Multi-Modal Hand-to-Mouth Gesture Recognition in Activity-Oriented RGB-Thermal Footage (Student Abstract)

Glenn Fernandes [1*], Meixi Lu [1*], Farzad Shahabi [1], Jiayi Zheng [1], Aggelos Katsaggelos [2], Nabil Alshurafa [1,3]

[1] Department of Computer Science, Northwestern University, Evanston, IL, USA
[2] Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA
[3] Department of Preventive Medicine, Northwestern University, Chicago, IL, USA
{glenn.fernandes, nabil}@northwestern.edu

Abstract

Health-risk behaviors such as overeating and smoking have a profound impact on public health, making their monitoring and mitigation critical. Wearable RGB-Thermal cameras are being employed to monitor these behaviors by capturing hand-to-mouth (HTM) gestures. However, detection models relying on single modalities—either RGB or thermal—often struggle to accurately distinguish between confounding gestures related to eating or smoking due to inherent sensor limitations, such as sensitivity to lighting conditions or thermal occlusions. We present a family of fusion models that integrate RGB and thermal video data using early-, decision-, and a novel mid-fusion architecture, RGB-Thermal Fusion Video Network (RTFVNet), designed to enhance the recognition of HTM gestures associated with eating and smoking. Our evaluation shows that while decision fusion achieves the highest F1-score of 88% (0.44 TFLOPs), RTFVNet offers an optimal balance between performance (85%) and complexity (0.37 TFLOPs) for gesture classification of eating, smoking, and non-gesture activities.

Introduction

RGB-Thermal wearable Activity-Oriented Camera systems (AOCs) (Fernandes et al. 2024; Pedram et al. 2023) are being used to monitor health-risk behaviors such as eating and smoking. These systems are typically worn as a necklace, with lenses oriented towards the mouth to capture HTM activities from a unique, non-egocentric perspective (Figure 1), which presents additional challenges for designing AI models for gesture classification and detection. While previous studies have focused on gesture recognition using either RGB or thermal data, little research has explored distinguishing multiple gestures like eating and smoking using both modalities in activity-oriented footage. RGB modality excels at capturing detailed visuals but struggles in low-light conditions, such as when eating in a dimly lit environment. Conversely, the thermal modality is adept at detecting heat signatures, making it particularly useful for recognizing eating-related HTM gestures, where food typically has a distinct temperature, and smoking-related HTM gestures, where the cigarette tip emits a specific heat signature. However, despite their effectiveness in low-visibility conditions, thermal sensors are affected by ambient temperature variations or occlusions, such as when smoking inside a vehicle or when the cigarette tip is blocked by ones hand. These limitations contribute to higher misclassification rates and reduced reliability in gesture recognition. By fusing RGB and thermal modalities, we integrate RGB's visual detail with thermal heat detection, significantly improving the performance and robustness of multi-class gesture classification, specifically distinguishing between eating, smoking, and non-gesture activities. We present the following key contributions:

1. A novel RGB-Thermal video dataset comprising 441.9 hours of data captured in free-living environments.
2. A family of fusion models, including early-, mid-, and decision-fusion approaches, for multi-class classification to distinguish between eating, smoking, and non-activity.
3. RTFVNet, a novel mid-fusion architecture tailored for multi-class HTM gesture recognition in AOC video footage.

Methodology

We used the state-of-the-art video classification architecture, TimeSformer (Bertasius, Wang, and Torresani 2021), to train models on individual RGB and thermal modalities for multi-class classification of eating, smoking, and non-gesture activities. For decision fusion, we combined the outputs from two separate models: the RGB-TimeSformer and the thermal-TimeSformer. In the early fusion approach, RGB and thermal frames were concatenated prior to being inputted into the TimeSformer model. Furthermore, we developed a novel mid-fusion architecture inspired by RTFNet (Sun, Zuo, and Liu 2019), a state-of-the-art architecture for RGB-Thermal-based segmentation. While RTFNet utilizes shortcut connections between encoders to pass feature maps from the thermal encoder to the RGB encoder, which is advantageous for segmentation tasks. To adapt it to our multi-class classification problem, we remove these cross-modal connections. This modification aims to ensure that each encoder specializes in extracting the most relevant features from its respective modality, minimizing cross-modal interference and enhancing the model's ability to leverage distinct modality-specific information prior to fusion (Nagrani et al. 2021).

*These authors contributed equally.

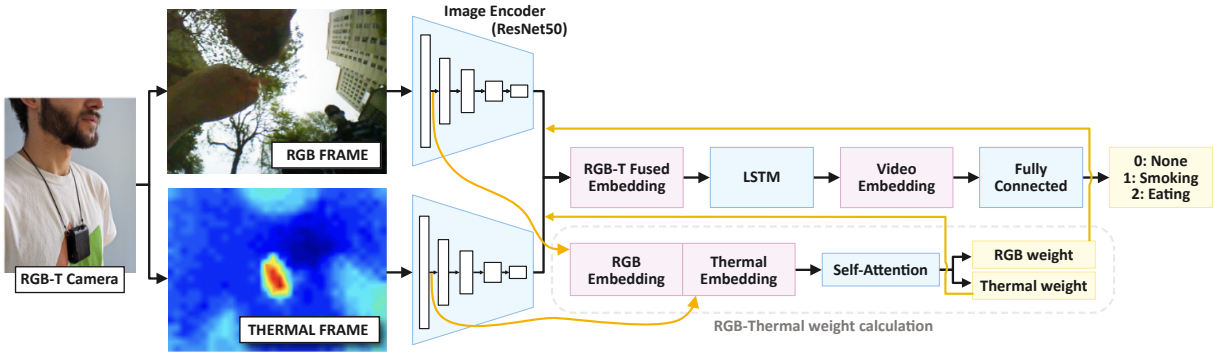

Figure 1: RTFVNet Model Architecture

RTFVNet. (Figure 1) In our novel RTFVNet model, the initial RGB and thermal frames are fed into an image encoder based on ResNet50, where only the early layers of the ResNet model are used to extract RGB and thermal embeddings separately. These embeddings are concatenated into an RGB-T embedding that captures complementary information from both modalities. A self-attention mechanism then computes the relative importance (weights) of each modality—RGB and thermal. The calculated weights are applied to the respective latent representations at the end of the encoding process, dynamically adjusting their contributions for each RGB-T image pair within the model. The latent vectors are fused and passed through an LSTM layer to capture temporal dependencies across frames. Finally, a fully connected layer outputs a classification into three categories: none, smoking, or eating; leveraging both spatial and temporal features for robust behavior detection. All models were evaluated using leave-one-participant-out cross-validation, with performance metrics in terms of F1-score, precision, recall and complexity measured by TFLOPs.

Experiments

Dataset. We collected 441.9 hours of RGB-Thermal data from 7 smokers using HabitSense (Fernandes et al. 2024), an RGB-Thermal activity-oriented, neck-worn camera system designed with lenses directed toward the wearer's face. The start and end points of eating and smoking gestures were annotated.

Experimental Results We evaluated the performance of our mid-fusion architecture, RTFVNet, using leave-one-participant-out cross-validation and compared it against single-modality and other fusion techniques based on F1-scores and computational costs (TFLOPs) [Table 1]. The RGB and thermal models achieved F1-scores of 82.6% and 79.6%, respectively, with computational costs of 0.22 TFLOPs each. While the decision fusion model achieved the highest performance with an F1-score of 87.7%, it incurred a significantly higher computational cost of 0.44 TFLOPs. Our RTFVNet model balanced performance and efficiency, achieving an F1-score of 84.8% with a reduced computational cost of 0.37 TFLOPs compared to decision fusion.

Model	F1-Score	Precision	Recall	TFLOPs
RGB	82.6 ± 2.4	84.0 ± 1.8	82.6 ± 2.3	0.22
Thermal	79.6 ± 2.0	80.7 ± 1.9	79.7 ± 2.0	0.22
Decision Fusion	87.7 ± 1.6	88.7 ± 1.4	87.7 ± 1.5	0.44
Early Fusion	83.8 ± 2.2	84.6 ± 2.2	83.9 ± 2.2	0.22
Mid (RTFVNet)	$\mathbf{84.8 \pm 2.2}$	$\mathbf{84.8 \pm 2.2}$	$\mathbf{85.6 \pm 2.1}$	**0.37**

Table 1: Performance metrics (%, mean ± standard error) for different models, and computational cost (TFLOPs).

Conclusion and Future Work

Our findings indicate that decision fusion is the most effective method for recognizing eating, smoking, and non-gesture activities. However, its computational expense, requiring two separate models, can pose challenges when designing real-time pipelines. In contrast, RTFVNet provides a relatively efficient solution with comparable performance using a single model. In future work, we plan to further optimize the mid-fusion approach for real-time use cases in health-risk behavior detection. We will also expand the dataset and explore advanced multimodal pretraining approaches (Girdhar et al. 2023), to improve both performance and robustness.

Acknowledgements

This work was supported by the National Institute of Health (NIH) under award numbers R21EB030305, R03DK127128, K25DK113242, and R01DK129843.

References

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Fernandes, G. J.; Zheng, J.; Pedram, M.; Romano, C.; Shahabi, F.; Rothrock, B.; Cohen, T.; Zhu, H.; Butani, T. S.; Hester, J.; et al. 2024. HabitSense: A Privacy-Aware, AI-Enhanced Multimodal Wearable Platform for mHealth Applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3): 1–48.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 15180–15190.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34: 14200–14213.

Pedram, M.; Fernandes, G.; Romano, C.; Wei, B.; Sen, S.; Hester, J.; and Alshurafa, N. 2023. Experience: Barriers and opportunities of wearables for eating research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–8.

Sun, Y.; Zuo, W.; and Liu, M. 2019. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3): 2576–2583.