris data," *Int. J. Pure Appl. Math.*, vol. 116, no. 9, pp. 319–325 (2017).

[16] C. Leelavathi, G. Thahniyath, and K. Rajendra 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).

[17] M. Iqbal, R. K. Sharma, V. Kumar, A. Kushwaha, H. K. Gupta, and M. A. Khan, "Explainable deep learning framework for multimodal emotion recognition using physiological signals," *J. Inf. Optim. Sci.*, vol. 47, no. 7, pp. 2681–2694 (2026), doi: 10.47974/JIOS-2324.

Received March, 2026