EEG for Emotion Recognition

Anchit Gupta
7035463
Saarland University
Saarbrücken, Germany
angu00006@uni-saarland.de

Roba ElQadi
7037144
Saarland University
Saarbrücken, Germany
roel00002@uni-saarland.de

Dinar Orazgaliyev
7056420
Saarland University
Saarbrücken, Germany
dior00002@uni-saarland.de

Paranjoy Gupta
7039391
Saarland University
Saarbrücken, Germany
pagu00003@uni-saarland.de

Pooja Kotresh Halannavar
7058527
Saarland University
Saarbrücken, Germany
poko00001@uni-saarland.de

Abstract

This project explores the feasibility of classifying emotions elicited by films using low-level properties of audio-visual content rather than traditional EEG signals. We extracted low-level audio and video features using datasets such as DREAMER and SEED-FRA. Machine learning models like XGBoost were employed to predict Arousal and Valence, achieving notable accuracy, especially in the Leave-One-Subject-Out (LOSO) validation. The results indicate that low-level features can effectively describe emotional states, with implications for emotion recognition technologies. However, challenges such as limited dataset size and the absence of direct EEG comparisons are acknowledged. Future work will focus on incorporating EEG data and expanding the dataset to enhance model performance and generalizability.

Keywords

Emotion Recognition, Low-level properties, Optic Flow, Histogram Analysis, Onset Strength, Event Rate, Spectral Contrast, Spectral Flatness, Zero Crossing Rate, Tempo, Valence, Arousal

1 Introduction

The ability to understand human emotions through technological means holds significant potential across various domains, including healthcare, entertainment, and education. Traditional methods for emotion recognition frequently rely on Electroencephalography (EEG) signals, which measure brain activity and are often presumed to correlate with emotional experiences. However, it remains unclear whether EEG signals genuinely reflect emotional states or if they primarily respond to low-level stimulus properties, such as brightness, color, or tempo.

This research seeks to explore the potential for classifying emotions elicited by film clips using low-level audio-visual features, rather than EEG data. By focusing on these features, the project aims to evaluate their effectiveness in predicting emotional states, specifically in terms of Arousal and Valence.

The motivation for this research lies in the prospect of enhancing emotion recognition technologies by leveraging more accessible and non-invasive data sources. Should this approach prove successful, it could pave the way for advancements in multiple fields where the interpretation of human emotions is of critical importance.

In this report, we detail the methodology employed, which involves the extraction of low-level features from film clips, the application of machine learning models, and the evaluation of these models using established datasets such as DREAMER [2] and SEED-FRA [4]. Additionally, we address the challenges inherent to this approach, including data variability and the complex nature of emotions, and propose potential avenues for future research to enhance the robustness and accuracy of emotion recognition systems.

2 Related Work

Previous studies have extensively explored the intersection of EEG signals and emotional recognition, particularly within audio-visual contexts. Early research by Koelstra et al. [3] established the DEAP dataset, linking EEG data with emotional responses to music videos, demonstrating the potential of EEG for emotion detection. However, the influence of low-level audiovisual features on EEG-based emotion prediction remains a critical concern. Yang et al. [5] showed that low-level acoustic features, such as tempo, can significantly modulate EEG patterns, raising questions about the specificity of emotion-related EEG markers. Similarly, Guo et al. [1] examined how optical flow in videos influences EEG signals, suggesting that what may appear as emotion-driven EEG responses might partially reflect underlying audiovisual properties. These studies underscore the necessity for a nuanced understanding of EEG's role in emotion recognition, especially in distinguishing genuine emotional responses from confounding audiovisual elements.

3 Methodology

3.1 Datasets

The training dataset used in this project was the DREAMER dataset, a multi-modal database containing EEG and ECG data from 23 participants. In order to elicit emotional responses, 18 film clips were used that were rated by participants in terms of affective

Seminar: "Machine Learning for Emotion Recognition", Apr. - Aug., 2024, Saarbrücken, DE

responses valence, arousal and dominance [2]. For the classification tasks, we extracted arousal and valence ratings, which were rated on 5 point scale.

To meet the binary classification requirement, a threshold of 3 was introduced to these ratings. Ratings greater than 3 were labeled as *Class* 1, while ratings 3 or below were labeled as *Class* 0. The final dataset was constructed by stacking the low-level feature data for each participant and pairing it with the corresponding binary labels, preserving the integrity of the physiological recordings and enabling standardized training.

The generalizability of the model was evaluated using the SEED-FRA dataset [4], which includes binary labels for valence but does not account for arousal. This dataset, consisting of 14 video clips, was used to assess the model's performance in classifying valence.

3.2 Low-Level Features

- **Histograms of Pixel Intensity** Video clip brightness is analyzed using histograms. Average skewness and kurtosis are calculated for each clip. Skewness indicates brightness distribution asymmetry, while kurtosis shows its peakedness. These metrics reveal brightness patterns and variations across clips.
- **Optic Flow**: Motion characteristics are quantified using three key metrics:
  - **Peak Magnitude:** Highest optic flow vector magnitude, showing maximum motion intensity.
  - **Average Magnitude:** Mean of optic flow vector magnitudes, indicating overall motion intensity.
  - **Standard Deviation of Magnitude:** Variability in optic flow magnitudes, reflecting motion consistency.
- **Event Rate**: This feature was extracted from the audio signals, which quantifies the frequency of distinct sound occurrences such as notes, spoken words, or transient noises. This metric, measured as the number of events per unit time, provides insights into the density and distribution of sonic activity throughout the audio content. It is calculated as the average of the total number of onsets detected per second for a given clip.
- **Onset Strength**: The abruptness and magnitude of energy change were measured at the beginning of sound events using an onset strength algorithm. For each clip, we calculated the Onset Strength as the average of the per-second average onset strengths, quantifying the typical intensity of sound beginnings.
- **Tempo**: The tempo, representing the speed of the audio, was extracted using an algorithm that calculates the number of beats per minute (BPM). This measurement indicates how quickly or slowly the rhythm progresses throughout the clip.
- **Spectral Contrast**: Spectral Contrast, which refers to the difference in amplitude between the peaks and valleys in an audio signal's frequency spectrum, was extracted using a spectral contrast detection algorithm. The Spectral Contrast Average for a clip is calculated as the average of the spectral contrast values across all frames.
- **Spectral Flatness**: Spectral Flatness quantifies how smooth the sound spectrum is. This feature was extracted using a spectral flatness detection algorithm. The Spectral Flatness Average for a clip is computed by averaging the spectral flatness values across all frames.
- **Zero crossing rate**: Zero-Crossing Rate (ZCR) measures the frequency at which an audio signal changes sign (crosses the zero amplitude line). The Zero-Crossing Rate Average for a clip is calculated by averaging the zero-crossing rates across all frames.

3.3 Pre-processing

Prior to model training, the film clips from the DREAMER and SEED-FRA datasets underwent several pre-processing steps:

- **Gray-scale Conversion:** All video clips were converted to gray-scale to reduce computational demands and emphasize key visual features most relevant to emotion recognition.
- **Dynamic Border Cropping:** For the SEED-FRA dataset, dynamic border cropping was an essential step, given that the videos were screen-recorded. This process removed unnecessary portions of the video frames (such as black bars) to improve the accuracy of low-level video feature extraction.

3.4 ML Models

We tested several machine learning models for emotion recognition, selecting each based on its suitability for classification tasks and compatibility with our dataset. Below are the models and the reasons for their selection:

- **K-Nearest Neighbors (KNN):** KNN is a simple algorithm well-suited for classification tasks, especially with complex decision boundaries. It classifies instances based on the labels of their closest neighbors in the feature space. KNN was chosen for its ease of use and interpretability, along with its ability to adapt to the data distribution without needing a predefined decision boundary.
- **Support Vector Machine (SVM):** SVM is a powerful algorithm often used for binary classification tasks. It finds the optimal hyperplane that maximizes the margin between classes. SVM was selected for its capacity to handle non-linear decision boundaries using the kernel trick, making it suitable for emotion recognition where feature-emotion relationships are complex.
- **Decision Tree:** Decision Trees are easy to interpret, as they split data based on feature values into branches leading to decisions. This model was chosen for its interpretability and its ability to capture non-linear relationships between features and emotions. Although Decision Trees can overfit small datasets, this issue can be mitigated with tuning or ensemble methods.
- **XGBoost:** XGBoost, an advanced gradient boosting technique, is known for speed, accuracy, and handling large datasets. It builds an ensemble of weak learners iteratively, improving predictions by correcting errors of previous models. XGBoost was chosen for its high performance in classification tasks and its ability to handle non-linearities while reducing overfitting through regularization. In our experiments, XGBoost delivered the highest accuracy in classifying emotional responses.

Each of these models was trained on binary classes derived from the DREAMER dataset, classifying emotional responses into high or low categories based on Valence and Arousal values.

3.5 Evaluation Methods

To assess the performance and generalization capability of our emotion recognition models, we employed two cross-validation techniques: Leave-One-Subject-Out (LOSO) and Leave-One-Trial-Out (LOTO). These methods were crucial in ensuring that our models could effectively generalize across different subjects and trials.

- **Leave-One-Subject-Out (LOSO) Cross-Validation:** In LOSO cross-validation, we aimed to evaluate the model's ability to generalize to new subjects. The dataset was divided such that the data from all but one subject were used for training, while the data from the left-out subject were used as the test set. This process was repeated for each subject in the dataset, ensuring that every subject's data was eventually used as a test set.
  Within each iteration:
  - Training Set: The data from all but one subject were used to train the model.
  - Validation Set: A portion of the training set was held out to fine-tune the model's hyperparameters and avoid overfitting.
  - Test Set: The data from the left-out subject served as the test set to evaluate the model's performance on unseen subjects.

This approach helped us assess how well the model could generalize to new subjects, providing a robust measure of its ability to handle subject variability.

- **Leave-One-Trial-Out (LOTO) Cross-Validation:** In LOTO cross-validation, the focus was on the model's ability to generalize across different trials within the same subject. For each subject, all but one trial were used for training, while the left-out trial was used for testing. This process was repeated for each trial, ensuring that every trial was used as a test set once.
  Within each iteration:
  - Training Set: The data from all but one trial of each subject were used for training.
  - Validation Set: A subset of the training data was used to validate and tune the model.
  - Test Set: The left-out trial served as the test set to evaluate the model's performance on unseen trials from the same subject.

This method provided insights into the model's consistency in recognizing emotions across different emotional states and conditions experienced by the same subject.

3.6 Evaluation Metrics

The primary metric for evaluating model performance was **accuracy**. Accuracy measures the proportion of correctly classified instances out of the total instances. In our binary classification task, it was essential to accurately categorize emotional responses (high or low) based on Valence and Arousal. This straightforward metric was particularly effective for assessing how well the models performed in the cross-validation setups described above.

3.7 Baseline Calculation

To assess our model's performance, we established baseline accuracies using the majority class method for both Leave-One-Trial-Out (LOTO) and Leave-One-Subject-Out (LOSO) cross-validation scenarios. This method sets the baseline accuracy as the percentage of the most frequent class.

For the DREAMER dataset, *Class* 0 was the majority class in both LOTO and LOSO settings, indicating a slight class imbalance. In contrast, the SEED-FRA dataset exhibits balanced classes for Valence, resulting in a baseline accuracy of 50%.

These baselines provide context for evaluating our model's performance, with accuracies significantly above these values indicating meaningful learning. The difference in class balance between datasets highlights the importance of considering dataset characteristics in model evaluation and may necessitate additional performance metrics for a comprehensive assessment.

4 Experimental Results and Analysis

4.1 Main Results

As evident from the results on the DREAMER dataset in Table 1, our model demonstrates substantially higher performance compared to the baseline accuracy for the Leave-One-Subject-Out (LOSO) evaluation setting. However, for Arousal prediction in the Leave-One-Trial-Out (LOTO) evaluation setting, the model's performance is marginally below the respective baseline. This discrepancy can be attributed to several factors in the LOTO scenario:

- **Limited generalization**: In LOTO, the model may struggle to generalize across different trials, possibly due to inter-trial variability in Arousal levels.
- **Temporal dependencies**: LOTO evaluations might not fully capture the temporal dynamics of Arousal, which could be crucial for accurate prediction.

Table 1: Mean Accuracy results for XGBoost model

Evaluation Type	Arousal	Valence
LOTO	55.56%	76.57%
LOSO	74.40%	81.16%
Baseline (Majority Class)	56.00%	61.00%

4.2 Generalizability Results

In evaluating our model's generalizability on the SEED-FRA dataset, we observed contrasting results across different cross-validation strategies. Using the best parameters from the Leave-One-Trial-Out (LOTO) evaluation, the model achieved 71.43% accuracy, indicating effective learning of trial-specific patterns. However, with the best parameters from the Leave-One-Subject-Out (LOSO) evaluation, the precision was 50%, which matches the baseline performance discussed in Section 3.7. This discrepancy highlights the challenges

in developing emotion recognition models that generalize well across both trials and subjects, particularly for complex datasets like SEED-FRA.

4.3 Correlation Analysis Results

Given that the ground truth arousal values in both the DREAMER and SEED-FRA datasets were continuous[1], we conducted a correlation analysis between the audio-visual features and arousal levels (Table 2). A notable observation emerged from this analysis: for both these datasets, we identified a consistent set of features that positively correlated with arousal values. This finding has several implications:

- **Cross-dataset consistency**: The uniformity in feature correlations across two distinct datasets suggests a robust relationship between these audio-visual characteristics and perceived arousal.
- **Potential for generalizable models**: This consistency indicates that these features may serve as reliable indicators of arousal across different experimental contexts and participant groups.
- **Insight into arousal mechanisms**: The identified features may provide valuable insights into the audio-visual cues that humans associate with elevated arousal states.

Table 2: Correlation analysis for Arousal (V=video, A=audio)

Positively Correlated	Negatively Correlated
Optic Flow Peak Magnitude (V)	Event Rate (A)
Optic Flow Average Magnitude (V)	Spectral Flatness (A)
Optic Flow Standard Deviation (V)	Zero Crossing Rate (A)
Histogram Skewness (V)	
Histogram Kurtosis (V)	
Onsets (A)	
Bpm (A)	
Spectral Contrast (A)	

4.4 KS-Test Results

In our experiments, we tested our model on a novel dataset originating from a distribution significantly different from that of our training data, which primarily includes data from the 2000s (DREAMER dataset), while the test data comprises samples from the 1980s and 1990s (SEED-FRA dataset). To determine whether the features in both datasets were derived from the same distribution, we employed the Kolmogorov–Smirnov test, a non-parametric statistical test commonly used to compare the distributions of two samples. This was a critical step, as ensuring similar feature distributions between the training and test datasets would enhance the model's generalizability to new, unseen data. The results indicated that seven features shared the same distribution across both datasets, while four features exhibited significant distributional differences in the test data compared to the training data. Consequently, we excluded the four features — 'peak magnitude','average magnitude','standard deviation of magnitude', and 'spectral flatness' — from further analysis, as their p-values were less than 0.002, indicating a lack of generalizability.

Following the exclusion of these features, we repeated the experiment using the XGBoost model to evaluate the performance on the DREAMER dataset after excluding the 4 features (Table 3), ensuring improved robustness and generalization to unseen data. When testing the best model parameters on the SEED-FRA dataset for Valence, we achieved 64.29% accuracy for LOTO and 57.14% accuracy for LOSO. Comparing these results with the testing results on the full feature set in Section 4.2, we observe that the LOTO model accuracy has dropped by 7%, while the LOSO model has jumped by 7%. While the performance dropped for LOTO, it still ensures generalizability, which can be interpreted as a trade-off between generalizability and accuracy.

Table 3: Mean Accuracy results after performing KS-Test

Evaluation Type	Arousal	Valence
LOTO	55.30%	67.15%
LOSO	74.39%	81.16%
Baseline (Majority Class)	56.00%	61.00%

5 Conclusions

When applying Leave-One-Subject-Out (LOSO) validation, our model achieved significant improvements over the baseline for both Arousal and Valence, with performance metrics of 81.16% for Valence and 74.40% for Arousal. In contrast, the Leave-One-Trial-Out (LOTO) validation yielded good results for Valence (76.57%) but under-performed for Arousal (55.56%). When assessing the generalizability of our model using the SEED-FRA dataset as a test set, the results were positive and surpassed the baseline, demonstrating the robustness of our approach. Additionally, while utilizing the Kolmogorov-Smirnov (KS) Test as a feature reduction method had a minimal negative impact on the results, it ensured that generalizable features were retained, making it a viable option, particularly for handling large datasets.

6 Limitations

The absence of EEG prediction results prevented us from comparing predictions derived from low-level features with those obtained from EEG data. Additionally, the limited size of the training data, with only 18 clips available in the DREAMER dataset, impacted our results in some scenarios, especially in the Leave-One-Trial-Out (LOTO) validation for Arousal.

7 Future Work

In future, obtaining EEG predictions and comparing them with our model will be crucial to determine whether EEG truly captures emotional states or merely reflects reactions to low-level properties. Additionally, expanding the dataset by including more video clips will be beneficial, as it will provide a larger number of data points for each emotion, potentially leading to improved results and more robust conclusions.

[1]Valence values are not continuous for SEED-FRA dataset

References

[1] Wenshan Guo, Qi Jiang, Yueming Li, and Xia Li. 2020. EEG-based emotion recognition considering video-induced low-level audiovisual features. *IEEE Access* 8 (2020), 183789–183801.

[2] Stamos Katsigiannis and Naeem Ramzan. 2017. *DREAMER: A Database for Emotion Recognition through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices.* https://doi.org/10.1109/JBHI.2017.2688239

[3] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ali Yazdani, Touradj Ebrahimi, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31.

[4] Wei Liu, Wei-Long Zheng, Ziyi Li, Si-Yuan Wu, Lu Gan, and Bao-Liang Lu. 2022. Identifying similarities and differences in emotion recognition with EEG and eye movements among Chinese, German, and French People. *Journal of Neural Engineering* 19, 2 (2022), 026012.

[5] Yi-Hsuan Yang, Yen-Ping Lin, Chun-Yu Wu, Chia-Gan Tsai, Jyh-Horng Chen, and Tzyy-Ping Jung. 2019. Music tempo-induced emotions and EEG responses. *Journal of Neural Engineering* 16, 5 (2019), 056026.