BAREM: A multimodal dataset of individuals interacting with an e-service platform

Romain Belmonte, Amel Aissaoui, Sofiane Mihoubi, Benjamin Allaert, José Mennesson, Ioan Marius Bilasco, Laurent Goncalves

BAREM: A multimodal dataset of individuals interacting with an e-service platform

Romain Belmonte
CRIStAL⋆
University of Lille
France

Amel Aissaoui
CRIStAL⋆
University of Lille
France

Sofiane Mihoubi
CRIStAL⋆
University of Lille
France

Benjamin Allaert
CRIStAL⋆
University of Lille
France

José Mennesson
IMT Lille-Douai†
CRIStAL⋆
France
jose.mennesson@imt-lille-douai.fr

Ioan Marius Bilasco
CRIStAL⋆
University of Lille
France
marius.bilasco@univ-lille.fr

Laurent Goncalves
Softeam
Docaposte
France
laurent.goncalves@softeam.fr

***Abstract*—The use of e-service platforms has become essential for many applications (administrative documents, online shopping, reservations). Although these platforms have improved significantly the user experience, unexpected and stressful situations can occur. Navigation problems (latency, missing information, poor ergonomics) are not always reported to the designers. To address this problem, we propose a multimodal dataset (video, audio, and physiological data) to help implicitly quantify the impact of navigation problems on users when using an e-service platform. A scenario has been designed to generate various navigation problems which can lead to changes in user behaviour. A baseline is proposed to spot changes in user behaviour, opening the way towards automatically qualifying user experiences while using e-service platforms.**

***Index Terms*—e-service platform, multimodal dataset, behaviour analysis, anomaly detection**

I. Introduction

The popularity of e-service platforms has grown steadily in recent years. They are used in many areas to facilitate the daily life of people (administrative documents, online shopping, reservations). However, it is difficult to design a platform that is easy to use for a wide range of users. Many parameters can have a strong impact on the user experience (internet connection, computer skills, web browser). These platforms are generally designed to cover as many of these situations as possible to ensure an optimal use. However, this does not prevent the occurrence of a number of problems (latency, missing information, poor ergonomics) likely to affect the user experience.

E-service platforms offer solutions to contact the support in charge of their design. However, in some cases, flaws affecting user experience do not significantly hinder the completion of the task and therefore are not necessarily reported by the user. Users often simply choose to leave the platform without taking the time to explain their difficulties. Lack of user feedback is a problem for the designers of e-service platforms as they cannot accurately identify situations that lead to a bad user experience. Hence, the designers miss the opportunity to fix them.

There is a need to qualify user interactions with e-service platforms that lead to negative behaviours (e.g., frustration) and provide feedback to the system. To enable the development of solutions, it is crucial to collect user sessions navigation logs and other user-related modalities. Video, audio, and physiological data are subject to strict GDPR regulations which makes them difficult to exploit in public settings. However, in a lab environment, the joint analysis of these modalities is possible and can provide valuable information about user experience.

To this end, we designed a new dataset, called BAREM[1], to support automatic detection of changes in user behaviour when interacting with an e-service. A school transportation e-service platform called e-Citiz[2], developed by Softeam, and currently deployed in several French cities, was used to put users under realistic conditions. A scenario composed of several tasks has been defined to expose users to various problems that can impact their experience at different levels of intensity. Multiple modalities (video, audio, and physiological) have been recorded to help analyze user behaviour.

The paper is structured as follows. In Section II, we review the related work, i.e., similar datasets and multimodal emotion analysis. In Section IV-B2, we provide details about the multimodal BAREM dataset. In Section IV, we propose a baseline to spot changes in user behaviour when using the school transportation e-service platform. Finally, we conclude and discuss future work in Section V.

II. Related work

Navigation problems when using e-service platforms generally lead to changes in user behaviour reflecting negative emo-

⋆Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
† IMT Lille-Douai, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France

[1]BAREM stands for Behaviour Analysis for Reverse Efficient Modeling. The dataset is available upon request. Please contact I. M. Bilasco.
[2]https://www.e-citiz.com/

TABLE I
DATASETS FOR MULTIMODAL USER BEHAVIOUR ANALYSIS.

Dataset	Subjects	Setting	Modalities	Annotations
DISFA [16]	27	Watching a video	Audio / Video	6 basic expressions + intensities
DEAP [12]	32	Watching a video	Audio / Video / Physiological	Valence / Arousal
AVEC [22]	292	Interacting with a computer	Audio / Video	Valence / Arousal
SEMAINE [17]	150	Chatting with a virtual agent	Audio / Video	Valence / Arousal
RECOLA [20]	23	Dyadic interactions	Audio / Video / Physiological	Valence / Arousal

tions such as frustration. Frustration is usually associated with an emotional state change close to anger or disappointment [11] and can be expressed in many ways. We found relevant to center our literature review around multimodal emotion analysis and associated datasets.

Datasets Over the past few years, many datasets which contain visual, audio, and other modalities such as physiological signals have been proposed for multimodal emotion analysis. The most relevant datasets are reported in Table I. Settings range from passive subjects watching a video to active subjects interacting on a computer.

Video - Emotion analysis from images and videos has been attracting a lot of interest lately. Current solutions are able to predict basic emotions by analyzing facial expressions [6]. The face also provide information regarding the level of frustration of individuals [7]. Besides expression recognition, another challenging task, more specific to video, is to identify when expressions occur, i.e., expression spotting [8], [14]. Gesture analysis [9], and more broadly scene motion analysis [23], can also help to recognize emotional states and identify behaviour changes. Frustration often leads to large movements with high intensities, distinct from more ordinary behaviours.

Audio - Emotional state estimation can also be achieved through audio modality. Several states such as happiness, anger, and sadness can be predicted by analyzing the voice of individuals (i.e., intonation, enunciation, and breathing) [21]. Although the trend is to learn features, e.g., by applying a convolutional neural network on spectrograms [19], traditional features are still commonly used today [4]. The most popular ones are probably Mel-Frequency Cepstral Coefficients (MFCCs) [13].

Physiological signals - Physiological modalities such as Electrodermal Activity (EDA), Electroencephalographic (EEG), and Blood Volume Pulse (BVP) can also be helpful when analyzing the emotional state of individuals [1]. Although such sensors are generally highly intrusive, they can provide valuable information to determine the level of engagement and frustration [5]. As an example, it has been applied to frustration recognition in drivers [10].

Since none of the available datasets fully meet our needs as they do not involve e-service interactions and navigation logs, we have collected our own dataset called BAREM. In this work, we focus more specifically on spotting changes in user behaviour when they are interacting with an e-service platform. We aim to identify automatically key moments in sequences where users have encountered a problem with the platform. The long term objective, out of the scope of the present work, is to be able to correlate navigation logs and user behaviour changes.

III. BAREM DATASET

In this section, the main properties of BAREM dataset are presented. We first explain the recording equipment and setup used. Then, we describe the scenario and its 4 different tasks designed to generate frustration situations and we provide some details on capture conditions. Finally, we discuss the manual annotation process to establish the ground truth. Table II provides an overview of the contents of the dataset.

TABLE II
DATASET CONTENT SUMMARY.

Subjects	18
Videos	72 (4 tasks per subject)
Setting	Interaction with a flawed e-service platform
Annotation	Manual annotation (presence/absence of changes in behaviour) Subject self-evaluation (level of frustration)
Modalities	RGB and depth videos (including face and upper body) Audio (voice, breath and keyboard input force) Screen captures Physiological signals (BVP, EEG, EDA) Navigation logs

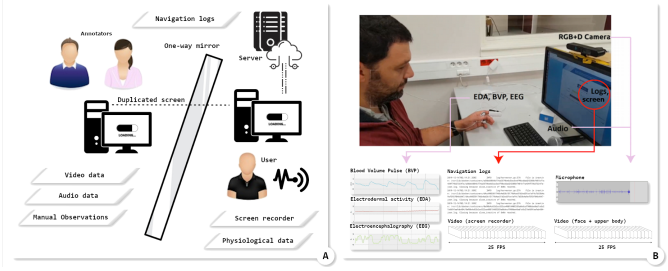

Fig. 1. A) The recording setup; B) A participant during the experiment and all the modalities collected during the scenario.

A. Equipment and Setup

Our experiments are based on an existing e-service platform called e-Citiz and developed by Softeam. We were able to make adjustments to the platform and retrieve all the data it generates according to the needs of our experiments. In order to monitor changes in subject behaviour (facial expression, body gesture, physiological variation), several modalities have been recorded and synchronized: video, audio, physiological signals (EEG, EDA, BVP), and navigation logs. The recording setup is shown in Figure 1 and the specifications of the sensors used are as follows:

Video - Video acquisition was conducted using a RGB-D Xtion pro live camera. The latter was put on top of the screen to record the face and upper body of the subject. RGB and 16-bit depth images were acquired at a frame rate of 25 frames per second. The subject screen was also captured to facilitate offline annotation.

Audio - To capture the sound emitted by the user (voice, breath, keyboard input force), a cardioid microphone was put in front of the keyboard and directed toward the subject.

Physiological signals Sensors from Biosignalsplux[3] were used to record 3 physiological signals from subjects at a frequency of 1000 Hz.

EDA sensor records the change in electrical skin properties resulting from a response of the sympathetic nervous system to a stimulus. EDA sensor is composed of two electrodes placed on hand index and ring fingers by using pre-gelled self-adhesive disposable electrodes. The resulting signal is the amplified difference between the two recorded signals.

BVP sensor measures cardiovascular dynamics by detecting changes in the arterial translucency. When the heart pumps blood the arteries become more opaque, allowing less light to pass from the emitter on the sensor through to the receiver. The sensor is placed on the index finger.

EEG sensor is composed of two electrodes that detects the electrical potentials in the specific scalp region with respect to a reference electrode, which is placed in a region of low muscular activity (behind the ear). It corresponds to the amplified difference between signals coming from the two electrodes.

Sensors are placed on the non-dominant hand as they have to be stationary during the whole acquisition to avoid noise.

Navigation logs Logs are generated by the web server and the e-Citiz platform. Each log file contains structured lines about timestamps and various events (user request, error, etc.).

Synchronisation A timestamp associated with the beginning and end of each acquired data is saved to ensure proper synchronization despite the difference in frequency between the different modalities. Each modality is then sampled so that the first data corresponds to the start time and the last data to the end time. However, timestamps associated with the logs depends on Softeam web server while timestamps associated with other data depends on our acquisition computer. Synchronization is later achieved by matching a specific screen capture with the corresponding log line.

B. Scenario and Capture

Data capture is based on a school transport reservation scenario running on e-Citiz e-service platform. The main difficulty when designing the scenario was to ensure that the subject experienced frustration. To do so, he or she must be involved. In order to challenge the user, a time constraint (about 20 minutes) is given to achieve the scenario. In addition, the scenario is located in a geographical area unfamiliar to the subject. By proposing challenging and frustrating tasks, we increase the involvement of the subjects and the chances of generating behaviour changes. During the scenario, each subject is asked to complete 4 distinct tasks to generate different levels of frustration. The 4 tasks are as follows:

Task 1 - The subject is asked to sign up on the platform using a given email and username. However, the username already exists and the registration leads to an error message. To complete the task, the subject must select another username.

Task 2 - The subject is asked to fill in the information about 2 children. Even if cumbersome, no flaws is encountered in this task. In this way, we give the subject a sense of accomplishment and reinforce the level of frustration in the next task.

Task 3 - The subject is asked to book school transportation for the first child. Some of the information must be obtained from a document provided at the beginning of the scenario. It is also necessary to upload a file available on the computer in two formats. However, the first attempt with the first file will lead to a format error. The subject will have to repeat the upload with the second file.

Task 4 - The subject is asked to book school transportation for the second child. However, essential information regarding the itinerary is missing, making it impossible to complete the task. This is intended to have a strong impact on the level of frustration of the subject.

Eighteen adults subjects, 11 men and 7 women, have been selected to take part in the experiments. Four are above 50 years old, 11 below 30, and 3 in between. During data capture, two annotators are located behind a one-way mirror. The subjects are, therefore, alone in a room, which helps to reduce the Hawthorne effect, i.e., the subject modifies his or her behaviour as a result of the observer reactions. In contrast, the two annotators can see the subject and monitor its behaviour changes in real time. They also have a copy of the screen. Frame-level annotation is performed indicating presence or absence of changes in behaviour with respect to a neutral state. Note that annotators don't have the ability to communicate and annotation were also refined offline[4]. At the end of the experiment, subjects are asked to fill out a self-evaluation form to report their level of frustration on each task.

C. Annotation Analysis

Inter-annotator agreement (Cohen's kappa) calculated using a sliding window of size 20 indicates moderate agreement (0.45) between annotators. The analysis of the self-evaluation of the subjects shows that:

- the average frustration for task 1 is relatively high (50%) despite its simplicity; it is highly dependent on the reactivity of the subject to get around the problem;
- although task 2 does not lead to any problems, an average frustration of 20% is still observed; this may be due to

[3]biosignalsplux, PLUX wireless biosignals S.A. (Lisbon, Portugal)

[4]Annotation refinement are alternative versions more focused on activation of events. No additional information has been added but 2 new annotators were involved.

the large amount of information that needs to be filled out;
- the average frustration for task 3 is about 50% with the same findings as for task 1;
- task 4 has the highest average frustration (80%) as the subject is unable to complete the task.

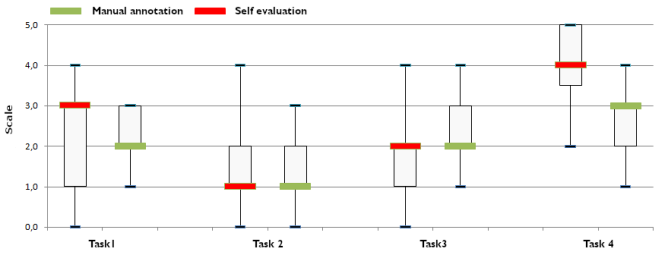

Fig. 2. Comparison between the self-evaluation of the subjects and our manual annotation.

A comparison was made between the self-evaluation of the subjects and our manual annotations. Figure 2 shows that the amount of frustration is proportional to the difficulty of the tasks in both cases, which contributes to validate our annotations reflecting behaviour changes.

We also gathered annotator observations to both define what a behaviour change in BAREM dataset means and how they are expressed. These observations are particularly useful to better understand the results in the experimental section. Here is what was mainly observed:

- changes in expression, both at the level of the mouth (pinching, smiling, tongue movements), and the eyes / eyebrows (squinting, lifting);
- global head / body movements (balancing forward / backward and right / left, moving forward to better see / concentrate, looking at the instruction sheet, intense breathing, throat movements);
- hand movements (scratching head, mouth, eyes);
- composition of these events.

IV. Behaviour Changes Spotting

In addition to BAREM dataset, we also provide a baseline to spot user behaviour changes. The latter is evaluated on all available modalities (video, audio, and physiological). We opted for an unsupervised approach since our manual annotations were conducted using only the audio and the visual feedback perceived by the annotators.

A. Baseline

The selected approach [18] consists of a two-class ordinal regression which involves two steps. The first step is to perform a first anomaly detection based on existing unsupervised approaches. We associate changes in behaviour with an anomaly, as the normal fulfilment of the task should not induce behaviour changes. Once this detection is done, subsets of what is called pseudo-normal and pseudo abnormal-data can be selected with reasonable confidence. These subsets are then used to train deep neural networks in an end-to-end fashion. An iterative learning which allows to significantly boost performance is proposed in the original work. For the sake of simplicity, the latter part is not included in our work.

1) Pseudo-Label: Two well known unsupervised anomaly detection approaches, Isolation Forest [15] and Local Outlier Factor [2], have been chosen to perform initial pseudo-label predictions.

As input for these approaches, we have extracted 2048 features from the last dense layer of a ResNeXt50 [24] pre-trained on ImageNet [3] for video data. A principal component analysis has been applied to keep only the 100 most significant components. Twenty MFCCs have been computed for audio data. Signal filtered with a butterworth filter have been used for physiological data.

The granularity of the pseudo-annotation is guided by the frame rate. At the video level, a prediction is made for each frame. At the audio level each segment have the same duration based on the frame rate (1/25 = 40 ms. The physiological annotation follow the same temporal unit.

Similar to [18], we have selected a percentage of the most normal and the most abnormal data based on the anomaly score obtained with these approaches. This allows to obtain reliable subset of data for each class.

2) Video Architecture: A frozen ResNeXt50, pre-trained on ImageNet, is used as image features extractor (see Figure 3). The latter is combined with a few fully-connected layers, which is trained with the pseudo-labels to perform anomaly score regression. The anomaly score, between 0 (no change in behaviour) and 1 (change in behaviour), is returned for each image of a video.

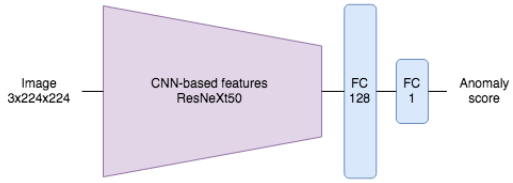

Fig. 3. The video architecture used in our experiments. The weights of ResNeXt50 have been frozen. FC means fully connected layer. A dropout of 0.2 and ELU activation have been used for all FC layers except the final one.

3) Other modalities: Regarding other modalities, signals are sampled to match the frame rate of the video. Preprocessing of the data is performed to extract MFCCs for audio and to filter physiological signals (see IV-A1). Each modality is then given as input to a neural network composed of a few fully connected layers (see Figure 4). Similar to the video architecture, training is based on the pseudo-labels and the last layer returns an anomaly score between 0 (no change in behaviour) and 1 (change in behaviour).

B. Experiments

In this section, we provide a performance evaluation of the proposed anomaly detection baseline on BAREM dataset, i.e.,

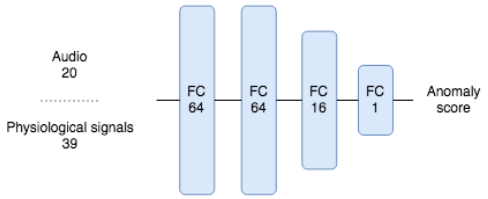

Fig. 4. The architecture used in our experiments for audio and physiological data. FC means fully connected layer. A dropout of 0.2 and ELU activation have been used for all FC layers except the final one.

its ability to spot behaviour changes. We also provide the necessary details for the replication of these experiments.

1) Implementation Details: All the models of the proposed baseline have been trained for 60 epochs with ADAM optimizer and a MSE loss function. The batch size has been set to 32 and the learning rate to 0.0001. A scheduler with a patience of 8 was used to reduce the learning rate in case of plateau.

Savitzky-Golay filter with a window of length 65 and polynomial of order 3 has been applied to smooth model predictions over time. These are then thresholded to define whether they correspond to an anomaly or not. The threshold is set, for each video, as follows:

$$Threshold = mean_p + v \times (max_p - mean_p)$$

with $mean_p$ the mean of the predictions, max_p the maximum prediction, and v a variable. Sequences with a length inferior to t are discarded as they are not relevant. v and t are set using validation data:

- 0.1 and 24 for video;
- 0.1 and 29 for audio;
- 0.1, 0.5, 0.01 and 42, 36, 20 for EEG, EDA, and BVP, respectively.

2) Dataset: All experiments included in this work are entirely based on the proposed BAREM dataset. 3 subsets have been defined:

- subjects 3 and 14 are used as unseen test data;
- subjects 8 and 11 are used as validation data;
- all other subjects are used both as training data (i.e., with pseudo-label) but also as test data given the unsupervised nature of the proposed baseline.

As explained in Section IV-A1, a percentage of the most normal and the most abnormal data (% normal / % abnormal) have been selected for each modality. After some investigation, we set the percentages to:

- 5/10 for video;
- 5/15 for audio;
- 20/10, 5/15, 5/15 for EEG, EDA, and BVP.

3) Evaluation Metrics: We propose an evaluation per interval commonly used in event spotting, such as facial expressions [14]. An interval is a succession of similar predictions. Temporal intersection over union ($tIoU$) between the predicted interval and the ground truth interval is used to define True Positive (TP) as follows:

$$TP = \frac{I_p \cap I_{gt}}{I_p \cup I_{gt}} \geq k$$

with I_p the predicted interval, I_{gt} the ground truth interval, and k a threshold set to 0.4 in our experiments. The latter allows addressing possible inaccuracies in the annotations. Given the number of ground truth intervals (nI_{gt}), the number of predicted intervals (nI_p), and the number of TP (nTP), we can define False Positive (FP) and False Negative (FN) as follows:

$$FP = nI_p - nTP$$
$$FN = nI_{gt} - nTP$$

Finally, to evaluate user behaviour changes detection performance, we compute Precision, Recall, and F1-Score as follows:

$$Precision = \frac{nTP}{nI_p}$$
$$Recall = \frac{nTP}{nI_{gt}}$$
$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

4) Results: Table III shows anomaly detection results on seen subjects, i.e. used for unsupervised anomaly detection (see Section IV-A1). Overall, F1-Score is low for all modalities mostly due to a large number of false positives. Note that this is a complex problem given the subtlety of abnormal events and their lack of universality (see Section III-C). As a coarse comparison, similar results are observed on related tasks such as facial expression spotting [8], [14], which tends to confirm the difficulty of behaviour changes spotting. Results are the lowest on physiological data. Besides the difficulty of the problem, this could be related to the ground truth defined from video and audio only. A different temporality of the anomalies might be observed on physiological data.

TABLE III
RESULTS ON SEEN TEST SUBJECTS (ALL EXCEPT S08, S11, S03, AND S14).

	Video	Audio	EEG	EDA	BVP
Total	570				
TP	132	60	21	53	35
FP	444	787	386	843	760
FN	438	510	549	517	535
Precision	0.232	0.105	0.037	0.093	0.061
Recall	0.229	0.071	0.052	0.059	0.044
F1-Score	0.23	0.085	0.043	0.072	0.051

To ensure reliable evaluation, we also report results on unseen subjects, i.e., not used for unsupervised anomaly detection. The latter are presented in Table IV. Findings remain similar to those of seen subjects. The variability observed between Table III and IV may be due to the small number of unseen subjects and their inherent difficulty.

V. CONCLUSION

In this work, we introduced a multimodal dataset (video, audio, physiological data) of 18 individuals interacting with a school transportation e-service platform. The capture was performed through 4 tasks with potential flaws for a total of about 20 minutes. Data have been manually annotated with 2 classes, i.e., normal and abnormal, depicting changes in

TABLE IV
RESULTS ON UNSEEN TEST SUBJECTS (S03 AND S14).

	Video	Audio	EEG	EDA	BVP
Total	106				
TP	22	18	6	27	0
FP	218	189	79	244	4
FN	84	88	100	79	106
Precision	0.208	0.170	0.057	0.255	0.0
Recall	0.092	0.087	0.071	0.1	0.0
F1-Score	0.127	0.115	0.063	0.143	0.0

user behaviour. Each subject also filled a form with a self-evaluation of their level of frustration on each task. Such a dataset could help in the automatic rating of satisfaction in people using e-service platforms. The benefits are numerous, e.g., guide the user in his task, help him to solve his problems, give feedback to the designers of the application.

A baseline to spot user behaviour changes was also proposed. The performance of the latter on BAREM dataset highlight the challenges with such data. A high number of false positives is mainly observed, which may be related, among others, to the subtlety of user behaviour changes.

The proposed baseline can be improved in many ways: better exploit temporal information, fusion of the modalities within the same model, online predictions, multi-scale processing, and more. It also seems interesting to consider the other modalities provided, especially navigation logs, and correlate them to the changes in user behavior. Finally, it appears relevant to study the possible delays that may occur between the different modalities.

REFERENCES

[1] Mouhannad Ali, Ahmad Haj Mosa, Fadi Al Machot, and Kyandoghere Kyamakya. Emotion recognition involving physiological and speech signals: A comprehensive review. In Recent advances in nonlinear dynamics and synchronization, pages 287–302. Springer, 2018.

[2] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pages 93–104, 2000.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[4] Shubham Dham, Anirudh Sharma, and Abhinav Dhall. Depression scale recognition from audio, visual and text analysis. arXiv preprint arXiv:1709.05865, 2017.

[5] Ashlee Edwards and Diane Kelly. Engaged or frustrated? disambiguating emotional state in search. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 125–134, 2017.

[6] Rosenberg Ekman. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997.

[7] Joseph F Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. Automatically recognizing facial indicators of frustration: a learning-centric analysis. In 2013 humaine association conference on affective computing and intelligent interaction, pages 159–165. IEEE, 2013.

[8] Ying He, Su-Jing Wang, Jingting Li, and Moi Hoon Yap. Spotting macro-and micro-expression intervals in long video sequences. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG), pages 238–244, Los Alamitos, CA, USA. IEEE Computer Society.

[9] Nathan Henderson, Jonathan Rowe, Luc Paquette, Ryan S Baker, and James Lester. Improving affect detection in game-based learning with multimodal data fusion. In International Conference on Artificial Intelligence in Education, pages 228–239. Springer, 2020.

[10] Klas Ihme, Anirudh Unni, Meng Zhang, Jochem W Rieger, and Meike Jipp. Recognizing frustration of drivers from face video recordings and brain activation measurements with functional near-infrared spectroscopy. Frontiers in human neuroscience, 12:327, 2018.

[11] Ashish Kapoor, Winslow Burleson, and Rosalind W Picard. Automatic prediction of frustration. International journal of human-computer studies, 65(8):724–736, 2007.

[12] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing, 3(1):18–31, 2011.

[13] Hyejin Koo, Soyeong Jeong, Sungjae Yoon, and Wonjong Kim. Development of speech emotion recognition algorithm using mfcc and prosody. In 2020 International Conference on Electronics, Information, and Communication (ICEIC), pages 1–4. IEEE, 2020.

[14] Jingting Li, Su-Jing Wang, Moi Hoon Yap, John See, Xiaopeng Hong, and Xiaobai Li. Megc2020-the third facial micro-expression grand challenge. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG), pages 234–237. IEEE Computer Society, 2020.

[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining, pages 413–422. IEEE, 2008.

[16] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing, 4(2):151–160, 2013.

[17] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE transactions on affective computing, 3(1):5–17, 2011.

[18] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12173–12182, 2020.

[19] Michalis Papakostas, Evaggelos Spyrou, Theodoros Giannakopoulos, Giorgos Siantikos, Dimitrios Sgouropoulos, Phivos Mylonas, and Fillia Makedon. Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. Computation, 5(2):26, 2017.

[20] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pages 1–8. IEEE, 2013.

[21] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. Speech communication, 40(1-2):227–256, 2003.

[22] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pages 3–10, 2013.

[23] Tian Wang, Meina Qiao, Aichun Zhu, Yida Niu, Ce Li, and Hichem Snoussi. Abnormal event detection via covariance matrix for optical flow based feature. Multimedia Tools and Applications, 77(13):17375–17395, 2018.

[24] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017.