CSIT940 Research Methodology
Annotated Bibliography

Yinqiao Li (yl800@uowmail.edu.au)
UOW Number: 8455569
CCNU Number: 2023124188
University of Wollongong
Central China Normal University

March 12, 2024

Utilizing Facial Expression and rPPG Signals for Emotion Recognition in Classroom Settings

The integration of facial expression analysis and physiological signals such as common physiological approaches like ECG (Electrocardiography), PPG (Photoplethysmography), specifically remote rPPG (Photoplethysmography) holds promise for enhancing the accuracy of affective computation. The multi-model approach aims to address the challenge of comprehensively measuring and characterizing learning emotions, which plays a profound significance in education psychology and education evaluation. By leveraging both facial expressions and physiological indicators like HR (Heart rate) and HRV (Heart rate variability), researchers aim to develop effective methods for understanding and quantifying students' emotional states during learning activities.

Understanding students' emotional states during learning is crucial for enhancing educational experiences and providing timely support. The integration of facial expression analysis and rPPG signals offers a holistic picture to in-class education evaluation, as it combines both visual and physiological signals. The selection of this topic stems from the shortness of the current single model evaluation approach in classroom settings, including several aspects.

From the perspective of feature extraction, existing methods exhibit poor adaptability to the complex learning environments of classrooms. Phenomena such as students' spatial distribution, movements, and occlusions in classroom videos can lead to unstable extraction of rPPG signals. Moreover, the significant disparity in the frequency of emotional states over prolonged classroom durations results in imbalanced categories, which in turn affects the accuracy of measurement outcomes.

From the perspective of feature fusion, there is a lack of interpretability when considering the classroom context. In theory, incorporating more modalities and features can lead to more accurate representation and recognition. However, there is a significant disparity in the difficulty and interpretability of fusing different features, especially in conventional classroom settings. The interpretability of feature fusion is crucial for a deeper understanding of emotional changes during classroom learning.

Literature A: Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals
Fan H, Zhang X, Xu Y, et al. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals[J]. Information Fusion, 2024, 104: 102161. [1]
Depression detection is vital for early diagnosis, prompting the development of automatic multimodal detection methods. This work proposes Transformer-based feature enhancement networks for multimodal depression detection, integrating video, audio, and rPPG signals. The research focuses on assessing the effectiveness of the proposed method on depression detection tasks. The method provides valuable insights into improving depression detection through multimodal approaches, which is beneficial for research in this area. The study's limitation lies in the focus on depression detection without addressing other mental health conditions. The experimental results demonstrate the validity of the proposed method in depression detection tasks, indicating its potential for practical applications. Although not the primary focus of the research, the proposed method offers valuable supplementary information for understanding depression detection mechanisms and enhancing research in this field. Incorporating the audio modality may not be very helpful for noisy classroom environments, but there are currently few applications of Transformer-enhanced networks for emotion detection. If lightweight implementation can be achieved, it would greatly benefit my

research.

Literature B: Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement

Liu X, Hill B, Jiang Z, et al. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement[C]. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023: 5008-5017. [2]

Liu et al. introduce EfficientPhys, a novel approach for camera-based physiological measurement, aiming to simplify and improve the accuracy of contactless vitals measurement. Unlike prior research, which relies on complex preprocessing steps and architectures, EfficientPhys eliminates the need for face detection, segmentation, normalization, and color space transformation. The proposed neural models achieve strong accuracy on multiple public datasets while significantly reducing computational requirements. The paper demonstrates the effectiveness of both transformer and convolutional backbone architectures and highlights a 33% improvement in efficiency with the most lightweight network. This article inspires me to implement lightweight solutions.

Literature C: Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements

Narayanswamy G, Liu Y, Yang Y, et al. Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 7914-7924. [3]

Girish et al. propoesd BigSmall, an efficient architecture for physiological and behavioral measurement inspired by human visual perception principles. BigSmall presents a novel multi-branch network with wrapping temporal shift modules, designed to achieve efficiency gains and comparable accuracy to task-optimized methods. The architecture integrates camera-based facial action, cardiac, and pulmonary measurement tasks into a unified model, demonstrating significant reductions in computational cost while maintaining state-of-the-art accuracy. The paper also proposes mixed spatial and temporal scales for efficient spatiotemporal modeling and develops a Wrapping Temporal Shift Module to enhance temporal feature representation. Extensive evaluations across multiple real-world video-based human physiology datasets validate the utility and effectiveness of BigSmall. The dual-branch network model proposed by the authors in this article fully utilizes the characteristics of feature extraction from different modalities. Adopting low-resolution input for rPPG extraction is highly instructive for reducing

model complexity.

Literature D: CNN-LSTM for automatic emotion recognition using contactless photoplythesmographic signals

Mellouk W, Handouzi W. CNN-LSTM for automatic emotion recognition using contactless photoplythesmographic signals[J]. Biomedical Signal Processing and Control, 2023, 85: 104907. [4]

In this article, Mellouk et al. extracted rPPG from facial videos and calculated HR for emotion classification. The paper presents a novel framework for automatic emotion recognition based on non-contact physiological signals, specifically photoplethysmographic (PPG) signals extracted from facial videos. The study leverages deep learning techniques, employing a one-dimensional convolutional neural network (1DCNN) and a long short-term memory (LSTM) network for classification after normalization and segmentation of the signals. Evaluation on the MAHNOB-HCI emotional database yielded recognition rates of 73.33% and 60% for binary classification of valence and arousal, respectively, with a signal segmentation of 4 seconds. The paper aims to contribute to the field of automatic emotion recognition by demonstrating the feasibility of contactless physiological signal extraction methods.

Literature E: Trusted emotion recognition based on multiple signals captured from video

Zhang J, Zheng K, Mazhar S, et al. Trusted emotion recognition based on multiple signals captured from video[J]. Expert Systems with Applications, 2023, 233: 120948. [5]

The paper addresses the challenge of recognizing fake emotions by proposing a multimodal approach leveraging facial expressions, eye states, and physiological signals captured from video. The method utilizes a graph neural network to extract spatial and spectral features from facial images, a model-based approach for decomposing RGB signals into heart rates, and a deep learning model for eye region segmentation. Fusion strategies are applied to evaluate emotion recognition performance based on multiple signals. Experimental validation on various datasets demonstrates the effectiveness of multimodal approaches, with eye state emerging as a reliable cue for trusted emotion recognition. The contributions of the research include the creation of a new dataset for training eye structure segmentation models and the development of a trusted emotion recognition framework integrating facial expressions, eye states, and noncontact physiological signals. The usability of extracting eye state features used in this paper for emotion recognition is

highly inspiring and relevant to my research.

References

[1] H. Fan, X. Zhang, Y. Xu, J. Fang, S. Zhang, X. Zhao, and J. Yu, "Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals," vol. 104. Elsevier, 2024, p. 102161.

[2] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 5008–5017.

[3] W. Mellouk and W. Handouzi, "Cnn-lstm for automatic emotion recognition using contactless photoplythesmographic signals," vol. 85. Elsevier, 2023, p. 104907.

[4] G. Narayanswamy, Y. Liu, Y. Yang, C. Ma, X. Liu, D. McDuff, and S. Patel, "Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7914–7924.

[5] J. Zhang, K. Zheng, S. Mazhar, X. Fu, and J. Kong, "Trusted emotion recognition based on multiple signals captured from video," vol. 233. Elsevier, 2023, p. 120948.