doi:10.3233/AISE240015

Contact-Free Emotion Recognition for Monitoring of Well-Being: Early Prospects and Future Ideas

Gašper SLAPNIČAR [a,b,1], Zoja ANŽUR [a], Sebastijan TROJER [a,c] and Mitja LUŠTREK [a,b]

[a] *Department of Intelligent Systems, Jožef Stefan Institute*
[b] *Jožef Stefan International Postgraduate School*
[c] *Faculty of Computer and Information Science, University of Ljubljana*

Abstract. Emotions are an essential constituent of well-being. They can be recognized using contact-free sensors such as cameras, based on facial expressions and physiological parameters, such as changes in temperature. We conducted an early evaluation of emotion recognition from RGB cameras using two datasets and highlighted challenges such as subject-specific relationship between facial expressions and emotions, as well as inconsistent expressions during the same emotional state. Additionally we confirmed the feasibility of measuring subtle changes in temperature between facial regions correlating to different emotional states, using a thermal camera. Finally we proposed ideas for future improvements relating to transfer learning and cross-dataset data curation, which could allow for improvements in performance leading towards practical implementation of well-being monitoring.

Keywords. contact-free monitoring, facial expressions, thermal camera, work well-being, job productivity, emotions

1. Introduction

In recent years, work conditions that influence physical and mental health have become increasingly important. Simultaneously, reports of increased stress and worsened mental health among workers are more and more common [1]. This indicates a need for better understanding of employee well-being and development of methods for its monitoring.

Technology plays an important role in this, however, its effects can be either positive or negative, depending on its implementation and use. On one hand employers can abuse technology for increased supervision and continuous monitoring that exerts additional pressure on workers. On the other hand, technology can positively encourage workers to be mindful of their well-being and health, especially in office environments where sitting in front of a screen for prolonged periods can have detrimental consequences [2].

In order for the users to accept such technology in their daily work routine, it must be as unobtrusive as possible, not interfering with their activities or requiring special interaction. Contact-free sensing is ideal for such scenarios, specifically RGB and thermal

[1] Corresponding Author: Gašper Slapničar, gasper.slapnicar@ijs.si.

cameras (can be mounted on the screen), microphone and application usage monitoring applications (can run in the background). Furthermore, it is important for such technology to preserve complete user privacy, as to both prevent abuse as well as increase user adherence.

In this paper we report an early feasibility study conducted within the Trust-ME project [3]. Our proposed pipeline is to obtain features from contact-free sensors, such as cameras, and then train emotion recognition models on existing datasets. This allows us to train initial models without the need for extensive data collection in the first phase. While such models are smaller and simpler compared to end-to-end approaches, they are more appropriate for real-time use due to much lower computational demands and simplify mechanisms of privacy preservation such as federated learning.

We initially identified psychological constructs of interest that relate to well-being and job productivity and the corresponding sensors that would allow for their measurement. This model is subject to refinement and out of the scope of this paper. In this paper we limited our initial investigation to the first phase – classification of emotional states, which are reported to be detectable from different facial expressions using RGB cameras [4]. We attempted to extract relevant facial features using state-of-the-art methods and train classification models to classify emotions related to well-being. Additionally, we confirmed the feasibility of observing physiological changes on users' faces using thermal cameras and also investigated how facial region of interest detection performs on thermal data of different quality. We report results of these early experiments, comment on the feasibility, challenges and drawbacks of existing approaches, and propose a direction for future work.

2. Related Work

In recent years, emotion recognition from video emerged as an increasingly common research topic in the literature [5]. Different methods and techniques have been proposed, many of them including an analysis of facial expressions from video. Ebrahimi Kahou et al. [6] proposed using convolutional and recurrent neural networks for facial expression analysis. In their study, they also used audio information beside visual for more accurate emotion recognition. In another example, Wu et al. [7] proposed a method that integrates the analysis of facial expressions with data on head pose and eye gaze. This way of complementing data allows for an implementation of an attention block that guides the use of facial features and utilizes the data to the greater extent. The proposed approach increased state-of-the-art accuracy.

Different types of data can be used for emotion recognition. A large body of literature is emerging on combining video data with data from other available sources. For example, Soleymani et al. [8] conducted a study using data from video (facial expressions) alongside data from electroencephalogram (EEG). Using this approach, continuous emotion recognition was enabled on participants watching emotionally colored videos. Due to the nature of the study, data on facial expressions turned out to be more useful compared to data gathered with the EEG. Another study used data from video and wearables (e.g., electrodermal activity, heart rate) [9], proposing an approach to recognise valence and arousal of experienced emotions from physiological signals. They utilized features both inside each instance and between different instances for the same video, adopting a correlational approach, achieving promising accuracy.

Shifting focus back to emotion recognition from video, some research investigated the role of context. Combining facial expression, tone and text derived from YouTube videos, Bhattacharya et al. [10] investigated how various contextual factors such as gender of the speaker and duration of the emotional episode affect multimodal emotion recognition. They concluded that the gender of the speaker played a moderating role in the multimodal features' performance, while effectivenes of the features varied across different durations of the episodes.

Last but not least, thermal imaging has also been investigated in emotion recognition. Aristizabal-Tique et al. [11] gathered thermographic data on participants in three different conditions - baseline, positive, and negative valence of emotions. Analysing the data, blood perfusion and average temperature was calculated for the regions of interest (ROIs). The results show high correlations between changes in temperature and changes in valence across conditions. Bhushan et al. [12] also utilized thermal data when they conducted a study protocol inducing complex emotions. Similarly, they focused on certain ROIs and found a pattern connecting experienced emotions to temperature changes.

Following the related work, we identified lack of work in unobtrusive contact-free monitoring of emotions in office environments, which could in turn help measure and potentially increase both well-being and satisfaction of workers, as well as positively influence their productivity.

3. Data

Data collection is generally complex and expensive in the sense of protocol design, subject recruitment and other administrative tasks. As our initial focus was early feasibility analysis limited to emotions, we opted to use freely available open-access datasets, covering a broad spectrum of data modalities, including contact-free sensors such as cameras [4]. Our ultimate goal is to leverage existing datasets for emotion recognition to build early models, and then use the outputs of these classifiers as inputs into additional models for well-being monitoring. The datasets that we chose best imitate our end-goal setup, as they provide suitable data modalities and labels from a variety of subjects in naturalistic settings.

3.1. Emotion Recognition using RGB Camera

We identified two datasets that include RGB image data and labels of interest for our task, specifically Facial Expression Recognition (FER) [13] Dataset and Bahcesehir University Multimodal Face Database of Spontaneous Affective and Mental States (BAUM-1) [14].

FER dataset is relatively old and was a cornerstone in early emotion recognition from visual cues of the face. It consists of over 30000 48x48 pixel grayscale images of faces, originating from sources like movies and online content. The faces have been automatically centred in images and occupy about the same amount of space in each image. Each image is accompanied with an emotion label, which is one of 7 possibilities: anger, disgust, fear, happiness, sadness, surprise and neutral. These are the emotions that are defined to be culturally universal by the Emotion Facial Action Coding System (EMFACS). The dataset was designed so that the expressions are as clear as possible. State-of-the-art

deep learning models achieve accuracies up to 75% using this dataset [15]. Since the individual images in this dataset are mostly independent and labelled individually, there is high confidence in correlation between the facial expressions on an image and its label.

BAUM-1 dataset also contains videos of human faces corresponding to 31 subjects. As the data originates from video, audio is also available, but out of the scope of this paper. The data was collected under naturalistic conditions (elicited emotions, but subjects were given time to express themselves freely), with the same EMFACS emotions labelled. However, several additional classes were labelled, including boredom, contempt, confusion, thinking, concentrating and bothered. These latter states are more subtle and difficult to distinguish, making this dataset inherently more challenging due to both greater number of classes, as well as their subtle nature. In total there were 12 distinct class labels in this dataset. As our ultimate goal is to continuously monitor knowledge workers at their workplace (an office with a PC), a broad spectrum of subtle and varied emotional states can be expected, so an early evaluation on a challenging dataset has its merit in establishing an initial baseline. Additional classes such as *thinking* are expected to be quite important and present in daily routine of knowledge workers. Furthermore, training models for specific sets of emotions has additional potential in then using their outputs as inputs into more complex models for well-being monitoring. State-of-the-art multi-modal deep learning models report accuracies of up to 77% on the BAUM-1 dataset [16]. Unlike in FER, the data in BAUM-1 was labelled on per-video basis, meaning a whole video was assigned the same label. Naturally this means that potentially a large variation of expressions can have the same label, making training a model more difficult, as we will see in Section 4.

3.2. Detection of Physiological Changes using Thermal Camera

There is consensus in literature that physiological changes are well-correlated with different psychological states. Thermal imaging can be used to detect changes in temperature between ROIs, which correspond to changes in psychological or emotional state. A recent example dataset containing such thermal recordings with corresponding emotion labels was provided by Aristizabal-Tique et al. [11] and was described in Section 2. Our aim was to use their thermal data for precise ROI segmentation and verify the segmentation performance on a consumer thermal camera with lower specifications (resolution), which would be feasible to use in our planned larger-scale experiments.

4. Experiments and Results

Given the prevalence of emotion recognition from facial expressions in literature, we decided to initially replicate this approach and verify it on the BAUM-1 dataset, as it contains more varied and subtle labels that closely resemble our target application (e.g., addition of *thinking* class). We devised a pipeline that initially preprocesses the video data. It then takes remaining facial images as inputs and detects keypoints that are used for computation of face blendshapes. These are complex features that have a direct meaning and interpretation (e.g., how much the corner of the mouth is raised). We investigated distributions of features and how their changes correlate to different labels. Finally we trained and evaluated a machine learning (ML) model for prediction of emotions.

4.1. RGB Data Preprocessing and Feature Analysis

After splitting the videos into individual frames, we then subsampled the images. While video is typically recorded at 30 frames per second, facial expressions and emotions do not change as rapidly as every 30 ms. Thus we decided to subsample the data by keeping a frame every 100 ms. This is still conservative sampling and could potentially be further reduced, but more plentiful data is generally desirable for training of ML models.

We then identified several classes that are psychologically similar and usually manifest in similar facial expressions as well (e.g., *contempt* and *anger*). We merged such classes in order to simplify the initial problem, given its inherent difficulty. We additionally removed the class *unsure*, as its meaning was not clear.

Afterwards, we computed the facial features from each image by using Google MediaPipe framework. The latter offers real-time face detection and landmark detection based on BlazeFace, a lightweight and well-performing face detector [17]. The framework provides 52 blendshapes for each face detected on an image, computed from keypoints shown in Figure 1. These blendshapes were then used as our set of facial features.

Following the computation of features, we conducted an analysis on the distribution of their values with different classes. For instance, we did an analysis of smiling facial expression, which we defined as the average of *mouthsmile* features from MediaPipe, shown in Figure 1. We investigated its expected correlation with the *happiness* class by plotting the value of this feature in different frames with different labels, as shown in Figure 2. For this specific feature we observed expected changes, as the distribution of its values became more uniform when looking at *happiness* images, while it was heavily centered around zero in all other cases. For majority of the features however, the distribution was narrowly centered around zero, meaning no consistent changes were observed across different emotions.

Figure 1. Face mesh with keypoints used to compute meaningful blendshapes. Red asterisk marks corners of mouth related to smiling.

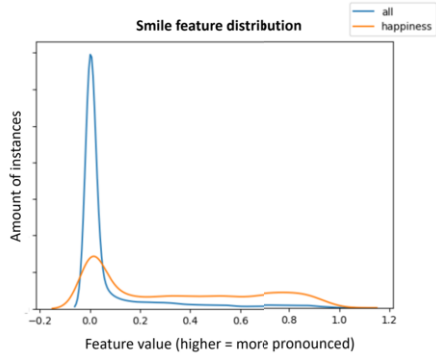

Figure 2. Distribution of values for the feature describing smiling.

There is a use for such distributions of feature values, as one can define a manual threshold-based classifier based on the distribution of a feature, which can be used to prune the data with the goal of obtaining clearer examples of emotions (based on some

expert knowledge correlating features with emotions). This can yield a subset of the initial dataset, which is expected to allow for a model to more easily learn relevant connections between features and classes.

4.2. Machine Learning Experiments

In the next step our aim was to train classification models for the datasets described previously and evaluate them robustly.

4.2.1. Classification of FER Dataset

The previously mentioned issue of a single label across whole recordings containing varied facial expressions is not present in the FER dataset, as the data is not sequential. Each image is instead an independent representation of the labelled emotion, meaning that a clearer and more consistent relationship between facial expressions and class labels is expected. We thus initially investigated classification performance of a light-weight RandomForest model (with default hyperparameters) on this dataset, using the faceblend features described earlier. RandomForest was chosen due to its fast training and relatively good performance on a variety of tasks, which make it ideal for early feasibility analysis.

Given the independence of instances, the data splitting procedure was rather simple, as we could simply randomly split all instances into the train and validation folds in a 5-fold cross validation (CV). We monitored accuracy, precision, recall and F1 score to get a robust overview of performance. The per-fold results are shown in Figure 3, but importantly they are not per subject, as we had no subject information in this dataset. The model achieved average accuracy and F1 score of 0.59 and 0.54 respectively, which substantially surpasses the baseline majority classifier.

Initial observations are very consistent performance across all folds, and relatively small discrepancies between different metrics (e.g., accuracy and recall or F1 score), compared to BAUM-1 results, which we present later. Additionally, the average performance is lower compared to BAUM-1 by about 0.1 in all metrics. This tells us that the performance is more robust in terms of predicting different classes and not converging towards the majority class, especially compared to some subjects in the BAUM-1 dataset (those, who have large discrepancy between accuracy and recall).

4.2.2. Classification of BAUM-1 Dataset

When evaluating models on the BAUM-1 dataset we had to be mindful to correctly split the data in order to avoid overfitting. Specifically, when dealing with sequential data with high sampling frequencies, subsequent frames are nearly identical and must not be split between the train and test data. This was partially alleviated with data subsampling, but even at 10 fps this problem persists. In the BAUM-1 dataset, an obvious workaround is to always split the data based on recordings, meaning that parts of the same video recording never appear in both the train and test sets. However, this option is not viable if only a single recording of a single emotion is present within a subjects data.

We first trained a completely general RandomForest classifier, again using the same facial features as inputs, and predicting the classes after merging similar emotions. Our first attempt was to use the leave-one-subject-out (LOSO) experiment, which is the most robust evaluation scheme and mimics the 5-fold CV conducted on FER (since we had no

subject information there). This experiment assumes that a general model can be trained, independent of subject. This yielded poor initial results with accuracies around 30%, directing us towards training person-specific models instead. Given the uniqueness of people and their expressions of emotions, it is not unexpected for person-specific models to be a more feasible alternative.

For further evaluation we decided to again use 5-fold CV, but within each subject data. We temporally split each recording of a subject (with the same class label) into 5 chunks, always taking 4 chunks of each recording for training and used the last one for testing. This temporal split ensured that we avoided overfitting in terms of subsequent (or close together) frames that we described earlier. We averaged the same five classification performance scores as in previous experiments, but for each subject. These results for all subjects of the BAUM-1 dataset are shown in Figure 4.

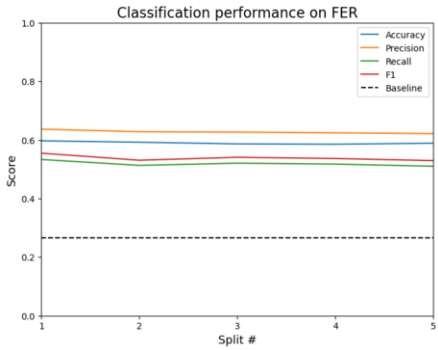

Figure 3. Classification performance scores across 5-folds of CV for FER dataset.

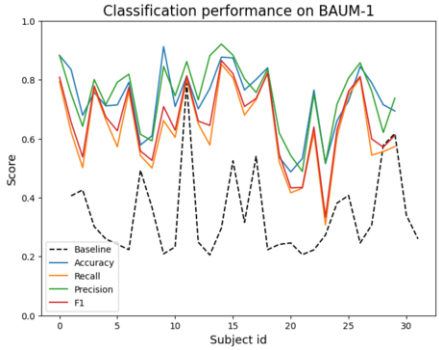

Figure 4. Average 5-fold CV classification performance scores for each subject in the BAUM-1 dataset.

We can observe notable variation in classification performance between subjects, ranging from 0.4 to over 0.9, with the average accuracy being 0.73, again substantially suprassing baseline for nearly all subjects. It should be kept in mind that we removed *unsure* class and merged *contempt* and *anger*, so this result is somewhat optimistic. The large variations shows that even within subjects robust and consistent performance is not guaranteed. When training person-specific models, one challenge was that each subject did not have individual recordings for all emotions in their data, but only a smaller (different) subset. This is important to keep in mind when interpreting results, as they are not directly comparable between subjects due to different classes and different amount of instances of each class present in the data. For example, if one only had classes that are difficult to distinguish, the results can be substantially lower compared to a subject that had very distinct classes.

Additionally, there are challenges due to the whole recording having the same label, despite great variations in subjects' facial expressions during that time. This is further confirmed by what we showed in Figure 2, where we saw that for class *happiness* the *smiling* feature still often takes values around zero, meaning the person is not really smiling. In results this can be seen in the fact that recall is often lower than precision, which can be explained by the model managing to learn some correct relationship between features and classes, and subsequently managed to detect this emotion when it was actually

present. However, when the emotion was in fact not present (despite the continuous label marking it), it did not detect it, but such errors cause the recall to be low.

Furthermore, the class balance between whatever classes were present for a given subject also varied greatly, which can be seen in the fact that F1 score is often substantially lower than accuracy for a given subject. In cases where the class distribution is more uniform the accuracy and F1 are closer (e.g., subject 7 in Figure 4) and in other cases they are further apart.

4.3. Thermal Image Analysis

As outlined in Section 2, thermal imaging was shown in literature to be an efficient approach to emotion recognition. Thus, we investigated the dataset provided by Aristizabal-Tique et al. [11], described in the Section 3.2, which contains videos of four emotional states (baseline, neutral, fear, happiness) for each of the three subjects. The videos were shot in front of a computer screen, similar to the setting that we intend to utilize in our upcoming study. Initially, we decided to replicate the approach and findings of Aristizabal-Tique et al. [11]. Furthermore, our aim was to test the employed methods on data coming from the FLIR Lepton thermal camera, which is more feasible for use in our upcoming larger-scale experiments due to its substantially lower price.

For initial visualisation and sanity check purposes, we first extracted traditional RGB heatmap representations of thermal images from each video, frame by frame alongside temperature data for each pixel of each frame.

After the initial steps, we proceeded to apply a pre-trained model for real-time face detection and landmark detection on thermal images [18]. This framework provides 54 facial landmarks for detecting various parts of the face. The model is based on an ensemble of regression trees and trained on an extensive high-resolution dataset containing 2,556 thermal images of 142 people. In the next step, we tested the generalization performance of this framework on data obtained with the FLIR Lepton. After initial testing, we noticed that the framework provides consistent results, regardless of the resolution of thermal images, as seen in Figure 5 and 6. More systematic investigating is still needed to determine the framework's robustness and consistency on low-resolution images.

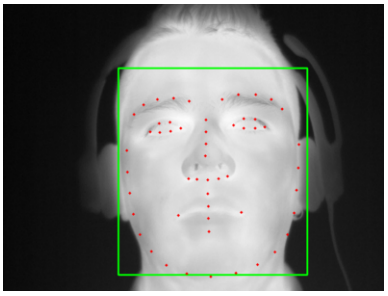

Figure 5. Facial landmarks on high-resolution thermal image from the [11] dataset.

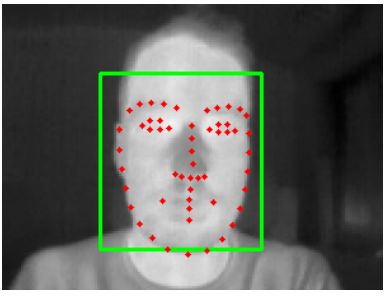

Figure 6. Facial landmarks on low-res image captured with a FLIR Lepton thermal camera.

In the next step after extracting relevant facial landmarks, we used them to identify ROIs. The ROIs (nose, forehead, and eyes) were chosen based of their known connection to psychological changes, as reported in related work [11,12]. When comparing average temperatures of the ROIs for each of the four conditions, we discovered some changes

across conditions and compared them to the work of Aristizabal-Tique et al. [11]. For two out of three subjects, our results were similar (though not the same) to the results reported by the mentioned study [11], as shown in Table 1. Discrepancies were found in some cases though (subject 3, not reported in this paper), highlighting the need for further more systematic analysis before practical application. In general, more thorough research is needed to determine the replicability and generalization possibilities of the study [11], which we intend to do in the next steps of our analysis.

Table 1. Calculated average temperatures [$^{\circ}C$] for the four ROIs across all conditions and absolute differences in T [$^{\circ}C$] between conditions fear - baseline and happiness - baseline for both subjects (S1 / S2).

ROI	ΔT_B Baseline	ΔT_F Fear	ΔT_H Happiness	ΔT_N Neutral	ΔT_{FB}	ΔT_{HB}
Nose	34.14 / 35.33	33.79 / 34.15	33.95 / 34.53	33.44 / 34.53	-0.35 / -1.18	-0.19 / -0.80
Forehead	34.88 / 35.39	34.77 / 35.43	34.78 / 35.47	34.72 / 35.34	-0.11 / 0.03	-0.10 / 0.07
Left eye	35.63 / 34.78	35.45 / 35.76	35.41 / 35.75	35.37 / 35.69	-0.18 / 0.98	-0.22 / 0.98
Right eye	35.64 / 35.29	35.71 / 35.95	35.40 / 35.83	35.40 / 35.80	0.07 / 0.65	-0.25 / 0.54

5. Discussion and Conclusions

In the early experiments presented in this paper we found that simple classifiers with facial keypoint-based features are feasible for emotion recognition, achieving average accuracies of 0.59 and 0.73, surpassing the baseline majority classifier. However, many challenges remain before practical application, which should be investigated.

We found that person-specific models are more feasible, however, variations between subjects are large. It should be investigated, especially for subjects with poor performance, whether the degradation comes from the fact that classes are similar in terms of facial expressions, or the fact that features are not informative within recordings (e.g., different facial expressions with the same label).

As this problem is not present in the FER dataset, it would be possible to attempt a transfer learning approach. Initially the model could be trained on the FER dataset and evaluate it on the BAUM-1, for the classes that overlap. In the next step it would be possible to consider the confidence of the FER model classifying BAUM-1 instances, and use this information to further prune the BAUM-1 dataset towards high-quality instances with consistent feature-class relationship. This would yield parts of the BAUM-1 data that are suitable for training a more robust model in combination with FER data.

In terms of thermal data, we confirmed temperature differences between ROIs when subject experience different emotional states. These changes will now be used as features and a classifier will be trained to predict different emotional states. Later we want to expand this in our own data collection, going from positive and negative emotions to more subtle psychological states present in work environments.

Finally, we also plan to extend our recording setup beyond RGB and thermal cameras, using the information from a microphone, an eye tracker and application usage to better model well-being and the psychological constructs comprising it. We expect the final performance to benefit from the fusion of data, potentially using several or even a single multi-modal model, like a branched neural network. Additionally, all results must be validated in terms of statistical significance.

In summary, we showed through our evaluation that many challenges remain before practical implementation of continuous monitoring of well-being for knowledge work-

ers at their workplace. The question remains how to achieve state-of-the-art recognition of complex and subtle emotions (e.g., BAUM-1 dataset) in a computationally effective manner. Another open question is how to best use different existing emotion datasets for well-being monitoring. Finally, the comparison between several smaller task-specific models and a larger end-to-end multi-modal approach warrants further investigation. The results reported in this paper are subject to further analysis and verification of statistical significance. They serve as an initial benchmark and a starting point for discussion and future work towards real-world implementation of well-being monitoring at work.

References

[1] Pagán-Castaño, E., Maseda-Moreno, A., Santos-Rojo, C. Wellbeing in work environments. 2020, Journal of Business Research, 115, 469-474.

[2] Daneshmandi, H., Choobineh, A., Ghaem, H., Karimi, M. Adverse effects of prolonged sitting behavior on the general health of office workers. 2017, Journal of lifestyle medicine, 7(2), 69.

[3] Online. Available at: https://dis.ijs.si/trust-me/. Accessed on 29 February 2024.

[4] Poria, S., Majumder, N., Mihalcea, R., Hovy, E. Emotion recognition in conversation: Research challenges, datasets, and recent advances. 2019, IEEE Access, 7, 100943-100953.

[5] Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., ... , Van den Noortgate, W. Towards measuring cognitive load through multimodal physiological data. 2021, Cognition, Technology and Work, 23, 567-585.

[6] Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C. Recurrent neural networks for emotion recognition in video. 2015, In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 467-474).

[7] Wu, S., Du, Z., Li, W., Huang, D., Wang, Y. Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. 2019, In 2019 International Conference on Multimodal Interaction (pp. 40-48).

[8] Soleymani, M., Asghari-Esfeden, S., Fu, Y., Pantic, M. Analysis of EEG signals and facial expressions for continuous emotion detection. 2015, IEEE Transactions on Affective Computing, 7(1), 17-28.

[9] Zhang, T., El Ali, A., Wang, C., Hanjalic, A., Cesar, P. Cornet: Fine-grained emotion recognition for video watching using wearable physiological sensors. 2020, Sensors, 21(1), 52.

[10] Bhattacharya, P., Gupta, R. K., Yang, Y. Exploring the contextual factors affecting multimodal emotion recognition in videos. 2021, IEEE Transactions on Affective Computing.

[11] Aristizabal-Tique, V. H., Henao-Pérez, M., López-Medina, D. C., Zambrano-Cruz, R., Díaz-Londoño, G. 2023, Facial thermal and blood perfusion patterns of human emotions: Proof-of-Concept. Journal of Thermal Biology, 112, 103464.

[12] Bhushan, B., Basu, S., Panigrahi, P. K., Dutta, S. Exploring the thermal signature of guilt, shame, and remorse. 2020, Frontiers in Psychology, 11, 580071.

[13] Giannopoulos, P., Perikos, I., Hatzilygeroudis, I. Deep learning approaches for facial emotion recognition: A case study on FER-2013. 2013, Advances in Hybridization of Intelligent Methods: Models, Systems and Applications, 1-16.

[14] Zhalehpour, S., Onder, O., Akhtar, Z., Erdem, C. E. BAUM-1: A spontaneous audio-visual face database of affective and mental states. 2016, IEEE Transactions on Affective Computing, 8(3), 300-313.

[15] Khaireddin, Y., Chen, Z. Facial emotion recognition: State of the art performance on FER2013. 2021, arXiv preprint arXiv:2105.03588.

[16] Hsu, J. H., Wu, C. H. Applying Segment-Level Attention on Bi-modal Transformer Encoder for Audio-Visual Emotion Recognition. 2023, IEEE Transactions on Affective Computing.

[17] Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., and Grundmann, M. Blazeface: Sub-millisecond neural face detection on mobile gpus. 2019, arXiv preprint arXiv:1907.05047.

[18] Abdrakhmanova, M., Kuzdeuov, A., Jarju, S., Khassanov, Y., Lewis, M., Varol, H. A. Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams. 2021, Sensors, 21(10), 3465.