ISSN(O): 2320-5407 | ISSN(P): 3107-4928

ISSN NO. 2320-5407

International Journal of Advanced Research

Publisher's Name: Jana Publication and Research LLP

www.journalijar.com

REVIEWER'S REPORT

Manuscript No.: IJAR-59759

Title: Asynchronous, Self-Supervised and Uncertainty-Aware Audio-Visual Fusion for Vector Surveillance in Resource-Constrained Environments.

Recommendation:
Accept as it is .....
Accept after minor revision.....
Accept after major revision ...✓.....
Do not accept (*Reasons below*) .....

Rating	Excel.	Good	Fair	Poor
Originality		✓		
Techn. Quality		✓		
Clarity			✓	
Significance			✓	

Reviewer's ID: JPR- **098**

Detailed Reviewer's Report

Decision: Accept after Major Revision

Major Revisions

1. Insufficient details about the datasets

The manuscript states that public datasets containing mosquito wingbeat recordings and RGB images were used. However, the authors have not clearly identified the datasets, their sources, sample sizes, number of mosquito species, number of audio recordings, and number of visual samples. The authors should provide complete dataset information, including dataset names, publicly accessible sources, sample distributions, and selection criteria.

2. Lack of clarity regarding audio-visual data synchronization

The proposed framework considers visual segments within a temporal window as positive matches for corresponding audio segments. However, the manuscript does not adequately explain how audio and visual samples belonging to the same mosquito or biological event are identified. The authors should clarify the pairing mechanism and explain how the proposed approach

ISSN(O): 2320-5407 | ISSN(P): 3107-4928
International Journal of Advanced Research
Publisher's Name: Jana Publication and Research LLP
www.journalijar.com

avoids incorrect positive associations when the modalities are temporally asynchronous.

3. Insufficient explanation of the proposed uncertainty estimation mechanism

The authors introduce separate uncertainty modules for audio and visual representations. However, the mathematical formulation and training mechanism of these modules are not sufficiently explained. In particular, the authors should clarify how the uncertainty values are estimated, how the uncertainty modules are optimized, and how their outputs influence the final multimodal representation.

4. Mathematical formulations require further clarification

Equations (1)– (4) are central to the proposed methodology. However, the mathematical definitions, variables, parameters, and optimization procedures require more detailed explanations. The authors should clearly define the temporal window, temperature parameter, similarity function, uncertainty values, and modality-weighting coefficients to improve the reproducibility of the proposed framework.

5. Missing implementation and hyperparameter details

The manuscript mentions ResNet-18, MobileNetV2, Adam optimization, a learning rate of (1 times 10^-4), cosine decay, and 100 training epochs. However, important implementation details, including batch size, input dimensions, Mel spectrogram configuration, projection-head dimensions, temperature value, temporal window size, and hardware specifications, are missing. These details should be provided to ensure experimental reproducibility.

6. Insufficient comparison with existing state-of-the-art approaches

The proposed framework is compared with audio-only, visual-only, and supervised multimodal models. However, the comparison with existing self-supervised audio-visual learning and multimodal fusion approaches is limited. The authors should include additional relevant baseline methods and explain

ISSN(O): 2320-5407 | ISSN(P): 3107-4928
International Journal of Advanced Research
Publisher's Name: Jana Publication and Research LLP
www.journalijar.com

how the proposed approach differs from existing multimodal contrastive learning and uncertainty-aware fusion techniques.

7. Lack of comprehensive experimental evaluation

The experimental evaluation primarily reports accuracy and F1-score. Additional evaluation metrics, such as precision, recall, specificity, and confusion matrices, should be included. Furthermore, the authors should report the number of experimental repetitions, mean performance, and standard deviation to demonstrate the reliability and consistency of the reported results.

8. Insufficient validation of robustness under degraded conditions

The manuscript claims improved robustness under acoustic noise, low-light conditions, and temporal misalignment. However, the experimental results do not provide sufficient quantitative evidence for all three conditions. The authors should report detailed performance results under different noise levels, lighting conditions, and temporal shifts. A comparative analysis with baseline models under identical degradation conditions is also necessary.

9. Ablation study is missing

The proposed framework combines multiple components, including self-supervised contrastive learning, asynchronous positive-pair construction, uncertainty estimation, and adaptive fusion. However, the manuscript does not provide an ablation study demonstrating the individual contribution of these components. The authors should evaluate the framework by removing or modifying individual components to establish their actual contribution to overall performance.

10. Computational efficiency and real-world deployment require stronger justification

The authors report an inference time of 42 ms per multimodal sample using a four-core CPU with 8 GB RAM. However, the manuscript does not provide sufficient information about the processor model, memory consumption, model size, parameter count of the complete framework, or computational complexity.

ISSN(O): 2320-5407 | ISSN(P): 3107-4928
International Journal of Advanced Research
Publisher's Name: Jana Publication and Research LLP
www.journalijar.com

REVIEWER'S REPORT

The authors should provide a more comprehensive computational analysis and discuss the practical feasibility of deploying the proposed framework on low-cost embedded devices.

Minor Revisions

1. Improvement of the abstract

The abstract provides a general overview of the proposed methodology and experimental findings. However, the dataset information and the principal methodological contribution should be presented more explicitly. The authors should improve the abstract to make the research novelty and experimental significance clearer.

2. Improvement of the research gap

The Introduction discusses several existing approaches to vector surveillance and multimodal learning. However, the specific research gap is not sufficiently distinguished from the general limitations of existing methods. The authors should clearly identify the limitations of existing asynchronous multimodal learning and uncertainty-aware fusion techniques.

3. Clarification of research contributions

The three contributions presented at the end of the Introduction are relatively broad. The authors should rewrite them to clearly distinguish the methodological novelty, uncertainty-aware fusion mechanism, and experimental contributions of the proposed framework.

4. Improvement of Table 1

Table 1 presents a comparison of candidate encoder architectures. However, the presentation and organization of the table require improvement. The authors should clearly distinguish audio encoders, visual encoders, and projection heads. Furthermore, the parameter counts and computational complexity should be verified and presented consistently.

5. Improvement of Figure 1

ISSN(O): 2320-5407 | ISSN(P): 3107-4928
International Journal of Advanced Research
Publisher's Name: Jana Publication and Research LLP
www.journalijar.com

REVIEWER'S REPORT

Figure 1 illustrates the overall architecture of the proposed framework. However, the figure should be made more informative and self-explanatory. The authors should clearly indicate the audio and visual processing branches, projection heads, uncertainty estimation modules, contrastive learning mechanism, and adaptive fusion stage.

6. Improvement of Algorithm 1

Algorithm 1 presents the training procedure of the proposed framework. However, some implementation steps require further clarification, particularly the construction of positive and negative pairs, uncertainty estimation, and parameter optimization. The authors should improve the algorithmic presentation to make the procedure easier to understand and reproduce.

7. Consistency of technical terminology

The manuscript uses terms such as self-supervised learning, contrastive learning, asynchronous matching, uncertainty estimation, and adaptive fusion. The authors should ensure consistent usage of these terms throughout the manuscript.

8. Improvement of result tables

Tables 2 and 3 present the classification performance and robustness results. The authors should improve the presentation by providing more detailed quantitative comparisons, consistent numerical precision, and clearly defined performance metrics. The qualitative descriptions such as Moderate, Good, and Very good should also be supported by numerical evidence.

9. Strengthening the discussion section

The Discussion section mainly explains the observed improvements in model performance. The authors should provide a deeper comparison with relevant existing studies and discuss the practical implications, limitations, and potential applications of the proposed framework in real-world vector surveillance.

10. Reference formatting and language quality

ISSN(O): 2320-5407 | ISSN(P): 3107-4928
International Journal of Advanced Research
Publisher's Name: Jana Publication and Research LLP
www.journalijar.com

REVIEWER'S REPORT

The authors should carefully check the reference list for consistency in citation style, journal abbreviations, publication details, DOI formatting, and preprint identification. Furthermore, the entire manuscript should be thoroughly proofread to correct grammatical errors, typographical mistakes, and sentence construction issues.