Fusing Physical and Cognitive Stimuli: An Eye Movement Emotion Recognition Framework Based on Hierarchical Attention Mechanism

Extended Abstract

ZhiLin Li
Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education,Guangxi Normal University,Guilin,541004,China
Guangxi Key Lab of Multi-Source Information Mining and Security,Guangxi Normal University,Guilin,541004,China
lizhilin@stu.gxnu.edu.cn

Xiaomei Tao*
Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education,Guangxi Normal University,Guilin,541004,China
Guangxi Key Lab of Multi-Source Information Mining and Security,Guangxi Normal University,Guilin,541004,China
xiaomei.tao@gxnu.edu.cn

Abstract

Eye movement signals, as physiological signals that are resistant to interference and closely related to emotions, have been widely applied in multimodal emotion recognition research. However, existing studies often focus on directly utilizing eye movement signals for emotion recognition, with few exploring the interaction between different video stimuli and eye movement signals from the perspective of Human-Computer Interaction (HCI). Research in cognitive neuroscience has revealed that eye movement signals are not only directly influenced by physical visual information such as light and brightness during observation but also by high-level visual features resulting from top-down cognitive processing of the stimulus materials in the brain. The sequential stimulation of these two types of visual features both affects eye movement signals and reflects the corresponding emotional states through these signals. Inspired by these findings, this study designs a hierarchical attention mechanism-based emotion recognition framework(HAMER) to simulate the process by which eye movement signals respond to stimuli, enabling interaction between video information and physiological signals in HCI. The framework demonstrates excellent performance on two emotion recognition datasets, VLMED and MAHNOB-HCI, which contain eye movement signals and video information, providing innovative perspectives and empirical support for emotion recognition based on physiological signals and video information in the field of HCI.

Keywords

Emotion recognition; Human-computer interaction; Cognitive Computing; Eye tracking; Attention mechanism;

ACM Reference Format:
ZhiLin Li and Xiaomei Tao. 2025. Fusing Physical and Cognitive Stimuli: An Eye Movement Emotion Recognition Framework Based on Hierarchical Attention Mechanism: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

*Corresponding author

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 Introduction

In Human–Computer Interaction (HCI), timely and accurate emotion recognition enables intelligent systems to deliver more intelligent and personalized responses based on the user's emotional state.Eye movement signals, as non-invasive physiological signals, have emerged as a crucial modality in emotion recognition due to their resistance to environmental noise and minimal interference[9]. Studies have demonstrated that eye movement signals, including fixations, saccades, blinks, and pupil size, are indicative of the user's emotional state[11].

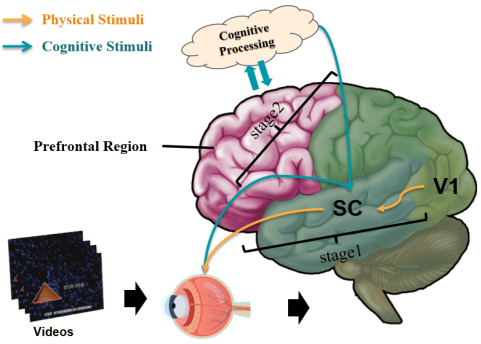

Figure 1: The process by which the brain processes different visual information and stimulates eye movement signals

Research in cognitive neuroscience and related fields has confirmed that eye movement signals are not only directly influenced by physical visual information, such as light and brightness, during observation, but also by cognitive features resulting from top-down cognitive processing in the brain.[6–8]. Figure 1 illustrates the process through which the brain processes visual information in different level and stimulates eye movements. After visual information enters the brain through the retina, in the first stage, physical visual information such as color, brightness, and saturation is directly processed in the primary visual cortex (V1). The processed physical visual features are then transmitted to the superior colliculus (a key structure in the brain responsible for processing visual information and controlling eye movements), which directly stimulates

the oculomotor nerves and influences the eye movement signals. In the second stage, high-level visual information undergoes cognitive processing in the prefrontal cortex and is then returned to the superior colliculus through a feedback mechanism, modulating the eye movement signals.

Studies have shown that the processing of primary visual information(i.e., the physical visual information in this paper) in the brain's primary visual cortex is completed almost instantaneously, very quickly and briefly[3, 5], while the cognitive processing of high-level visual brain regions takes several hundred milliseconds to a few seconds to complete[13]. Therefore, the direct stimulation of the oculomotor nerves by physical visual information and the modulation of the oculomotor nerves by high-level visual information after cognitive processing in the brain is a sequential and continuously iterative process.In the scenario of watching a video, the physical visual features of the video, such as color, light, and brightness, directly affect eye movement characteristics, while the semantic information in the video undergoes cognitive processing in the brain and then influences the eye movement behavior.

Inspired by the results of the above cognitive neuroscience research, this study integrates physical and cognitive stimuli to propose a hierarchical attention mechanism-based emotion recognition framework(HAMER) .The framework uses video frame information such as brightness, color, and saturation to simulate the physical visual features that stimulate eye movement signals, and utilizes the picture description of the video frame to simulate high-level cognitive features formed in the brain through cognitive processing.By sequentially capturing the effects of physical and cognitive features on eye movement signals, the framework simulates the process of eye movement signal response to stimuli, facilitating Human–Computer Interaction (HCI) between video information and physiological signals. Ultimately, the model achieves accurate emotion classification of the participants.

2 PREVIOUS WORK

Physiological signals can objectively reflect the emotional state of the subject without being easily controlled or disguised by subjective awareness. Therefore, many emotion recognition studies have focused on recognizing the emotional state of subjects by integrating different physiological signals. For example, Tao et al.[12]used video stream sequences to capture facial expressions and eye movement data as bimodal inputs into a data flow framework. Fu et al.[4] proposed a cross-modal guiding neural network for electroencephalogram (EEG) and eye movement signals, using EEG features to guide eye movement feature extraction. Zhao et al.[16] introduced a gradient neural network (DGNN) model that integrates eye movement and EEG signals. These studies achieved good results in emotion recognition accuracy. However, these works were limited to considering only the participants' physiological signals, neglecting the influence of the stimulus materials on the participants during the emotion generation process. Therefore, they did not achieve Human–Computer Interaction (HCI) between video information and physiological signals, which limited the effectiveness of the models. Additionally, emotional expression and interpretation often depend on specific social and situational contexts, and interpreting physiological signals out of context may lead to misinterpretation [2].

Therefore, another line of research (our earlier work) attempted to combine video information with physiological signals. Ye et al.[14], for example, introduced video information such as brightness, emotional induction time, and video click rate into the physiological signals, achieving good results across multiple datasets and demonstrating that video information also impacts learners. Bao et al. [1, 10] analyzed the impact of stimulus materials on the subjects by calculating the pixel change rate of each frame and MFCC coefficients as audiovisual features of the stimuli. However, these works only directly fused video information with physiological signals, without considering the deep interrelation between video information and physiological signals during modality fusion. Therefore, this paper proposes an emotion recognition model that integrates different video information and eye movement signals, aiming to perform modality fusion from the perspective of Human–Computer Interaction (HCI), making the model more aligned with the interactive process between various video information and eye movement signals in real-world scenarios.

3 CURRENT WORK

To simulate the sequential and continuous physical and cognitive stimulus process of video content on eye movement signals, we extracted video frame information such as brightness, hue, and saturation as the physical visual features perceived by the participants. Meanwhile, we employed the Zhang's method[15] to obtain video frame description information as the cognitive features formed in the participants' brain after cognitive processing. We then used LSTM to extract the temporal features of eye movement signals and physical visual features. Finally, we designed a hierarchical attention mechanism that sequentially extracts the stimulus effects of primary physical visual features and high-level cognitive features on eye movement signals, retaining the results of each extraction for input into the next layer of the framework. The model achieved good performance on the VLMED and MAHNOB-HCI public datasets.

4 FUTURE WORK

The feedback mechanisms in the nervous system are much more complex than described in this paper, and there are many other factors that influence physiological signals, such as prior knowledge. Therefore, in future work, we will also consider individual prior knowledge to provide a more accurate and complete description of the process. Additionally, changes in scene lighting (or display modes in the video, such as images with prominent visual features) can also affect saccade patterns and pupil size variations,which have not been considered in this study.As a result,in future work,we will need to preprocess and remove lighting effects to mitigate this issue.Lastly,the current experiment lacks a sufficient number of videos in the two datasets, the effectiveness of the model in more complex scenes and datasets with a greater number of videos remains uncertain. Hence, in future work, we will validate our findings on more complex datasets.

Acknowledgments

This work was supported by the National Natural Science(No.61906051 and No.62267001).

References

[1] Jindi Bao, Xiaomei Tao, and Yinghui Zhou. 2022. An emotion recognition method based on eye movement and audiovisual features in MOOC learning environment. *IEEE Transactions on Computational Social Systems* 11, 1 (2022), 171–183.

[2] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.

[3] Russell L DeValois and Karen K DeValois. 1990. Spatial vision. (1990).

[4] Baole Fu, Wenhao Chu, Chunrui Gu, and Yinhua Liu. 2024. Cross-modal Guiding Neural Network for Multimodal Emotion Recognition from EEG and Eye Movement Signals. *IEEE Journal of Biomedical and Health Informatics* (2024).

[5] David H Hubel and Torsten N Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology* 160, 1 (1962), 106.

[6] Peter König, Niklas Wilming, Tim C Kietzmann, Jose P Ossandón, Selim Onat, Benedikt V Ehinger, Ricardo R Gameiro, and Kai Kaspar. 2016. Eye movements as a window to cognitive processes. *Journal of eye movement research* 9, 5 (2016), 1–16.

[7] Lisa M Kroell and Martin Rolfs. 2022. Foveal vision anticipates defining features of eye movement targets. *Elife* 11 (2022), e78106.

[8] Xue Liu, Hongren Huang, Terrance P Snutch, Peng Cao, Liping Wang, and Feng Wang. 2022. The superior colliculus: cell types, connectivity, and behavior. *Neuroscience Bulletin* 38, 12 (2022), 1519–1540.

[9] Jaromir Przybyło, Eliasz Kańtoch, and Piotr Augustyniak. 2019. Eyetracking-based assessment of affect-related decay of human performance in visual tasks. *Future Generation Computer Systems* 92 (2019), 504–515.

[10] Xianhao Shen, Jindi Bao, Xiaomei Tao, and Ze Li. 2022. Research on emotion recognition method based on adaptive window and fine-grained features in MOOC learning. *Sensors* 22, 19 (2022), 7321.

[11] Vasileios Skaramagkas, Giorgos Giannakakis, Emmanouil Ktistakis, Dimitris Manousos, Ioannis Karatzanis, Nikolaos S Tachos, Evanthia Tripoliti, Kostas Marias, Dimitrios I Fotiadis, and Manolis Tsiknakis. 2021. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering* 16 (2021), 260–277.

[12] Xiaomei Tao, Shengxi Liu, and Xinyi Chen. 2020. Dual flow framework on bimodality emotion recognition based on facial expression and eye movement. In *2020 International Conference on Artificial Intelligence and Education (ICAIE)*. IEEE, 127–133.

[13] Anne M Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cognitive psychology* 12, 1 (1980), 97–136.

[14] Hanmin Ye, Yinghui Zhou, and Xiaomei Tao. 2023. A Method of Multimodal Emotion Recognition in Video Learning Based on Knowledge Enhancement. *Computer Systems Science & Engineering* 47, 2 (2023).

[15] Yuan Zhang, Xiaomei Tao, Hanxu Ai, Tao Chen, and Yanling Gan. 2024. Multimodal Emotion Recognition by Fusing Video Semantic in MOOC Learning Scenarios. *arXiv preprint arXiv:2404.07484* (2024).

[16] Li-Ming Zhao, Rui Li, Wei-Long Zheng, and Bao-Liang Lu. 2019. Classification of five emotions from EEG and eye movement signals: complementary representation properties. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 611–614.