A machine learning approach to predict emotional arousal and valence from gaze extracted features

Vasileios Skaramagkas*, Emmanouil Ktistakis*†, Dimitris Manousos*, Nikolaos S. Tachos‡,
Eleni Kazantzaki*, Evanthia E. Tripoliti§, Dimitrios I. Fotiadis‡§ and Manolis Tsiknakis*¶
*Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH),
GR-700 13 Heraklion, Crete, Greece, Email: vskaramag@ics.forth.gr
†Laboratory of Optics and Vision, School of Medicine, University of Crete,
GR-710 13 Heraklion, Crete, Greece
‡Dept. of Biomedical Research, Institute of Molecular Biology and Biotechnology (FORTH),
GR-451 10, Ioannina, Greece
§Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems,
University of Ioannina, GR-451 10, Ioannina, Greece
¶Dept. of Electrical and Computer Engineering, Hellenic Mediterranean University,
GR-710 04 Heraklion, Crete, Greece

***Abstract*—In the last years, many studies have been investigating emotional arousal and valence. Most of them have focused on the use of physiological signals such as EEG or EMG, cardiovascular measures or skin conductance. However, eye related features have proven to be very helpful and easy to use metrics, especially pupil size and blink activity. The aim of this study is to predict emotional arousal and valence levels which are induced during emotionally charged situations from eye related features. For this reason, we performed an experimental study where the participants watched emotion-eliciting videos and self-assessed their emotions, while their eye movements were being recorded. In this work, several classifiers such as KNN, SVM, Naive Bayes, Trees and Ensemble methods were trained and tested. Finally, emotional arousal and valence levels were predicted with 85 and 91% efficiency, respectively.**

I. Introduction

Among the various dimensional models of affect, the 2D arousal-valence emotion space of Russell [1] is the most commonly used one. Emotional valence describes the extent to which an emotion is positive or negative [2], whereas arousal refers to the level of calmness (i.e., low arousal) or excitation (i.e., high arousal) elicited by a stimulus [3].

Physiological signals combined with eye-related metrics are the most commonly used modality in order to estimate one's emotional state [4]. However, there are several studies that have used eye features as the only predictor to identify emotional arousal and valence levels. These studies attempt to solve either multi-class [5], [6], [7] or binary classification problems [8], [9] with success rates for multi-class cases remaining below 80%, while the binary classification approaches have proven to be more effective reaching up to 93% prediction success. Nevertheless, none of the aforementioned studies attempts to investigate in parallel the discrimination between the arousal and valence levels.

A meta-analysis of the related studies has shown that the gaze extracted features that better indicate emotional arousal are pupil diameter and blink duration [3].

This work reports the results of a study in which participants watched emotion-evoking video clips during which, an eye tracker was used to capture eye motion and activity. The features extracted from all acquired gaze signals were used to train and evaluate a set of different classification algorithms, including decision trees, discriminant analysis, support vector machine (SVM), k-nearest neighbors (kNN) and ensemble learning algorithms, aiming to accurately classify the various arousal and valence levels.

II. Protocol for data collection

In the present study, 37 participants (22 female, 15 male) with mean age 29 (SD:7) years were enrolled. Binocular visual acuity at 80 cm was measured before each trial (mean VA: -0.10±0.07 logMAR). Mean illuminance at cornea when screen was on, was 450 (SD: 24) lux.

Fig. 1. The experimental setup

Two video clips for each of the 4 emotions (happiness, sadness, anger and disgust) were obtained from the public database FilmStim [10]. Two more video clips served as neutral videos thus creating a total of 10 videos watched by each participant. The video clips were presented in a randomised order. After each video clip, participants were presented with a questionnaire for the self-assessment of the above mentioned emotions in a scale from 1-10. The design of the study is shown in Fig. 2. We only accepted self-assessments of 5 or higher as a true indication of the presence of a specific emotion. A self-assessment score lower than 5 was treated as emotionally neutral. In parallel, we estimated the level of arousal and valence for each of the emotions (Table I), based on [11].

The video clips were presented on a computer screen at 80cm distance from the study participant as shown in Fig. 1. All measurements were performed with the participants seated on a chair with their head stabilized by means of a chin and head rest to minimize head movements. Eye tracking measurements were recorded using the Pupil Labs "Pupil Core" eye and gaze tracker.

TABLE I
AROUSAL AND VALENCE LEVEL BASED ON EMOTION

Emotion	Arousal level	Valence level
happiness	medium	positive
anger	high	negative
disgust	high	negative
sadness	low	negative
neutral	medium	neutral

From the 10 emotion-evoking videos watched by each of the participants and after removing the invalid recordings where the participants looked away from the screen or closed their eyes in order to avoid a certain video scene, a total of 362 examples were collected. The classes of valence and arousal in which the data were split can be seen in Table II. The "positive" valence and "low" arousal classes constitute minority classes as they contain a few number of examples.

TABLE II
EMOTION CLASSIFICATION CATEGORIES

Emotional state	Classes	Sample size
Arousal	not high / high	(233 / 129)
Valence	not positive / positive	(326 / 36)
Arousal	low / medium / high	(32 / 201 / 129)
Valence	negative / neutral / positive	(159 / 167 / 36)

The study protocol was approved by the Ethics Committee of FORTH and all participants have signed written consent.

III. METHODOLOGY

A. Feature extraction

The feature extraction procedure is extremely significant in order to be able to efficiently discriminate among the various emotional arousal and valence levels solely from the low level eye and gaze metrics collected from the eye tracker. For each recording sequence, the raw gaze points from Pupil Core are processed and analyzed to ensure that the participants have watched the whole video each time and did not look away from the screen neither close their eyes for duration longer than the average blink time to avoid watching. If these prerequisites are satisfied, fixations and saccades are identified based on the I-VT algorithm proposed by Salvucci et al. 2000 [12] and fixation and saccade related features are calculated. Furthermore, pupil diameter and blink timings computed from the eye tracker contribute to further extract pupil and blink related features. In total, 29 eye and gaze features are extracted and are presented in Table III.

TABLE III
FEATURES EXTRACTED FROM EYE AND GAZE DATA

Fixations	Saccades	Blinks	Pupil
duration*	duration*	frequency	diameter*
total duration	velocity*	duration*	difference from
frequency	frequency		baseline metric
	amplitude*		

* max, min, mean and median values are calculated.

B. Data processing

The problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary. One way to solve this problem is to oversample the examples in the minority class. An improvement on duplicating examples from the minority class is to synthesizing new examples from the minority class. This is a type of data augmentation for tabular data and can be very effective. We employed the widely used approach to synthesizing new examples called the Synthetic Minority Oversampling TEchnique, or SMOTE [13].

In addition, due to the fact that many machine learning algorithms perform better when numerical input variables are scaled to a standard range [14] we performed data scaling using the MinMax Scaler since our data are not normally distributed. Furthermore, MinMax Scaler rescales the dataset in such a manner that all feature values are in the same range (0–1).

C. Feature selection

We tested and performed feature selection with two different approaches. During the first approach we used a regularization method and compared it with the ANOVA test and then we obtained the feature importance from ensemble methods. The second approach involved the LASSO regularization analysis, a regression analysis method that performs both variable selection and regularization, thus improving accuracy and interpretability [15].

D. Training and testing

For the classification procedure we split the data into training and testing, with the number of the test data amounting 20% of the total number of examples. The classifiers we chose are split into two groups; the ones used for binary and the ones

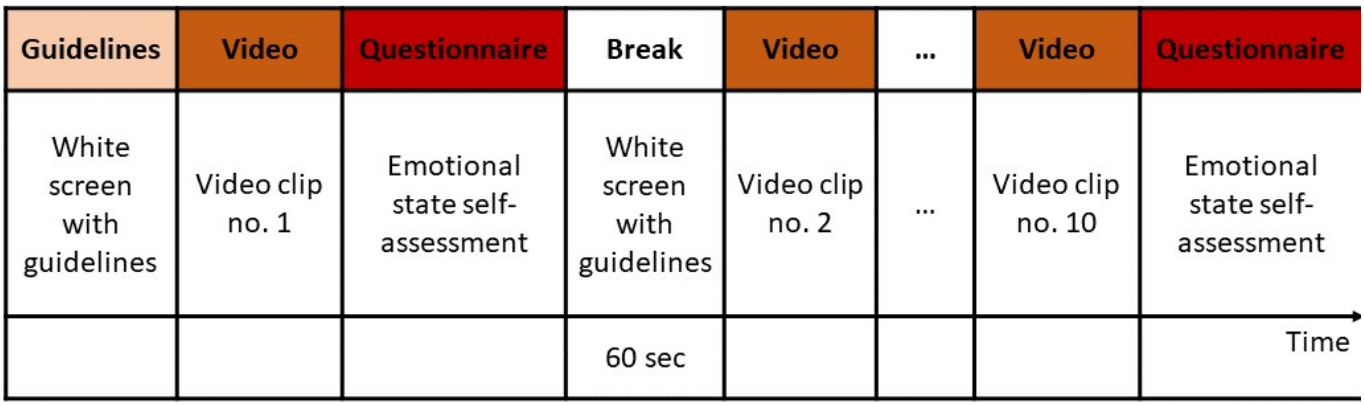

Fig. 2. Design of the study

used for multiclass classification. For the binary classification problem we train and test the following classifiers:

- neighbors.KNeighborsClassifier (n = 7, metric = 'euclidean', weights = 'uniform')
- svm.LinearSVC (kernels = 'linear')
- ensemble.GradientBoosting (n_estimators = 12)
- naive_bayes.GaussianNB
- ensemble.RandomForestClassifier (max_features = 10, n_estimators = 12)
- linear_model.Perceptron (alpha = 0.0001, max_iter = 500)

Then, for the multi-class process we train and test the classifiers enlisted below:

- naïve bayes.BernoulliNB
- tree.DecisionTreeClassifier (n = 23)
- tree.ExtraTreeClassifier (criterion = 'gini')
- ensemble.ExtraTreesClassifier (n_estimators = 100)
- ensemble.GradientBoosting (n_estimators = 10)
- naive_bayes.GaussianNB
- svm.LinearSVC (setting multi_class = 'crammer_singer', kernels = 'linear')
- linear_model.LogisticRegression (setting multi_class = 'multinomial', penalty = 'l1', C = 10)
- ensemble.RandomForestClassifier (max_features = 10, n_estimators = 10)

Furthermore, and in order to choose an algorithm that is able to learn from training data how to recognize the classes of the target variable by minimizing the error function, we need to tune each classifiers' hyperparameters properly [16]. There are many hyperparameters and there is no general rule about their efficiency and suitability, so we had to find the right combination that fitted our data better. Therefore, we performed a RandomSearch where the machine iterated 1000 times through training data to find the combination of parameters that maximizes the accuracy.

E. Model evaluation

We evaluated the models using the metrics of accuracy, precision, recall and f1-score. In the multi-class classification, these 3 metrics are calculated on a per-class basis. Moreover, the models were validated using a k-fold cross-validation (k = 10) to check how well they are capable of being trained and predict unseen data. Finally, for a more comprehensive representation of our results we calculated and plotted confusion matrices and ROC curves for each fold, thus illustrating how the ability of the classifier changes as its discrimination threshold is varied. For the multi-class classification problems, we calculated the ROC AUC for all classes using One - vs - One (0V0) strategy, which is a heuristic method for using binary classification algorithms for multi-class classification [17].

IV. Emotional level identification results

A. Received Dataset

Annotation of the data was performed based on the level of valence and arousal acquired from the self-questionnaire filled by each participant. Overall, we performed four classification attempts that can be observed in Table II. The first two concern the investigation of the presence of high arousal and positive valence (binary approach), respectively, while the other two refer to an additional attempt to discriminate emotional states among their respective levels (multi-class approach) as seen at Table II.

The training examples were 290 and their respective numbers for each class before the oversampling process are presented again in Table II. For all the classification trials the training algorithms were tested in a Python 3.6 environment and the respective training success rates were extracted for each trial. The models with the higher accuracy were stored and used later for predictions. The test data included a total of 72 examples. For these the prediction rate arousal and valence states, recall, precision and f1-score were estimated.

B. Results

Figs. 3 - 6 list the name of each classification process, the classifier that performed best in terms of accuracy, the feature selection method as well as the precision, recall, f1-score and accuracy of the chosen model.

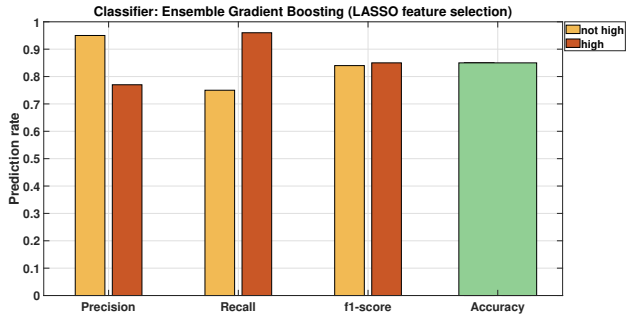

Fig. 3. Emotional Arousal binary classification.

As can be observed from Fig. 3, the Ensemble Gradient Boosting classifier proved to be superior over the other classification models tested. The model achieved to predict the presence of high emotional arousal with 85% success rate. The features for this procedure were selected using LASSO analysis. Furthermore, the model managed to predict 95% of positive instances of "high" emotional arousal that were actually correct, while the respective percentage for the "not high" examples was 77%. The recall percentages for "high" and "not high" examples were 75 and 96% respectively. Finally, the f1-score is approximately 85% for both "high" and "not high" instances.

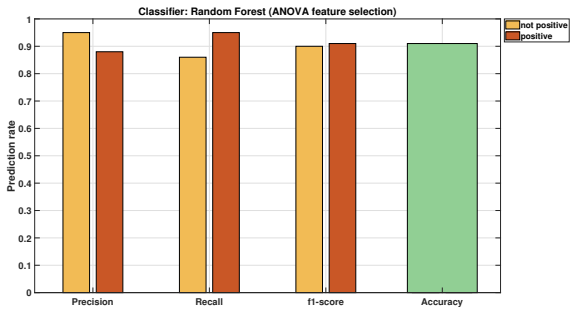

Fig. 4. Emotional Valence binary classification.

The results of our attempt to discriminate among positive and not positive emotional valence are presented in Fig 4. 9 out of 10 instances were classified correctly by the Random Forest classifier, while ANOVA analysis was used for the selection of the dominant features. In this binary classification problem, the recall rate of "positive" instances was found to be 95% and the respective precision rate achieved was 88%. Moreover, the 95% of the positively classified "not positive" cases were relevant. The f1-score for this classification trial remains over 90% for both classes.

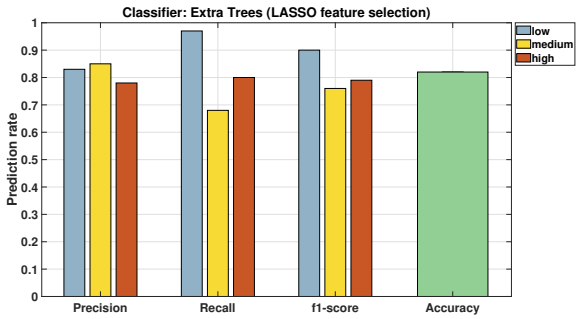

Fig. 5. Emotional Arousal multi-class classification.

In the next classification procedure, we attempted to identify between the three levels of emotional arousal, low, medium and high. The results obtained are shown in Fig. 5. As can be observed, in this multi-class problem we achieved a classification accuracy 82%. The best classification algorithm in terms of accuracy was the Extra Trees and the selection of the dominant features was performed by LASSO analysis. The positive predictive value of "medium" arousal level was found to be 85% and 83% for the "low" level. In parallel, the Extra Trees achieved satisfying sensitivity and f1-score rates for the "low" level with 97 and 90% respectively.

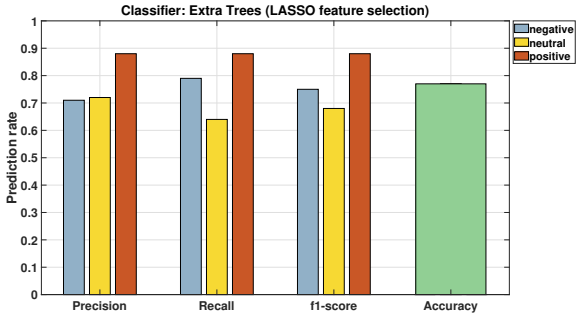

Fig. 6. Emotional Valence multi-class classification.

The final classification problem focused on distinguishing between three levels of emotional valence i.e. negative, neutral and positive. In this second multi-class classification attempt the Extra Trees classifier was once again the most efficient in terms of prediction success rate reaching up to 77% correct predictions. In detail, the "positive" instances of emotional valence reached precision, recall and f1-score percentages of 88% while the respective scores for the other two valence levels remained slightly lower.

Overall, the binary classification of emotional valence into "not positive" or "positive" using the Random Forest model achieved the best prediction rate 91%. However, when the "neutral" class was added to create a multi-class problem, this percentage was reduced by 14%. In addition, the precision, sensitivity and f1-score of the prediction of the "neutral" valence class were relatively lower than the other two classes. Finally, regarding the emotional arousal level recognition, the

success rates of the binary and multi-class problems differed only by 3% while the identification of the "high" arousal level, which demonstrates significant emotional charge, was correct in 85% of cases.

V. Discussion

In the present manuscript, we report our work focused on classifying emotional arousal and valence into their relevant levels, using eye and gaze tracking features. To this goal, an experimental trial was performed, for collecting eye and gaze tracking data from subjects watching emotion-evoking video-clips and self-accessing their emotions.

For each study participant, several eye and gaze related identification parameters were extracted, feature selection and data processing techniques were implemented and machine learning models were trained. A number of classifiers were tested and the best performing classifiers were identified.

From the results presented in Section IV-B, the highest success rate was observed during binary classification between not positive and positive emotional valence, with the Random Forest algorithm outperforming all others achieving 91% accuracy as it is able to effectively reduce the risk of over fitting, balance the error for unbalanced data, and determine the importance of features quickly. However, the inclusion of the "neutral" class proved to be challenging leading to a significant decrease in our system's performance. Regarding the emotional arousal level estimation, the binary as well as the multi-class identification tasks provided promising results reaching up to 85% correct predictions.

In this work, we have verified that using only eye and gaze metrics to estimate emotional arousal and valence produces similar results to [8], [9]. Furthermore, when comparing our results to those of [5], [6], [7], it must be pointed out that our results encourage the development of an emotion identification system with high discretization ability.

VI. Future work

The results presented in this article demonstrate the potential of utilizing machine learning optimization for discriminating between the various emotional states, while reinforcing the imperative need for future research. Towards this direction we aim to extend the dataset by adding more participants into the study. Furthermore, we are currently investigating new models as well as the applicability of deep learning methods in order to create a model for synchronously estimating emotional arousal and valence levels with high efficiency. Finally, we plan to compare our findings with research works that utilize multimodal approaches, i.e employ additional biosignals and investigate the necessity and the potential of combining eye and gaze data with other biometrics for increased performance with respect to the computational cost.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826429 (Project: SeeFar). This paper reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

References

[1] J. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161-1178, 1980, doi: 10.1037/h0077714.

[2] F. Citron, M. Gray, H. Critchley, B. Weekes and E. Ferstl, "Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia*, vol. 56, no. 100, pp. 79-89, Apr. 2014, doi: 10.1016/j.neuropsychologia.2014.01.002.

[3] V. Skaramagkas et al., "Review of eye tracking metrics involved in emotional and cognitive processes," in *IEEE Reviews in Biomedical Engineering*, 2021, doi: 10.1109/RBME.2021.3066072.

[4] J. Zhai, A. B. Barreto, C. Chin and C. Li, "Realization of stress detection using psychophysiological signals for improvement of human-computer interactions," in *Proc. IEEE SoutheastCon*, 2005, pp. 415-420, doi: 10.1109/SECON.2005.1423280.

[5] C. Aracena, S. Basterrech, V. Snáel and J. Velásquez, "Neural Networks for Emotion Recognition Based on Eye Tracking Data," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 2632-2637, doi: 10.1109/SMC.2015.460.

[6] L. J. Zheng, J. Mountstephens and J. Teo, "Multiclass Emotion Classification Using Pupil Size in VR: Tuning Support Vector Machines to Improve Performance," *Journal of Physics: Conference Series. 1529.*, doi: 052062. 10.1088/1742-6596/1529/5/052062.

[7] P. Tarnowski, M. Kołodziej and A. Majkowski, "Eye-Tracking Analysis for Emotion Recognition." *Computational Intelligence and Neuroscience*, pp. 1-13, 2020, doi: 10.1155/2020/2909267.

[8] S. Alghowinem, R. Goecke, M. Wagner, G. Parker and M. Breakspear, "Eye movement analysis for depression detection," in *2013 IEEE International Conference on Image Processing*, 2013, pp. 4220-4224, doi: 10.1109/ICIP.2013.6738869.

[9] S. Al-gawwam and M. Benaissa, "Depression Detection From Eye Blink Features," *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2018, pp. 388-392, doi: 10.1109/ISSPIT.2018.8642682.

[10] A. Schaefer, F. Nils, X. Sanchez and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers", *Cognition and Emotion*, vol. 24, no. 7, pp. 1153-1172, doi: 10.1080/02699930903274322.

[11] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, 2021, doi: 10.1038/s42256-020-00280-0.

[12] D. D. Salvucci and J. H. "Goldberg, Identifying fixations and saccades in eye-tracking protocols", in *ETRA '00: Proceedings of the 2000 symposium on Eye tracking research and applications*, November 2000, pp. 71–78, doi: 10.1145/355017.355028.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. Philip Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[14] M.M. Ahsan, M.A.P Mahmud, P.K. Saha, K.D. Gupta and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 52, 2021, doi:10.3390/technologies9030052.

[15] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 2016, pp. 18-20.

[16] J. Wu, X. Y. Chen, H. Zhang, L. Xiong, H. Lei and S. Deng, "Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26-40, 2019.

[17] Y. Liu, J. Bi and Z. Fan, "A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm," *Information Sciences*, 2017, doi: 10.1016/j.ins.2017.02.016.