https://doi.org/10.2352/EI.2026.38.10.HVEI-226

Enhancing Emotion Estimation Accuracy through Integrated Analysis of Heart Rate and Pupil Signals

Tsukasa Yano; Graduate School of Engineering, Chiba University; Chiba, Japan
Midori Tanaka; Graduate School of Informatics, Chiba University; Chiba, Japan
Takahiko Horiuchi; Graduate School of Informatics, Chiba University; Chiba, Japan

Abstract

Emotion recognition using physiological signals is often limited by unimodal analysis, which fails to capture interactions across physiological systems. This study proposes a multimodal framework that integrates heart rate (HR) and pupil diameter signals, with a particular focus on modeling cross-modal interactions. We introduce composite features that explicitly represent relationships between HR and pupil dynamics, combined with a two-step feature optimization strategy using correlation-based reduction and mutual information ranking. Experiments were conducted on an emotion-elicitation dataset with three emotional states (Joy, Neutral, Sad), using multiple classifiers and cross-validation schemes. The proposed method achieved a classification accuracy of 91.1%, significantly outperforming HR-only (61.1%) and pupil-only (72.2%) approaches. Feature analysis revealed that cross-modal descriptors, particularly an entropy-based interaction feature, contributed most to performance improvement. These results demonstrate that explicitly modeling cross-modal physiological interactions provides an effective strategy for enhancing emotion recognition accuracy.

Introduction

Emotion recognition plays an important role in human-computer interaction, affective computing, and healthcare applications. Physiological signals provide an objective means of estimating emotional states because they reflect autonomic nervous system responses that are difficult to control voluntarily. Among various biosignals, heart rate (HR) and pupil diameter are known to be sensitive to emotional arousal and cognitive processes. Previous studies have explored emotion recognition using individual physiological modalities, such as heart rate variability [1] or pupil dynamics [2]. While these approaches have demonstrated effectiveness, their performance is inherently limited because emotional responses emerge from coordinated activity across multiple physiological systems [3]. As a result, multimodal approaches that integrate multiple biosignals have attracted increasing attention. However, existing multimodal methods still face two key challenges. First, feature sets derived from multiple physiological signals often contain redundant information due to strong correlations, which can degrade model stability and performance. Second, and more importantly, interactions between modalities are rarely modeled explicitly. Most approaches simply concatenate features from different signals, failing to capture the coordinated but non-identical dynamics that characterize emotional responses.

To address these limitations, this study focuses on explicitly modeling cross-modal interactions between physiological signals. We propose a multimodal emotion recognition framework that integrates HR and pupil diameter signals and introduces composite features designed to represent relationships between the two modalities. In addition, a two-step feature optimization strategy is employed, combining correlation-based feature reduction to remove redundancy and mutual information (MI) ranking to identify informative features. The proposed framework is evaluated using an emotion-elicitation dataset consisting of three emotional states (Joy, Neutral, and Sad), under multiple cross-validation schemes. Experimental results demonstrate that incorporating cross-modal interaction features significantly improves classification performance compared with unimodal and simple multimodal approaches.

The main contributions of this study are as follows: (1) the introduction of composite features that explicitly model cross-modal physiological interactions, and (2) the demonstration that such interaction modeling substantially enhances emotion recognition accuracy.

Method

Framework Overview

This study proposes a cross-modal interaction modeling framework for physiological emotion recognition using heart rate (HR) and pupil diameter signals. The framework consists of three stages: (1) feature extraction from each modality, (2) construction of interaction-aware composite features, and (3) feature selection and classification. Unlike conventional approaches that treat modalities independently or simply concatenate features, the proposed method explicitly models relationships between modalities.

Feature Extraction

From each recorded time series segment, statistical and nonlinear descriptors were extracted for both HR and pupil signals. For heart rate signals, features included statistical measures such as mean, standard deviation, and variability, as well as nonlinear descriptors such as entropy and spectral power components. For pupil signals, features included mean pupil diameter, variability measures, maximum and minimum values, and entropy-based descriptors representing temporal complexity. In total, 85 candidate features were initially extracted from these physiological modalities.

Cross-Modal Interaction Features

To capture relationships between modalities, composite features are constructed by combining HR and pupil features. Specifically, ratio-based and difference-based features are introduced to represent relative changes between modalities. For example, the ratio of HR entropy to pupil entropy is used to quantify differences in signal complexity. These features are designed to reflect coordinated physiological responses associated with emotional states, which cannot be captured by unimodal features alone.

Feature Selection

The extracted feature set may contain redundant variables that negatively affect model stability. To address this issue, a two-step feature selection strategy was applied. First, correlation-based reduction was performed. If the absolute correlation coefficient between two features exceeded a predefined threshold, one of the features was removed to reduce redundancy. Second, mutual information (MI) analysis was used to evaluate the relevance of each feature with respect to the emotion labels. Features were ranked based on MI scores, and the top-ranked features were selected for classification. This procedure produces a compact feature set that is both informative and non-redundant.

Classification

To evaluate the effectiveness of the proposed features, five widely used machine learning classifiers were employed: K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Gradient Boosting, and AdaBoost. These models were selected because they represent diverse learning principles, including distance-based, tree-based, and ensemble methods. The hyperparameters were set based on preliminary experiments to achieve stable performance. Unless otherwise specified, the remaining parameters were set to their default values. Table 1 summarizes the main hyperparameters used for each classifier.

Table 1. Hyperparameter configuration of the machine learning classifiers used in this study.

Model	Main Hyperparameters
KNN	n_neighbors = 5, weights = distance, p = 1, algorithm = ball_tree
Random Forest	n_estimators = 90, oob_score = True
Decision Tree	max_depth = 6
Gradient Boosting	n_estimators = 120, max_depth = 6, learning_rate = 0.05, min_samples_split = 4, subsample = 0.5
AdaBoost	n_estimators = 200, learning_rate = 0.07

Experiment

Participants

A total of ten university students in their twenties participated in the experiment after providing informed consent. All participants had normal or corrected-to-normal vision and no reported neurological or cardiovascular disorders.

Stimuli and Experimental Procedure

An emotion-elicitation experiment was conducted to collect synchronized heart rate (HR) and pupil diameter signals. Participants viewed emotion-inducing video clips in a dark laboratory environment. The stimuli were presented on an EIZO CG314 monitor, and the viewing distance between the participant and the display was approximately 70 cm. To reduce motion artifacts during measurement, participants' heads were stabilized using a chin rest.

The emotional stimuli consisted of twelve short video clips designed to evoke three emotional states: Neutral, Joy, and Sad. Each emotional category contained four clips, and each participant viewed all twelve clips.

Data Acquisition

Physiological signals were collected during an emotion-elicitation experiment using video stimuli. As mentioned in the Participants section, ten healthy university students in their twenties participated in the experiment after providing informed consent. Heart rate signals were measured using an optical heart rate sensor (Polar Verity Sense), which records HR at a sampling rate of 1 Hz. Pupil diameter and gaze information were recorded using an eye-tracking device (Tobii Pro Glasses 3), which measures pupil dynamics at approximately 50 Hz. The experiment was conducted in a dark room to minimize environmental interference. Participants viewed emotion-eliciting video clips displayed on a monitor positioned approximately 70 cm from the participant. A chin rest was used to stabilize head movements during measurement. The video stimuli consisted of clips designed to evoke three emotional states: Joy, Neutral, and Sad.

Evaluation Protocol

The stimulus presentation order followed a fixed sequence of Neutral → Joy → Sad, which was repeated four times during the experiment, as illustrated in Figure 1. Before stimulus presentation, a three-minute adaptation period was provided to allow participants to adjust to the experimental environment. The video stimuli were selected from a validated empathy-eliciting stimulus dataset reported in Ota et al. [4]

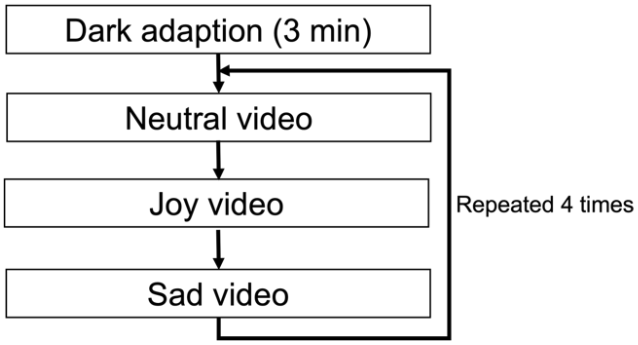


Figure 1 *Experimental procedure for emotion elicitation.*

Results

Feature Reduction

The proposed framework was evaluated using several feature configurations to investigate the effects of multimodal integration and feature optimization. Initially, 85 candidate features were extracted from heart rate (HR) and pupil diameter signals. To reduce redundancy among features, correlation-based feature reduction was applied. When the correlation threshold was set to $|r| > 0.9$, the number of features decreased to 54. Using $|r| > 0.8$ further reduced the feature set to 43, while a stricter threshold of $|r| > 0.7$ reduced the number to 27 features. These results indicate that a considerable part of the extracted features contained redundant information, and that correlation-based reduction effectively produced a more compact feature representation.

Classification Performance

Table 2 summarizes the classification accuracy obtained under different feature configurations. The primary classification results were obtained using Leave-One-Out (LOO) cross-validation. The results demonstrate that multimodal integration substantially improves emotion recognition performance. Using only heart rate (HR) features resulted in a classification accuracy of 61.1%, while

pupil-only features achieved 72.2%. When features from both modalities were combined without cross-modal descriptors, the accuracy increased to 85.4%, indicating the advantage of integrating physiological signals. Introducing composite cross-modal features further improved the performance to 87.8%. These features represent relationships between HR and pupil responses and allow the model to capture interactions between physiological modalities that cannot be observed when each signal is analyzed independently.

The highest performance was obtained when composite features were combined with correlation-based feature reduction using the threshold $|r| > 0.8$. Under this configuration, the Random Forest classifier achieved the best accuracy of 91.1%, suggesting that removing redundant variables improves model stability and discriminative power.

Table 2 Classification accuracy comparison across different feature configurations.

Feature configuration	Accuracy
HR features only	61.1%
Pupil features only	72.2%
HR + Pupil features	85.4%
Composite features	87.8%
composite + correlation reduction (|r| > 0.8)	91.1%

To further evaluate the generalization ability of the proposed framework, additional validation strategies were applied, including Leave-One-Subject-Out (LOSO) and Within-Subject Leave-One-Out (WS-LOO) cross-validation. The proposed multimodal model achieved an accuracy of 91.1% under both LOO and LOSO validation. In contrast, WS-LOO resulted in a lower accuracy of 77.1%. This difference may reflect the higher variability between individual stimulus presentations within the same participant, which makes trial-level prediction more challenging than subject-level generalization. The high LOSO accuracy suggests that the proposed framework generalizes well to unseen participants, indicating that the extracted physiological features capture emotion-related patterns that are not strongly dependent on specific individuals.

Statistical Reliability

To assess the statistical reliability of the classification results, a bootstrap analysis was conducted. The bootstrap procedure was repeated 2000 times by resampling the LOSO-based classification results at the subject level with replacement, thereby evaluating the stability of performance across subjects. The 95% confidence intervals were estimated using the percentile method. For the best-performing configuration (Random Forest with composite features and correlation reduction $|r| > 0.8$), the observed accuracy was 0.911, with a 95% confidence interval of [0.833, 0.978]. In comparison, unimodal configurations produced lower performance. The HR-only condition achieved an observed accuracy of 0.656 with a confidence interval of [0.589, 0.733], while the pupil-only condition achieved 0.656 with a confidence interval of [0.567, 0.744].

Figure 2 presents the observed accuracies together with their corresponding 95% confidence intervals for the three feature configurations. The multimodal composite model shows a clearly higher accuracy than the unimodal baselines, and the separation of the confidence intervals indicates that the performance improvement is consistently supported across bootstrap resamples.

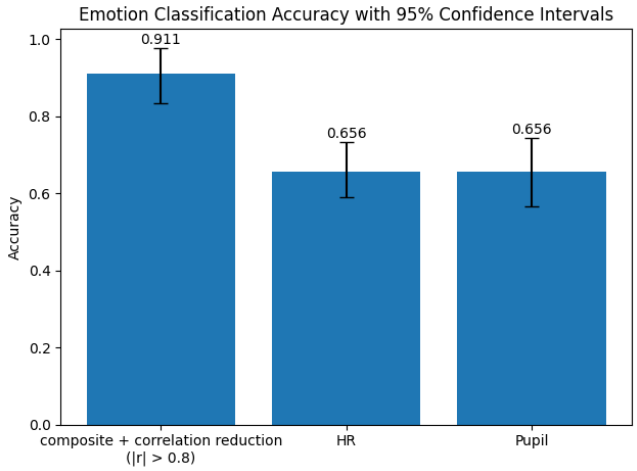


***Figure 2** Bootstrap-based Comparison of Emotion Classification Accuracy Across Feature Configurations.*

Class-wise Evaluation

To further analyze classification performance, class-wise evaluation metrics were calculated for the best-performing model configuration. Table 3 summarizes the precision, recall, and F1-score for each emotional category.

Table 3 Class-wise evaluation metrics (precision, recall, and F1-score) for the best-performing model configuration.

Class	Precision	Recall	F1-score
Joy	0.867	0.867	0.867
Sad	0.931	0.900	0.915
Neutral	0.935	0.967	0.951

The results indicate strong classification performance across all emotional states. The Neutral class achieved the highest recall (0.967), suggesting that neutral emotional states produce relatively stable physiological responses. The Sad class also showed high precision and F1-score values, indicating good separability from the other classes. In contrast, the Joy class exhibited slightly lower performance compared with the other categories, suggesting partial overlap in physiological responses between joyful and other emotional states.

Most Informative Features After Correlation-Based Reduction

To investigate which features contributed most to the classification performance, mutual information (MI) ranking was applied after correlation-based feature reduction. Table 4 lists the five most informative features.

Table 4 Top five features ranked by mutual information after correlation-based feature reduction.

Rank	Feature	Category
1	HR-Pupil Entropy Ratio	Composite
2	Pupil Entropy	Pupil
3	HR First-Difference Energy	HR
4	HR High-Frequency Power	HR
5	Pupil Diameter Variability	Pupil

The highest-ranked feature was the HR-Pupil entropy ratio, which represents the interaction between heart rate and pupil dynamics. This result suggests that cross-modal physiological relationships provide strong discriminative information for emotion recognition. In addition, the presence of both HR-based and pupil-based features among the top-ranked variables indicates that combining information from multiple physiological modalities contributes to improved classification performance.

Among the unimodal features, pupil entropy measures the irregularity of the pupil diameter series and is computed using Shannon entropy applied to pupil signal distribution. HR first-difference energy quantifies short-term heart rate fluctuations by summing the squared differences between consecutive heart rate samples. HR high-frequency power represents the spectral energy of the heart rate signal in the high-frequency band (0.15–0.4 Hz), which is commonly associated with parasympathetic nervous system activity [5]. Finally, pupil diameter variability is defined as the standard deviation of the pupil diameter time series, reflecting the magnitude of pupil fluctuations during stimulus presentation.

Discussion

The experimental results demonstrate that modeling cross-modal interactions between heart rate (HR) and pupil signals plays a central role in improving emotion recognition performance. In particular, the strong contribution of composite features, such as the HR-pupil entropy ratio, suggests that emotional states are better characterized by relationships between physiological signals rather than by each modality independently.

One possible explanation for this result is that HR and pupil responses reflect different aspects of autonomic and cognitive processes. Heart rate is primarily associated with autonomic regulation, including sympathetic and parasympathetic activity, whereas pupil dynamics are influenced not only by autonomic responses but also by cognitive and attentional processes. Therefore, emotional states may not be fully represented by the magnitude or variability of a single signal, but rather by the coordination and relative changes between multiple physiological systems. The composite features introduced in this study capture such coordinated dynamics, enabling the model to distinguish emotional states more effectively. In addition, the effectiveness of entropy-based interaction features suggests that temporal complexity plays an important role in emotion representation. Differences in entropy between HR and pupil signals may reflect variations in the stability and irregularity of physiological responses under different emotional conditions. By modeling these relative differences, the proposed approach captures information that is not available from individual signals alone.

The results also indicate that simple feature concatenation is insufficient for fully exploiting multimodal information. Although combining HR and pupil features improves performance, the introduction of explicit interaction features leads to further gains. This supports the idea that multimodal integration should not only aggregate information but also represent relationships between modalities. Furthermore, correlation-based feature reduction contributes to performance improvement by removing redundant variables that may obscure meaningful cross-modal relationships. This suggests that effective interaction modeling requires not only the design of composite features but also careful control of feature redundancy. Despite these findings, several limitations should be noted. First, the number of participants was relatively small, which may limit the generalizability of the results. Second, the stimulus presentation followed a fixed order, which may introduce order effects such as habituation or fatigue. Future studies should employ randomized stimulus sequences and larger, more diverse datasets to validate the robustness of the proposed approach.

Overall, the results suggest that explicitly modeling cross-modal physiological interactions provides a more informative representation of emotional states than unimodal or naively integrated features. This highlights the importance of interaction-aware design in multimodal emotion recognition systems.

Conclusion

This study investigated the role of cross-modal interaction modeling in physiological emotion recognition using heart rate and pupil diameter signals. The results demonstrate that explicitly representing relationships between modalities, rather than relying on unimodal features or simple feature concatenation, significantly improves classification performance. In particular, the effectiveness of composite features, such as the HR-pupil entropy ratio, indicates that emotional states are more accurately characterized by coordinated physiological dynamics than by individual signal properties. These findings highlight the importance of interaction-aware feature design for capturing the underlying structure of emotional responses.

The study also shows that appropriate feature selection is essential for maximizing the benefits of interaction modeling, as reducing redundancy helps reveal meaningful cross-modal relationships. However, the generalizability of the results is limited by the small sample size and the fixed stimulus presentation order. Future work should validate the proposed framework using larger and more diverse datasets, as well as randomized experimental designs. In addition, extending the approach to more complex models, such as deep learning architectures, may further improve the integration of cross-modal information.

Overall, this work demonstrates that modeling cross-modal physiological interactions provides an effective and scalable approach for improving emotion recognition performance.

References

[1] L. Shu et al., "Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet," Sensors, vol. 20, no. 3, p. 718, 2020.

[2] P. Tarnowski et al., "Eye-tracking Analysis for Emotion Recognition," Computational Intelligence and Neuroscience, vol. 2020, pp. 1-14, 2020.

[3] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review," IEEE Transactions on Affective Computing, vol. 1, no. 1, pp. 18–37, 2010.

[4] A. Ota, A. Nakane, S. Kumano, A. Murata, and S. Shimizu, "Validation study of empathy-eliciting video stimulus," Proc. IEICE HCG Symposium, No. A-4-3, 2023.

[5] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," Frontiers in Public Health, vol. 5, Art. 258, 2017.

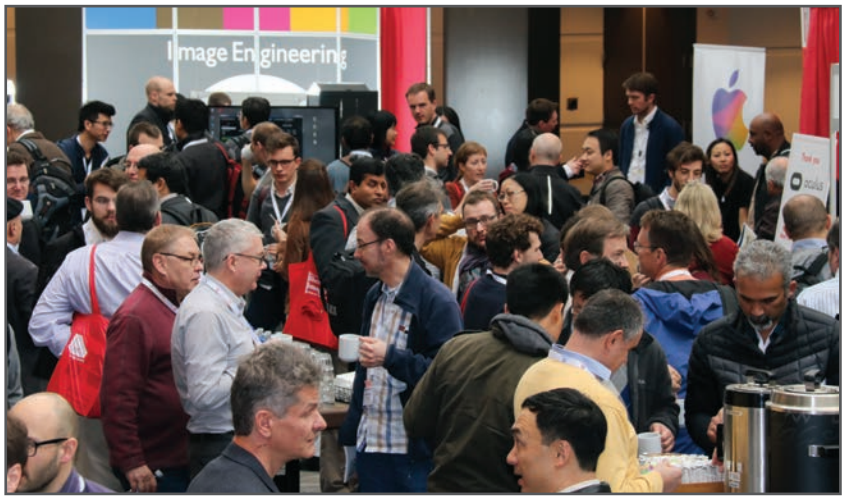
Author Biography

Tsukasa Yano received his Bachelor of Engineering from Chiba University in 2024. He is currently a master's program student in the Graduate School of Engineering at Chiba University. His research focuses on emotion recognition using physiological signals, with particular emphasis on the integrated analysis of heart rate and pupil diameter data. His work aims to improve the accuracy and robustness of emotion estimation through multimodal approaches.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org