Feature-Based Emotion Classification Using Eye-Tracking Data

Tomi Božak
tb85088@student.uni-lj.si
Jožef Stefan Institute
Ljubljana, Slovenia

Mitja Luštrek
mitja.lutsrek@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Gašper Slapničar
gasper.slapnicar@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

The field of emotion recognition from eye-tracking data is well-established and offers near-real-time insights into human affective states. It is less obtrusive than some other modalities, such as electroencephalogram (EEG), electrocardiogram (ECG) and galvanic skin response (GSR), which are often used in emotion recognition tasks. This study examined the practical feasibility of emotion recognition using an eye-tracker with a lower frequency than that typically employed in similar research. Using ocular features, we explored the efficacy of classical machine learning (ML) models in classifying four emotions (anger, disgust, sadness, and tenderness) as well as neutral and "undefined" emotions. The features included gaze direction, pupil size, saccadic movements, fixations, and blink data. The data from the "emotional State Estimation based on Eye-tracking database" was preprocessed and segmented into various time windows, with 22 features extracted for model training. Feature importance analysis revealed that pupil size and fixation duration were most important for emotion classification. The efficacy of different window lengths (1 to 10 seconds) was evaluated using Leave-One-Subject-Out (LOSO) and 10-fold cross-validation (CV). The results demonstrated that accuracies of up to 0.76 could be achieved with 10-fold CV when differentiating between positive, negative, and neutral emotions. The analysis of model performance across different window lengths revealed that longer time windows generally resulted in improved model performance. When the data was split using a marginally personalised 10-fold CV within video, the Random Forest Classifier (RF) achieved an accuracy of 0.60 in differentiating between the six aforementioned emotions. Some challenges remain, particularly in regard to data granularity, model generalization across subjects and the impact of downsampling on feature dynamics.

Keywords

eye-tracking, emotion recognition, machine learning

1 Introduction

Emotion recognition is a vibrant area of research, leveraging diverse data sources such as images [11], audio [16], and also, ocular features like pupil dilation, gaze direction, blinks, and saccadic movements [3, 8, 12]. Such eye-related features provide valuable insights into emotional states, offering a less-invasive and real-time approach to understanding human affective responses. Most studies that tried to predict emotions from these eye-related features relied not only on eye-tracking data but also on EEG [8, 12]. We hypothesized that eye-tracking data is a valuable modality for multi-modal emotion recognition on its own, with potential applications in real-world scenarios like office work, driving, and psychological assessments, as well as in estimating well-being. Our motivation was to explore eye-tracker-based predictive models as an essential component in such practical applications.

The primary objective of our study was to validate existing findings on the performance of classical ML models for emotion classification from eye-tracking data, using the models – Support Vector Machine (SVM) and k-Nearest Neighbors (KNN) – and features already explored in the literature [9, 15] as well as exploring classifiers not so frequently used in this field – such as RF and XGBoost (XGB). Additionally, we aimed to explore the potential of emotion recognition at lower sampling frequencies available in most non-professional eye trackers. For the early feasibility study, we used an existing dataset, which collected data using a wearable eye-tracker but findings could possibly be extended to high-quality unobtrusive contact-free trackers. Our research also focused on understanding the impact of individual features and window lengths on model performance.

2 Related Work

In literature, various physiological signals have been employed for emotion recognition, with a particular focus on modalities such as EEG, GSR, and eye-tracking systems [1, 6, 9]. Researchers have explored both uni- and multi-modal approaches, finding that the integration of multiple modalities can significantly enhance emotion recognition accuracy. Lu et al. achieved 0.78 accuracy with eye-related features recorded with eye-tracking glasses – which are not contact-free but record at relatively low frequencies of 60 Hz or 120 Hz. They predicted positive, negative and neutral classes with SVM. Interestingly, they observed a 0.10 increase in accuracy when combining eye-related and EEG features [12]. Similarly, Guo et al. observed a more substantial gain, with accuracy improving by 0.20 when integrating EEG, eye-tracking, and eye images, as opposed to using only eye-tracking data [7].

The features derived from eye-tracking have been widely used in ML algorithms to detect emotional states [2, 7, 12, 15]. However, most studies have traditionally categorized emotions into broad groups like positive, negative, and neutral [12, 14]. Pupil size, in particular, has emerged as a valuable indicator for distinguishing between positive and negative emotions [2, 7, 12] . Recent efforts have begun to refine these broad categories, identifying more specific emotions like happiness, sadness, fear, anger, etc. [2, 7, 15]. Although current methods can effectively identify certain emotions such as sadness and fear, further research is needed to reliably differentiate between others like disgust, joy, and surprise [2].

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.9988

3 Methodology

3.1 Data

In our research, we used the "emotional State Estimation based on Eye-tracking database" (eSEEd) [13]. The eSEEd comprises data from 48 participants, each of whom watched 10 carefully selected videos intended to evoke specific emotional responses. After viewing each video, participants ranked their emotions – anger, disgust, sadness, and tenderness – on a scale from 0 to 10. Tenderness, however, is not regarded as one of the basic emotions, but it has been widely utilized in emotion research in recent years [13]. Since the participants had ranked all four emotions for every video, a labelling problem emerged when multiple emotions shared the highest score, in our case, leading to "undefined" labels. In our study, emotions were mapped by applying a set of extraction rules in the following order: if the highest-ranked emotion is below four, the response is labelled as neutral; if multiple emotions share the highest rank, the label is undefined; otherwise, the emotion with the highest rank is chosen. The boundary of four was chosen because the original study on eSEEd constructed this rule and we adapted it from there [13]. Although the initial study design aimed for an even distribution of emotions, neutral responses dominate, representing about one-fourth of the labels (depending on window length).

3.1.1 Data Preprocessing. We have preprocessed the data to make it more suitable for our future research and to reduce its size. We wanted to study the performance of data with a relatively low frequency rate of 60 Hz, which is used by relatively affordable mid-tier eye-trackers, like Tobii Pro Spark. Firstly, the features that were uninformative or could be misleading (e.g. raw tracker signal and timestamps) were removed, and the following set of features was preserved: 2D screen coordinates of gaze points (for standard deviation (std) of screen gaze coordinates), 3D coordinates of gaze points (exclusively for saccade calculations), pupil sizes (a and b of the pupil ellipse), and eye IDs (each eye has its own pupil size features). Secondly, rows containing any NaN values were removed, as there were no large consecutive blocks of such rows and downsampling of the data was planned. Finally, we further downsampled the data to 60 Hz, matching the sampling frequency of a mid-tier eye-tracker. However, we acknowledge that downsampling might lead to the loss of high-frequency information, which could be important for capturing subtle dynamics in gaze behaviour and pupil responses. This is particularly relevant considering that recent studies, such as those by Collins et al. [3] and the SEED project [4, 17], have utilized data collected at much higher frequencies to preserve these subtle dynamics. Therefore, while downsampling makes the data more meaningful to our research and more computationally manageable, it is important to keep in mind the reduced temporal resolution when discussing the results.

Following the preprocessing, window segmentation was applied to the data. This step is essential for analyzing temporal patterns within the data, as it allows for the capture of trends and behaviours over specific time intervals. By segmenting the data into windows, we can improve the robustness of feature extraction and model training, enabling the detection of meaningful patterns that might be obscured in raw, unsegmented data. Additionally, with window segmentation, the number of training instances increases which is commonly better for learning more robust ML models and conducting rigorous evaluation. Hence, multiple window lengths were examined, namely: 1, 3, 5 and 10 s. We used 50% sliding window overlap. From each window, we computed 22 features, belonging to the following groups:

(1) gaze coordinates on screen: std of x and y coordinates
(2) pupil ellipse sizes of a and b for each eye: mean, std
(3) blinks: number; mean and std of duration (all 0 if no blinks)
(4) saccades: number; mean speed; mean, std, total duration
(5) fixations: number; mean, std, total duration

Saccade and, implicitly, fixation calculations were done using existing code based on the algorithm proposed by Engbert et al. [5, 10]. The algorithm calculates the velocity and acceleration of eye movements by using a velocity threshold identification method to detect saccades based on continuous 3D gaze data. In our study we define fixation (interval) as an absence of a saccade (interval), thus one fixation is declared between every two saccades (and before the first and after the last one).

As mentioned previously, our data was imbalanced in terms of class distribution, namely the distributions for anger, disgust, sadness, neutral, tenderness and undefined were 8.7%, 13.6%, 17.5%, 25.7%, 15.8% and 18.7%, respectively. Notably, for the 1 s window length, the number of windows was 67,181, whilst for the 10 s window, the number of instances decreased to 6,507.

3.2 Experiments

We initially examined feature correlation matrices to identify potential correlations between features, as well as between feature and class. Then, we compared the following classifiers from the Scikit-learn library: Random Forest (RF), Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and XGBoost (XGB) from the XGBoost library, as well as an ensemble method majority vote of the aforementioned classifiers. We compared all results against a baseline majority classifier. Each model was trained and tested using its default hyperparameters. To evaluate the models' performance, we implemented multiple CV techniques.

The first CV technique was Leave-One-Subject-Out (LOSO). Secondly, we implemented a marginally personalised 10-fold CV "within video." In this approach, a standard 10-fold CV was performed where 90% of temporally sequential windows were used for training and 10% for testing. The splits were done separately for each video within every subject. All the training data from every video was combined to train a single model, and all the test data was combined to evaluate the model, ensuring that the model was exposed to data from all subjects and videos. We named the experiment "marginally" personalised because most training data does not come from any single subject and is thus not very personalised. Finally, we explored a completely personalised 10-fold CV "within subject." Here, training and testing were done only on data of one subject. In all three CV methods, the instances were never shuffled to preserve temporal and subject sequential information and to minimize overfitting.

We attempted to merge certain classes in a way to group negative emotions – anger, disgust, and sadness – under the category "negative," while labelling tenderness as "positive." The label for neutral remained unchanged, while the undefined label was changed to "negative" because it always resulted from multiple negative emotions scoring equally. Lastly, the feature importances were analysed for different combinations of data splits and models in order to identify potential consistently important features.

4 Results

The results described in the following subsections are summarised in Table 1.

4.1 Feature Correlations

The first important observation from correlation matrices was that no output class is closely correlated to any other singular feature. Secondly, we noticed some strong correlations, for example, a 1.0 correlation between a number of fixations and a number of saccades, because one simply equals the other increased by one. More importantly, we noticed little-to-no correlation between features that proved to be most important in some best-performing models, meaning each of these features brought some novel information to the model. The only exceptions of important features being correlated are the features representing the mean size of a pupil i.e., ellipse a and b axes, which are expected to be correlated. They were correlated more than 0.8. However, we decided not to remove any features because we assessed the feature count of 22 to be well-balanced in relation to the number of instances.

4.2 Leave-One-Subject-Out

With the goal of training a robust general model for our dataset, we first applied the LOSO CV technique. The best performance was achieved by RF on 10 s windows, yielding an accuracy of 0.28 ± 0.13 and an F1-score of 0.28 ± 0.16. It outperformed the majority classifier by 0.03 in accuracy and 0.13 in F1-score. In a subsequent experiment, the negative emotions were grouped. This adjustment led to an overall increase in performance. However, with such grouping the majority classifier score also increased to 0.59 accuracy, which is the same as the best-performing model.

Further analysis revealed that high accuracy mainly implied the subject predominantly reported "neutral" feelings and low accuracy implied little-to-no "neutral" labels. However, not every subject with a high "neutral" count achieved outstanding results and not every subject with a wide range of emotions yielded poor results. A comparison was made between the number of windows in the left-out subject to their performance and no correlation was found. 10 s window length performed better than the shorter windows with lengths 1-5 s. We also tested longer (60 s) windows and the resulting accuracies were higher than those from 10 s windows, but we evaluated that the number of instances was insufficient for the results to be representative.

4.3 Marginally Personalised 10-fold Cross-Validation Within Video

Given that the LOSO yielded relatively poor results, the next step was to explore 10-fold CV. Experiments showed an average accuracy of 0.60 ± 0.07 and an F1-score of 0.60 ± 0.08, produced with RF on 10 s windows, the best-performing model. This should be compared to the results given by the majority classifier – average accuracy of 0.21 ± 0.01 and F1-score of 0.07 ± 0.01. With negative emotions grouped, the accuracy and F1-score raised to 0.76 ± 0.04 and 0.73 ± 0.04, respectively, for the best-performing XGB on 10 s windows. The majority class classifier yielded an accuracy of 0.66 ± 0.02 and an F1-score of 0.52 ± 0.02.

4.4 Personalised 10-fold Cross-Validation

Even though 10-fold CV within video resulted in much better performance compared to LOSO, we wanted to see the performance of completely personalised models. All the models performed similarly well, with the absolute best being RF on 10 s windows which outperformed the majority classifier by 0.05 and 0.13 for accuracy and F1-score, respectively. When grouping the negative emotions, we observe an absolute improvement in models' performance, but a relative decline toward the majority classifier benchmark. The best model, in this case, did not surpass the majority classifier in terms of accuracy, with the majority classifier achieving 0.67 ± 0.16 accuracy and 0.61 ± 0.16 F1-score, while SVM, the best-performing model, scored an accuracy of 0.64 ± 0.13 and an F1-score of 0.63 ± 0.12.

4.5 Feature Importances

Following the completion of model training, we analyzed the feature importances of the best-performing models. For RF this was calculated based on the Mean Decrease in Impurity, summing the impurity reduction each feature contributes across all trees; and for XGB, feature importances were calculated using the "weight" metric, which counts the number of times each feature is used to split the data across all trees. For SVM we did not calculate feature importances. In the completely personalised 10-fold experiments, feature importances varied significantly across different subjects and even between different runs within the same subject, specifically with RF, as the random state was not fixed. In contrast, feature importance was notably consistent in experiments where models were trained on data from multiple subjects, such as in the LOSO and the 10-fold within video, even with a variable random state of the RF model.

The most important features of best-performing models were those related to average pupil sizes, followed by fixation duration. These results partially align with those of Collins et al., who found features relating to pupil diameter and saccades statistically significant [3].

5 Conclusion

Our research explored emotion classification using eye-tracking data with classical ML models and hand-crafted features. The data was downsampled to a lower-than-standard frequency i.e., to 60 Hz, which was more realistic for consumer contact-free eye-tracker data. This made the problem harder, making it not directly comparable with other studies working on eSEEd, but valuable from a practical perspective.

Window segmentation significantly impacted model performance, with the best results constantly obtained using the largest window length. This suggests that longer observation periods capture more comprehensive information, making smaller windows less effective for emotion classification. We hypothesize that this does not transfer to realistic scenarios, as users might experience emotions in short bursts while being neutral for the majority of the time. In specifically designed cases where emotion is consistently induced for longer periods of time (like our dataset), this is more expected.

The LOSO validation strategy, which tests model generalization across different subjects, yielded poor results. The variability in performance across subjects indicates the challenge of capturing general relationships between eye features and emotions. While both 10-fold CV approaches showed an increase in performance, their generalizability is limited. Completely personalised 10-fold showed worse results than the marginally personalised one presumably because of the low number of videos per emotion within an individual subject.

Table 1: Best-performing models and their corresponding results along the results of the Majority Class Classifier for the same parameters. Window lengths are 10 s.

Settings	**Model Acc**	Model F1	Majority Class Acc	Majority Class F1
LOSO, RF	**0.28 ± 0.13**	0.28 ± 0.16	0.25 ± 0.25	0.15 ± 0.26
LOSO, SVM, negative emotions grouped	**0.59 ± 0.19**	0.46 ± 0.18	0.59 ± 0.19	0.46 ± 0.18
10-fold within video, RF	**0.60 ± 0.07**	0.60 ± 0.08	0.21 ± 0.01	0.07 ± 0.01
10-fold within video, XGB, negative emotions grouped	**0.76 ± 0.04**	0.73 ± 0.04	0.66 ± 0.02	0.52 ± 0.02
10-fold within subject, RF	**0.38 ± 0.20**	0.42 ± 0.19	0.33 ± 0.26	0.29 ± 0.26
10-fold within subject, SVM, negative emotions grouped	**0.64 ± 0.13**	0.63 ± 0.12	0.67 ± 0.16	0.61 ± 0.16

An important issue with the eSEEd data is that all participants watched the same 10 emotion-evoking videos in the exact same order. This uniformity raises concerns that, given the small number of videos (two intended[1] per emotion), the models might learn to associate features unrelated to emotions, such as video dynamics or illumination. We circumvented the problem with video dynamics by dropping the mean gaze coordinate features and not using them in our experiments.

Despite these challenges, our experiments offer valuable insights into the feasibility of emotion recognition from low-frequency eye-tracker data, providing a foundation for future work. We opted for classical models initially due to their explainability, lower computational complexity, and efficiency, which are in our opinion essential for understanding the data before transitioning to more complex deep learning models.

In future work, several enhancements could be explored to improve the robustness and accuracy of emotion classification models using eye-tracking data. One approach could involve analyzing distinct fixation areas as an additional feature, potentially offering deeper insights into visual attention patterns. Moreover, considering that each emotion is (in some cases) represented by two videos, a valuable experiment would be to train models on one video and test on the other. This could help assess the model's ability to generalize across different stimuli within the same emotional category.

Further analysis could focus on demographic factors by examining the LOSO results for potential correlations between model predictions and participant characteristics such as gender, age, and education. This might reveal underlying biases or trends that affect model performance. Additionally, rather than downsampling and removing rows with missing data, future work could explore retaining or imputing these rows.

Furthermore, exploring the training of neural networks on raw, non-downsampled data from multiple modalities is another promising direction, as other studies already observed promising results with such approaches. Moreover, we should address the issue of overlapping emotions which could involve developing a multiclass output model, reflecting a real-world scenario where multiple emotions can be present simultaneously. This approach could also help reduce the number of undefined labels, increasing the amount of useful data.

Acknowledgements

This work was supported by bilateral Weave project, funded by the Slovenian Agency of Research and Innovation (ARIS) under grant agreement N1-0319 and by the Swiss National Science Foundation (SNSF) under grant agreement 214991.

[1]The average percentage of the videos for which the participants had reported the target emotion (also known as the "hit rate") was 71.8% [13].

The authors acknowledge the use of OpenAI's ChatGPT for generating text suggestions during the preparation of this paper. All the generated content has been reviewed and edited by the authors to ensure accuracy and relevance to the research.

References

[1] Zeeshan Ahmad and Naimul Khan. 2022. A survey on physiological signal-based emotion recognition. *Bioengineering 2022*. https://www.mdpi.com/2306-5354/9/11/688.

[2] Aracena Claudio, Basterrech Sebastián, Snáel Václav, and Velásquez Juan. 2015. Neural networks for emotion recognition based on eye tracking data. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2632–2637. DOI: 10.1109/smc.2015.460.

[3] Mackenzie L. Collins and T. Claire Davies. 2023. Emotion differentiation through features of eye-tracking and pupil diameter for monitoring well-being. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. DOI: 10.1109/embc40787.2023.10340178.

[4] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. 2013. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 81–84.

[5] Ralf Engbert, Lars Rothkegel, Daniel Backhaus, and Hans A. Trukenbrod. 2016. Evaluation of velocity-based saccade detection in the smi-etg 2w system. *Technical report, Allgemeine und Biologische Psychologie, Uni-versität Potsdam*.

[6] Atefeh Goshvarpour, Ataollah Abbasi, and Ateke Goshvarpour. 2017. An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomed J. 2017*. DOI: 10.1016/j.bj.2017.11.001.

[7] Jiang-Jian Guo, Rong Zhou, Li-Ming Zhao, and Bao-Liang Lu. 2019. Multimodal emotion recognition from eye image, eye movement and EEG using deep neural networks. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3071–3074. DOI: 10.1109/embc.2019.8856563.

[8] Robert Jenke, Angelika Peer, and Martin Buss. 2014. Feature extraction and selection for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 5, 3, 327–339. DOI: 10.1109/taffc.2014.2339834.

[9] Lim Jia Zheng, Mountstephens James, and Jason Teo. 2020. Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors 2020*. https://doi.org/10.3390/s20082384.

[10] Fjorda Kazazi. 2022. Detect saccades and saccade mean velocity in python from data collected in pupil labs eye tracker. Accessed: 25. 7. 2024. https://www.fjordakazazi.com/detect_saccades.

[11] Yousif Khaireddin and Zhuofa Chen. 2021. Facial emotion recognition: state of the art performance on FER2013. *arXiv preprint arXiv:2105.03588*. DOI: 10.48550/arXiv.2105.03588.

[12] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. 2015. Combining eye movements and EEG to enhance emotion recognition. In *Ijcai*. Vol. 15. Buenos Aires, 1170–1176.

[13] Vasileios Skaramagkas and Emmanouil Ktistakis. 2023. Esee-d: emotional state estimation based on eye-tracking dataset. *Brain Sciences*, 13, 4. DOI: 10.3390/brainsci13040589.

[14] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3, 2, 211–223. DOI: 10.1109/t-affc.2011.37.

[15] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz Jan Rak. 2020. Eye-tracking analysis for emotion recognition. *Computational Intelligence and Neuroscience*. https://onlinelibrary.wiley.com/doi/10.1155/2020/2909267.

[16] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2018. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28, 10, 3030–3043. DOI: 10.1109/tcsvt.2017.2719043.

[17] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7, 3, 162–175. DOI: 10.1109/TAMD.2015.2431497.