Thermal Frustration in the Wild: The Contextual Dependence of Multi-modal Affect Detection

Youssef Mohamed[1], Maria Teresa Parreira[1]

Abstract— **Multi-modal affect detection systems are substantially impacted by contextual changes, both due to differences in human affect and to the perception systems that are used to capture data. In this work, we developed a multi-modal frustration detection system using thermal imaging in the wild, and compared its performance when used in a lab environment. A Gaussian Naive Bayes model was created to classify if a person is in a frustration state or a non-frustration state. To train the model, we collected a data set of 81 participants interacting with a frustration inducing robot in the wild. The model was trained using three feature modalities: RGB, thermal imaging and multi-modal (RGB + thermal). RGB features included 18 facial action units (AUs), while thermal features included four facial regions: nose, forehead, cheek and lower lip. When using only RBG features, the model reached an accuracy of 92% , 70% on thermal features only and 71% when both features are combined. In addition, we investigated if models trained on data from the wild could be used to predict frustration on data collected in the lab and vice-versa. The model that only used RGB features reached optimal accuracy of 70% when trained on data collected in the lab and tested on data collected in the wild. Using only thermal features yielded maximum accuracy of 57% when trained on data collected in the wild and tested on the data collected in the lab.**

Human-robot interaction; Thermal imaging; Frustration; Action units;

I. Introduction

In a data driven approach for multi-modal affect detection, the context and the environment which the data is collected becomes an integral part of the developed systems. Research in environmental psychology [1] suggests that environmental impact on human emotions can be detected using objective measurements like electroencephalography (EEG), functional magnetic resonance imaging (fMRI) and near infrared spectroscopy (fNIRS), indicating that these responses are caused by a spontaneous physiological reaction. Nonetheless, the use of those methods in human-robot interaction (HRI) can be intrusive and unnatural when detecting affective states.

It is important to note that even when using the most elaborate set of sensors to detect stress and frustration [2], the systems are typically deployed in a controlled lab environment [3], [4] with no consideration for other contextual impacts that can drastically affect the predictions made. Hence, it is essential to start deploying systems in an uncontrolled environment, to realistically evaluate the robustness of those systems, and to understand the realistic nature of the affective states predicted by them in different environments.

[1]Youssef Mohamed, Maria Teresa Parreira are affiliated with KTH: Royal institute of technology, Stockholm, Sweden ymo@kth.se and iolanda@kth.se

Further, the use of subjective measurements like self-assessments is important for the general understanding of the perceived impact of different environments on human emotions and affect. However, the subjectivity of the human consciousness makes it challenging to compare and justify those conclusions [5], in particular in times when instantaneous decisions about the interactions are needed.

Feeling frustrated while using robots is inevitable, as robots are prone to behavioural errors such as social norm violations, or technical errors like speech recognition failures [6], [7]. If left unmitigated, the feeling of frustration will have a negative effect on the acceptance of the robot in the environment [7]. Furthermore, frustration can be associated with lower levels of productivity [8], motivation [9], and trust [6], and higher levels of aggression [10], [11]. Hence, robots should have the ability to detect frustration and mitigate its effects in real time.

In previous work [Anonymous], we successfully collected data in a lab environment to train a frustration detection system based on thermal imaging, RGB features and Electrodermal activity (EDA). To acknowledge the importance of using systems in different environments, in this work we deploy a similar system in the wild, to investigate and explore the differences of the signals detected in the wild compared to the lab. In addition, we address the challenges in detecting frustration due to repeated failure [12] during collecting data in the wild.

This will be achieved by:

- Conducting an in the wild data collection extracting thermal and RGB features;
- Creating a machine learning model that can predict frustration in an uncontrolled environment;
- Testing the transferability of the models between lab and wild settings.

II. Related Work

Transferability in the affective computing community has been an ongoing challenge since the start of the field, due to the variability of the environment and how differently people react to stimuli. In this work we address these challenges using thermal imaging to detect frustration.

A. Effects of the Environment

[13] argues that robotic design decisions should be based more on in the wild observational analysis rather than controlled lab settings. The authors have analyzed data from two different interactions: deploying a robot in a conference setting and using a ”roboceptionist” in an entrance of the

institute. The observational analysis proved that environmental aspects can easily change the aspects of the interactions, perception of participants towards the robots and, most importantly, the behaviour of the participants towards the robot. Conversely to comparing two different in the wild settings, a direct comparison of a lab versus wild setting was made in [14], between the same setting of a robot performing animated-like behaviours in both a lab setting and in a public networking event. Both qualitative and quantitative data were collected, which have shown that people have reacted to the robot differently in the two conditions, especially when it comes to gazing behaviour towards the robot.

These studies show the effect of the environment on people's perception of the robot and highlight the importance of investigating the impact of changing the environment on robotic perception systems and human emotions.

B. Frustration Detection

One of the most commonly experienced affective states in HRI is frustration [15]. Therefore, several approaches have been implemented to detect it. For example, in [2] the authors used a set of sensors, including skin conductance, pupil trackers, posture and pressure sensors to predict frustration. The authors recruited 24 participants to interact with a tutoring virtual agent while doing a "towers of hanoi" activity. The best performing model was a Gaussian Process model which reached an accuracy of 79%. Taylor et al. [16] have used a similar approach by making use of three wearable sensors to detect frustration: ElectroDermal Activity (EDA), heat flux and skin temperature. The participants were instructed to play a modified version of the game "Breakout", on which the researchers had introduced some latency to induce frustration. Naïve Bayesian models were trained to classify frustration, reaching an accuracy of 80%. In both of these works, although the models have reached high accuracies, the sensors used can be intrusive and are impractical in more socially dynamic environments. An alternative approach, as presented by Bosch et al. [17], utilizes an RGB camera to detect facial action units (FAUs) during a physics-based playground activity. The study involved participants applying basic physics principles to solve a puzzle. The authors extracted both face-only and interaction-only features, which resulted in AUC scores above chance for detecting emotional states such as boredom, confusion, frustration, delight, and engagement.

C. Affective State Detection and Thermal Imaging

Vision-based cameras are commonly used for action unit and body movement extraction. For instance, [18] used a Microsoft Kinect for six basic emotions prediction. A uni-modal neural network was trained on both the facial expression and body movement streams using late fusion. 93% was the accuracy achieved by the neural network.

Although the use of RGB cameras can lead to high performing models, these cameras are dependant on lighting conditions of the recorded dataset, and other environmental conditions. Self-reported measures and conflicting facial expression labels are other factors that these models can be heavily affected by [19].

Alternatively, thermal cameras use far infra-red to measure the radiation emitted by warm objects, which is independent of reflected light [20]. Hence, thermal imaging can be used to overcome RGB cameras' limitations, as the thermal spectrum is not affected by light presence and it is able to record objective measures such as changes in skin temperature [21]. It has been established in the literature that stress and cognitive load have apparent effects on skin temperature [22]–[26], motivating the use of thermal imaging for affective state detection in HRI scenarios. In [27], a thermal camera was mounted on a Meka robot to measure facial temperature variations while playing a card-based quiz game with the robot. The authors tested different environmental setups with the positioning of the robot. They concluded that significant effects can be seen on the nose temperature of the participant when the robot is positioned closer to their personal space, causing a higher stress response.

The present study is based on our previous work on thermal imaging and frustration in a lab environment [Anonymous]. Through interacting with a robot, participants would experience two types of frustration: cognitive load-induced frustration and failure-induced frustration. The latter occurs when a person fails to overcome the cause of the failure [28], namely due to apparent technical failure in a robot. We collected thermal imaging data, facial expression (action units, AU) through RGB imaging and EDA. We concluded that thermal imaging can be used on its own to detect frustration in both conditions, with similar model accuracies to models trained on RGB features. When data from the failure-induced frustration case was used as training input, the model reached an accuracy of 81% with just RGB features , 64% using only thermal features, 55% using EDA, and 74% when using all modalities. Furthermore, the highest accuracy for the thermal data was reached using three facial regions of interest: nose, forehead and lower lip.

To the best of our knowledge, there is no study that develops non-intrusive frustration detection systems in an in the wild (uncontrolled) environment. Further, we test model transferability of these systems between lab and wild environments.

III. Methodology

In this section, we discuss the details of the two datasets used in this study and how the data was collected in both the lab (controlled) and in the wild (uncontrolled) environments. Model creation will also be addressed.

A. Datasets

In the present work, we want to develop frustration detection models that are trained on data collected "in the wild" (uncontrolled environment). We design a frustration-inducing interaction and collect thermal and facial action-units data from each participant, similarly to what was conducted in [Anonymous]. As this study is used as a reference for

TABLE I: Total number instances obtained from the data collected in the lab, for failure-induced frustration (F) and non-frustration instances (NF).

	No. of Instances	
Window (s)	**F**	**NF**
1	1127	1010
3.5	322	303
7	161	151

the present work, we describe its data collection process, task and results below.

1) In The Lab: A study on frustration detection on users interacting with a social robot in a lab setting was carried out in [Anonymous]. The dataset includes data from 18 participants in a total time of 180 minutes of interaction in a task that leads to failure-induced frustration. The age of the participants ranged from 21 to 46 years *(M = 27.80, SD = 6.18)*.

In this study, two tasks were presented to the participants, separated by a resting period. The instructions and guidance during the tasks were provided by a NAO robot. A thermal camera and an RGB camera were mounted on the table for data collection. Participants were equipped with a wrist EDA sensor. One task consisted of a quiz where the answers of the participant would be falsified by the robot, even if they were correct, leading to frustration by failure. In the other task, the participant had to program a virtual robot to go from one point to the other and alternating with a mental rotation task at the sound of a buzzer, which induced frustration due to cognitive load [28]. The data from cognitive load-induced frustration task was not used in the current work, as the type of frustration induced is different from the one induced in the wild (Section III-A.2).

Facial expression, temperature of facial regions and EDA data were collected and all the data was aggregated into non-overlapping windows of 1, 3.5 or 7 seconds. Within each window, features such as average, maximum and minimum temperatures of every facial region considered were computed. These samples were used to train a frustration detection system.

In the present work, we aim to explore failure-induced frustration. As such, we will only consider results from the failure-induced frustration task in the previous work. Frustration (F) instances were considered whenever the robot failed to acknowledge the participant's answers as correct. Non-frustration (NF) instances were collected during the resting period between the two tasks. The total number of training samples for each window size is shown in Table I.

2) In The Wild: We replicated the data collection process as much as possible from [Anonymous], with small adjustments in order to adapt the system to an uncontrolled environment.

a) Task Description: Data collection took place over one week, at the University library. A Furhat robot would

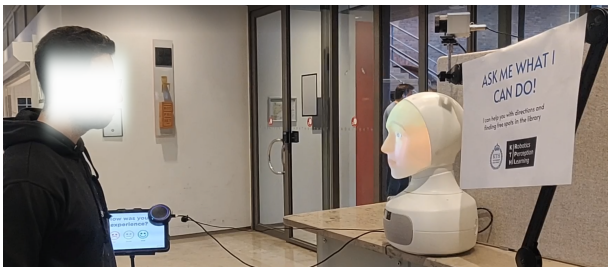

Fig. 1: The interaction setting. Visitors of the library are prompted to speak to a Furhat robot by a sign saying "Ask me what I can do!". A tablet is on a stand on the side of the participant, allowing for self-reporting at the end of the interaction.

TABLE II: Interaction reply examples.

Keyword	**Reply**
Books	The books are on the shelves.
Room	Which room?
Gallery	Which gallery?
Reception	I am the reception.

be acting as a receptionist in the library foyer, providing information to anyone that approached it. The robot's protocol for answers is keyword-based. The robot would be able to identify keywords within basic questions about the library, but it was purposely designed to give a failed response. Responses were decided based on the feedback received from the pilot study. A sample of the interaction keywords is shown in Table II. This aimed to lead the participant to an interaction loop which would be attributed to robot failure (see Figure 2). The task selected in the wild was adapted to be more sustainable than the task in the lab, as participants are less likely to interact with the robot for a longer period of time, although both tasks are designed to induce frustration due to failure.

The interaction was considered to be finished when the participant gazed away from the robot. At that moment, the robot prompted the participant to fill out a 2-step survey. The first screen asks "How was your experience?" and has only three options: good, neutral and bad. In the case that the participant chooses "bad", 4 more options come up with the text "I felt...". The options are: not sure, angry, disappointed or frustrated, as can be seen in Figure 3.

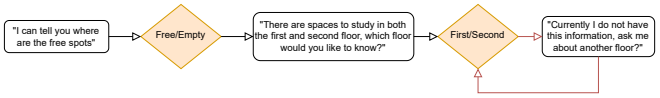

Fig. 2: Frustration loop example for keywords "free" or "empty". Participant keywords are in orange, robot replies triggered are in the white rectangles. Other loops similar to this example arise depending on the topic.

Fig. 3: 2-step survey seen on the tablet.

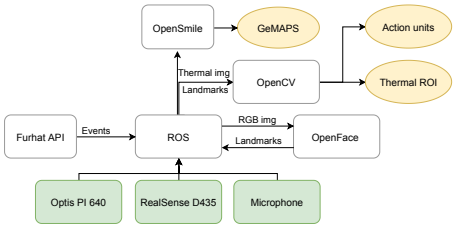

Fig. 4: System architecture.

b) Participants: We collected data from 90 interactions, with a total time of 80 minutes. 9 participants were discarded from the data, as they were wearing face masks, leaving 81 interactions with total time of 73 minutes. In interactions which included a group of people, the system would record the data of the closest person to the robot. When their face was not detected anymore, another recording would start with a new participant ID. Participants' expected age range is 18-40.

c) System Implementation: The system architecture (Fig. 4) was composed of two cameras mutually calibrated (thermal IR camera: `Optris PI 640`[1] and RGB-D camera: `RealSense D435`[2]), a directional microphone and a `Furhat`[3] robot. All of the mentioned components were synchronized in real-time using Robotic Operating System (ROS). In addition, OpenCV was used for image processing and camera calibration.

Thermal data was collected at a rate of 15 frames per second (fps). RGB camera data (action units) was collected at the same fps. The features for the thermal data were computed for all the four facial ROIs (regions of interest): nose, forehead, cheek and lower lip. As for the action units extracted, they corresponded to the Facial Action Coding System (FACS): 1 (inner brow raiser), 2 (outer brow raiser), 4 (brow lowerer), 5 (upper lid raiser), 6 (cheek raiser), 7 (lid tightener), 9 (nose wrinkler), 10 (upper lip raiser), 12 (lip corner puller), 14 (dimpler), 15 (lip corner depressor), 17 (chin raiser), 20 (lip stretcher), 23 (lip tightener), 25 (lips part), 26 (jaw drop), 28 (lip suck), and 45 (blink). Anonymized data is available upon request.

d) Labeling: Data from participants that evaluated the interaction as "good" or "neutral" in the survey were not included in training data of the model. Data labeling for the remaining participants was done based on the findings in [29]. The authors used the FACS to define frustration, which concluded that the activation of action units AU01, AU02 and AU14 is associated directly with frustration. In addition, [30] concluded that people also resort to smiling as a coping mechanism for frustration. This was considered while labeling the data for frustration instances, especially after the occurrence of failure.

[1]https://www.optris.global/thermal-imager-optris-pi-640

[2]`https://www.intelrealsense.com/depth-camera-d435/`

[3]https://furhatrobotics.com/

TABLE III: Extracted features for each modality.

Modality	Features
Thermal	ROIs temperature average ROIs temperature change ROIs temperature maximum ROIs temperature minimum
RGB	AU Intensity average AU intensity change AU maximum intensity AU minimum intensity

e) Feature Extraction: Thermal imaging and RGB data were concatenated and synchronized. After data collection, one annotator labelled each interaction according to the state of the participant (frustrated, F, or not frustrated). Assuming a data set with a total I data points (multiple data points per interaction), and a total of N measurements associated with each data point, we have each collected data point d_i associated with a set of measurements $M_i = [m_0, ..., m_N]$ (thermal data and RGB data) as well as with a label l_i (F or NF). Feature extraction for classification is performed through a sliding window of predefined length ($L = 1, 3.5$ or $7s$) and hop length $h = 0.5s$. Every instance (window) X_j used for training and testing is a feature vector that is calculated from data points d within that window. We obtain each feature f by getting the average, feature change (difference between the starting and ending value within the window), maximum or minimum values of each measurement m. The label Y_j of that instance is given by the most common label l within that window. The features extracted from the data are shown in Table III.

While we maintain the window length used in our previous work, we choose to overlap windows because that is a closer approach to how a real-use system would operate. This process is illustrated in Figure 5.

In Table IV the total number of instances obtained for each window length is shown.

f) Model: A Gaussian Naive Bayes (GNB) model was trained on the data collected from the wild. The choice was based on testing multiple machine learning algorithms: Random Forest Classifier (RFC), Support Vector Machine (SVM) and K-nearest Neighbor (KNN). The models testing was done using a Grid Search Cross Validation (GSCV) algorithm, which tested each of the models on a range of hyper-parameters, outputting the ones with the highest accuracy. None of the tested algorithms showed accuracies above chance, only the GNB showed higher performance.

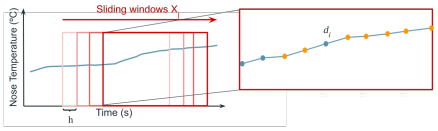

Fig. 5: Schematic view of feature extraction. Measurement m_i (in this case, nose temperature), is associated with a set of data points d, each labelled according to the frustration state of the participant (in this case, blue for non-frustration NF, orange for frustration F). One instance X_j for training is composed of features which are calculated over the set of data points d, such as average nose temperature. Y_j, the label of this training instance, is given by the most common label (in this example, $Y_j = F$).

TABLE IV: Total number of instances used for training in the cases of failure induced frustration (F) and baseline (NF) in the wild.

	No. of Instances	
Window (s)	**F**	**NF**
1	312	4296
3.5	257	4024
7	175	3724

Furthermore, the use of GNB has been proven to perform well for affect detection using facial expressions in larger datasets [31], [32].

The data is split into 3 participant-independent folds (2:1 training-validation split). Each fold does not have overlapping participants, in order to test the model on unseen participant data.We use cross-validation to report the performance Section IV.

The models were trained based on the best features selected by the Sequential Forward Floating Selection (SFFS) algorithm. The SFFS is a wrapper method that uses several greedy search methods to select the features that would yield the highest accuracy in the model. The method was adopted due to its wide use in the affective computing literature [33]–[35], over its more simple counterpart, sequential forward selection, which does not exclude the features once they are selected.

We train models on three different modalities: thermal data features, RGB data features and both thermal and RGB features combined. Each of the modalities is tested on three different window sizes for feature extraction, as is mentioned above (1, 3.5 and $7s$). That is, we obtain a total of 9 frustration detection models of varying performance.

IV. Results

This section addresses the performance and feature selection of the models trained on data collected in the wild. The results shown are the average performance on the validation sets of the 3 folds discussed in Section III-A.2.f. We also address data collection and transferability of the models between environments.

A. Model Performance

The GNB model was trained and tested on separate participant groups with no overlapping participant data between them from data collected in the wild. Table V shows performance (average accuracy and F1-score across folds) of each modality and window size. It can be seen that the accuracy of the thermal data reaches the best performance of 69% in the 7-second window, while the RGB reaches the optimal performance in the 1-second window of 92%. Similarly, when combining both the RGB and thermal modalities, performance is best for the 1-second window with a 71% accuracy.

B. Feature Selection

We train models on three different modalities: thermal data features, RGB data features and both thermal and RGB features combined. Each of the modalities is tested on three different window sizes for feature extraction, as is mentioned above (1, 3.5 and $7s$). That is, we obtain a total of 9 frustration detection models of varying performance.

1) Selected Thermal Features: The features selected in the thermal modality can be seen in Table VI. All the features concerning the lower lip region were selected, as well as average and maximum temperatures in the nose region, maximum temperature in the forehead region and minimum temperature in the cheek region.

2) Selected Action Unit Features: The features selected in the action units modality can be seen in Table VII. The average activation of AUs was not selected, but the maximum activation of multiple AUs appears to be significant towards frustration detection. The net intensity variation within the window (change) as well as the minimum intensity of some action units are also selected.

3) Selected Multi-modal Features: The features selected in the combined thermal and action units modality can be seen in Table VIII. RGB-based features are the majority of the selected features; similarly to the RBG only modality, the maximum intensity of multiple AUs was deemed significant for frustration detection. Thermal-based features are less prominent, with only the maximum temperature of the nose and forehead regions being included in the selected features.

C. Transferability

Testing model transferability is essential to broaden the understanding of environmental change within data collection. As such, we test the models from each setting (in the wild and lab environments) on the data from the alternative setting.

The best performing model from [Anonymous] (lab setting) is a KNN model, which used non-overlapping windows for feature extraction. Contrastingly, the best performing model trained on in the wild data is a GNB model (described in Section III-A.2.f). We test all the modalities and window sizes on the alternative setting data, so transferability is tested

TABLE V: Model performance for each modality and for each window size.

Modality	**Thermal**			**RGB**			**Thermal + RGB**		
Window Size (s)	1	3.5	**7**	**1**	3.5	7	**1**	3.5	7
Accuracy (%)	53	60	**69**	**92**	68	62	**71**	52	63
F1 (%)	53	68	**76**	**89**	74	56	71	58	**72**

TABLE VI: Selected features in the thermal-based modality for the 7-second window length.

Feature	**Average**	**Change**	**Max**	**Min**
Thermal Regions	Nose, Lowerlip	Lowerlip	Nose, Forehead, Lower lip	Lower lip, Cheek

TABLE VII: Selected features in the RGB-based modality for the 1-second window length.

Feature	**Average**	**Change**	**Max**	**Min**
Action Units	None	AU45, AU04, AU17	AU45, AU02, AU09 AU25, AU26	AU01, AU02, AU17

TABLE VIII: Selected features in the multi-modal (thermal + AU features) 1-second window length.

Feature	**Average**	**Change**	**Max**	**Min**
Thermal Regions	None	None	Nose, Forehead	None
Action Units	AU45, AU01, AU17	AU12	AU01, AU02, AU09, AU23, AU26	AU02, AU25

on a total of 18 models (9 trained on data from [Anonymous] and tested on data from the present work, and 9 vice-versa).

We reproduce the results from [Anonymous] for readability in Table IX. Transferability testing for each setting lead to accuracies below chance on all modalities except one, which we describe below.

1) Wild to Lab: The GNB model tested on data collected in a lab environment lead to a majority of poorly performing models, with only one modality - thermal-based models - resulting in accuracies above chance. Table X shows the performance of these models, with the 3.5-second window reaching an accuracy of 57%.

2) Lab to Wild: The KNN model tested on data collected in a wild environment also lead to a majority of poorly performing models, with again only one modality - RGB models - resulting in accuracies above chance. Table XI shows the performance of these models, with the 3.5-second window reaching an accuracy of 70%.

D. Data Collection in Uncontrolled Environments

We conduct a small observational analysis of the process of data collection in this study, which we can compare with data from the previous work. Signal-to-noise ratio is expected to decrease in uncontrolled environments, both due to participant behavior (high variance in body position, movement, and interaction duration across participants) and conditions of the environment (light and audio noise, other humans present).

In order to visualize differences between data collected in the lab [Anonymous] and in the wild, we consider the average features in the thermal data. For each participant, we characterize the distribution of temperature values in each region by taking its mean and standard deviation values. While physiological reactions to frustration cause a variation in facial temperature and expression, high deviations from the mean value of a feature (high standard deviation) may be indicative of noisy data collection.

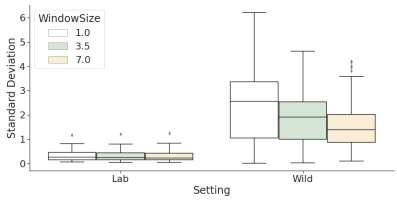

Fig. 6: Standard deviation across participants for average lower lip temperature, for each setting and window size (in seconds).

Figure 6 shows the standard deviations of the average lower lip temperature across participants and for all the window sizes. We choose this feature as it is among the selected features for the models in both studies.

V. Discussion

Data collection in uncontrolled environments is associated with a wide set of challenges, as has been discussed. This affects the amount of data collected, for instance due to partial or full occlusions of the participants' faces, but also on the quality of the data. Fig. 6 illustrates how data collected in the wild differs from in-lab data collection, with noticeably higher values for standard deviations in data collected in the wild. While this is expected given the environmental factors, we note that the task and participants are different, so slight differences in these distributions were to be expected. Given the contributions of this paper in expanding knowledge about use of affective state detection systems in uncontrolled environments, we still find these analyses to be relevant.

With this work, we intended to develop a frustration detection system for in the wild human-robot interactions. We test different modalities and processing methods (window

TABLE IX: Performance of frustration detection models from [Anonymous] (lab environment) for each modality and for each window size.

Modality	**Thermal**			**RGB**			**Thermal + RGB**		
Window Size (s)	1	3.5	**7**	1	**3.5**	7	**1**	3.5	7
Accuracy (%)	65	65	**69**	68	**81**	71	None	**75**	75
F1 (%)	60	60	**64**	68	**81**	71	None	**75**	75

TABLE X: Performance of the thermal-based modality models trained on wild data and tested on lab data.

Window size (s)	1s	**3.5s**	7s
Accuracy (%)	55	**57**	55
F1 (%)	53	**58**	53

TABLE XI: Performance of the RGB-based modality models trained on lab data and tested on wild data.

Window size (s)	1s	**3.5s**	7s
Accuracy (%)	60	**70**	60
F1 (%)	72	**78**	72

sizes), with varying performances. The thermal-based modality achieves accuracies up to 69%, with increasing accuracies as window sizes increase. Interestingly, in [Anonymous] similar conclusions were made on data collected in a lab environment. These findings coincide with Ekman's findings [36] that physiological reactions to a stimuli (such as temperature changes in facial regions) occur within 5-15 seconds.

Notably, the thermal features selected for the best performing model (7-second window, Table VI) differ from the findings in [Anonymous] for the frustration due to failure task. The cheek region, which was absent from all the models in the previous work, was selected in the case of some wild data, but only its minimum value within each window. In [37], [38] it was concluded that the cheek can warm up due to disgust and sadness.

The RGB-based modality performed in a contrasting manner as to previous results. Table V shows that performance peaks at a 1-second window length, but the accuracy of the model dips to 68% in the 3.5 second window. From the discussion on Section IV-D, and due to varying light conditions throughout the day, lower signal-to-noise ratios can be expected for RGB features, which will also vary across window sizes. Further, group-independent cross validation creates a pessimistic bias [39] with high variance across folds [40] due to noise and participant data variation.

Most action units selected from the RGB features are related to brow raising (AU04, AU01, AU02), blinking (AU45), chin raising (AU17) or lip movement (AU25). Brow raising can be an indication of focus as well as frustration [41]. According to the literature [42]–[44], there is no common consensus on which AUs relate to frustration, as it is task-dependent. However, some of the AUs commonly identified are AU09, AU10, AU23 and AU25, which can be seen among the features selected.

Finally, multi-modal model shows higher accuracies in the 1-second window size model. Overall, this model performs slightly better than a thermal-only frustration detection (71 % accuracy versus a 69% accuracy), with the smaller window size, possibly indicating the influence of RGB-based features towards classification.

When combining both thermal and RGB features, similar action units to the RGB-based modality were selected, with the addition of AU12 which correlates to smiling. As based on [30], smiling can be an indication of frustration. Only two thermal regions were selected (nose and forehead), which points to these regions being the strongest indicators of frustration [3], [38], [45], even when using RGB features.

Transferability of models is a key aspect in internal state detection, as it shows if the models can be deployed on different environments or if their usability is exclusive to the environment in which the training data was collected. In Tables X and XI we show the feasibility of applying models from uncontrolled (resp. controlled) environments on data from controlled (resp. uncontrolled) settings. Table X, which describes results from a model trained on wild data and tested with data from the lab environment, shows higher performance in the 3.5-second modality on thermal data, with accuracies only slightly below (57% compared to 60%, obtained when tested on data from the wild setting). Interestingly, the opposite model (lab data trained model, tested on wild data) performs best in the RGB-based modality in the 3.5-second window (70%). This window size lead to the worst performance among the RGB modality for the "wild trained, wild tested" models.

Lower performances in transferability testing can be caused by noisy data. For data collected in the wild, more advanced processing techniques might be needed to determine the baseline of each participant, in order to allow for transferability.

Overall, this study has shown that with short window sizes, frustration detection models in uncontrolled environments perform better when using only RGB features, whereas in the case of larger window sizes, using thermal imaging can yield better performance (See Table V). For thermal imaging to outperform RGB features in short window sizes, more processing might be needed, in order to better consider the person's baseline temperature and to account for the noise in the environment. As such, in environments where RGB features cannot be extracted (such as low light environments), thermal-based models are still able to detect frustration, although trading-off on performance. Finally, we see indications that models trained with data from a lab environment can be used to detect frustration in the wild only in the

RGB-based modality; while a model trained on thermal data collected in an uncontrolled environment has the potential to still detect frustration above chance with data from a lab environment.

VI. Limitations

While throughout this work we establish comparisons with frustration detection in controlled versus uncontrolled environments, it is important to recognize that the two settings are significantly different. Though the interaction with the robot, on both studies, is designed to elicit frustration due to failure, the tasks and task duration are not the same. The task in [Anonymous] is not reproducible in a in the wild scenario, yet the task in this work was designed to elicit the same type of frustration (failure-induced).

In the wild there are also multiple uncontrollable variables. For example, the participant's temperatures are expected to be affected by the weather and the activity that the participant was engaged in, and lighting conditions vary. Data is also imbalanced. In this work, because we establish some comparisons with prior work, performance metrics are replicated and expressed as the average performance across all the folds. Alternatively, the use of balanced accuracy [46] as a metric is an additional solution that can be used along with cross validation to address the imbalance of data. While we cannot measure the age range of the participants, users were mostly students or University faculty. This could have effects on the generalizability of our models.

Finally, the use of multiple annotators and considering the pair-wise Cohen kappa value to measure the inter-annotator agreement could have increased the robustness of the labeling method. For this study, one annotator was used due to the direct relationship of frustration to failure - if failure occurred, it was relatively clear to observe the frustration signs discussed in Section III-A.2.d.

VII. Conclusion

In this work we aimed to create a frustration detection system that can be used robustly in the wild using thermal imaging. The system creation process included data collection in an uncontrolled environment, training and testing machine learning algorithms and comparing the performance to previously created models from data collected in a lab environment [Anonymous].

The use of partially overlapping windows when extracting training samples approximates the system from a real-world scenario, as detection can start at any moment and should be continuous. The features extracted from the data in the wild included eighteen action units and four facial thermal regions. Future work may include the use of more modalities, such as audio-based features. More regions of the face may be considered in thermal imaging data, namely by separating both sides of the face, as [47] suggested that the face is thermally asymmetric.

We argue for the importance of testing transferability in affective state detection systems, and test performance of two model types on data from environment that differ from the training data. The models proved to be transferable under certain conditions, but data processing may be improved in order to allow for better transferability of the models.

Our main contribution is the development of a frustration detection system from data collected in an uncontrolled environment. We were able to achieve high performances, with the RGB-based modality achieving a high accuracy of 92% in a 1-second window, and the thermal-based model peaking in performance at the 7-second window, with an accuracy of 69%.

References

[1] Isabella Bower, Richard Tucker, and Peter G Enticott. Impact of built environment design on emotion measured via neurophysiological correlates and subjective indicators: A systematic review. *Journal of Environmental Psychology*, 66:101344, 2019.

[2] Ashish Kapoor, Winslow Burleson, and Rosalind W. Picard. Automatic prediction of frustration. *International Journal of Human Computer Studies*, 65(8):724–736, aug 2007.

[3] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. Cognitive Heat. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20, sep 2017.

[4] Youssef Mohamed, Giulia Ballardini, Maria Teresa Parreira, Séverin Lemaignan, and Iolanda Leite. Automatic frustration detection using thermal imaging. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, pages 451–460, 2022.

[5] Eve A Edelstein and Eduardo Macagno. Form follows function: bridging neuroscience and architecture. In *Sustainable environmental design in architecture*, pages 27–41. Springer, 2012.

[6] Moaed A Abd, Iker Gonzalez, Mehrdad Nojoumian, and Erik D Engeberg. Trust, satisfaction and frustration measurements during human-robot interaction. In *Proceedings of the 30th Florida Conference on Recent Advances in RoboticsMay 11-12*, volume 2107, 2017.

[7] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in Psychology*, 6:931, jul 2015.

[8] Jonathan Lazar, Adam Jones, and Ben Shneiderman. Workplace user frustration with computers: An exploratory investigation of the causes and severity. *Behaviour and Information Technology*, 25(3):239–251, may 2006.

[9] Bernard Weiner. An Attributional Theory of Achievement Motivation and Emotion. *Psychological Review*, 92(4):548–573, oct 1985.

[10] Suzy Fox and Paul E. Spector. A model of work frustration-aggression. *Journal of Organizational Behavior*, 20(6):915–931, 1999.

[11] John Dollard, Neal E. Miller, Leonard W. Doob, O. H. Mowrer, and Robert R. Sears. *Frustration and aggression.* Yale University Press, oct 1939.

[12] Suhaib Aslam, Kim Gouweleeuw, Gijs Verhoeven, and Nynke Zwart. Classification of Disappointment and Frustration Elicited by Human-Computer Interaction: Towards Affective HCI. Number August, 2019.

[13] Selma Sabanovic, Marek P Michalowski, and Reid Simmons. Robots in the wild: Observing human-robot social interaction outside the lab. In *9th IEEE International Workshop on Advanced Motion Control, 2006.*, pages 596–601. IEEE, 2006.

[14] Andreas Kornmaaler Hansen, Juliane Nilsson, Elizabeth Ann Jochum, and Damith Herath. On the importance of posture and the interaction environment: Exploring agency, animacy and presence in the lab vs wild using mixed-methods. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 227–229, 2020.

[15] Alexandra Weidemann and Nele Rußwinkel. The Role of Frustration in Human–Robot Interaction – What Is Needed for a Successful Collaboration? *Frontiers in Psychology*, 12:707, mar 2021.

[16] Brandon Taylor, Anind Dey, Daniel Siewiorek, and Asim Smailagic. Using physiological sensors to detect levels of user frustration induced by system delays. In *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 517–528. Association for Computing Machinery, Inc, sep 2015.

[17] Nigel Bosch, Huili Chen, Sidney D'Mello, Ryan Baker, and Valerie Shute. Accuracy vs. availability heuristic in multimodal affect detection in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 267–274, 2015.

[18] Athanasios Psaltis, Kyriaki Kaza, Kiriakos Stefanidis, Spyridon Thermos, Konstantinos C. Apostolakis, Kosmas Dimitropoulos, and Petros Daras. Multimodal affective state recognition in serious games applications. In *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 435–439, 2016.

[19] Angeliki Fydanaki and Zeno Geradts. Evaluating OpenFace: an open-source automatic facial comparison algorithm for forensics. *https://doi.org/10.1080/20961790.2018.1523703*, 3(3):202–209, jul 2018.

[20] J. M. Lloyd. *Thermal Imaging Systems*. Springer US, Boston, MA, 1975.

[21] Thu Nguyen, Khang Tran, and Hung Nguyen. Towards Thermal Region of Interest for Human Emotion Estimation. In *Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering, KSE 2018*, pages 152–157. Institute of Electrical and Electronics Engineers Inc., dec 2018.

[22] Stephanos Ioannou, Vittorio Gallese, and Arcangelo Merla. Thermal infrared imaging in psychophysiology: Potentialities and limits. *Psychophysiology*, 51(10):951–963, oct 2014.

[23] Youngjun Cho, Nadia Bianchi-Berthouze, Manuel Oliveira, Catherine Holloway, and Simon Julier. Nose heat: exploring stress-induced nasal thermal variability through mobile thermal imaging. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 566–572. IEEE, 2019.

[24] Youngjun Cho, Simon J Julier, and Nadia Bianchi-Berthouze. Instant Stress: Detection of Perceived Mental Stress Through Smartphone Photoplethysmography and Thermal Imaging. *JMIR Ment Health*, 6(4):e10140, apr 2019.

[25] Veronika Engert, Arcangelo Merla, Joshua A Grant, Daniela Cardone, Anita Tusche, and Tania Singer. Exploring the use of thermal infrared imaging in human stress research. *PloS one*, 9(3):e90782, 2014.

[26] Hans JA Veltman and Wouter WK Vos. Facial temperature as a measure of mental workload. In *2005 International Symposium on Aviation Psychology*, page 777, 2005.

[27] Mihaela Sorostinean, François Ferland, and Adriana Tapus. Reliable stress measurement using face temperature variation with a thermal camera in human-robot interaction. In *IEEE-RAS International Conference on Humanoid Robots*, volume 2015-Decem, pages 14–19. IEEE Computer Society, dec 2015.

[28] Barry Kort, Rob Reilly, and Rosalind W Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proceedings IEEE international conference on advanced learning technologies*, pages 43–46. IEEE, 2001.

[29] Scotty D Craig, Sidney D'Mello, Amy Witherspoon, and Art Graesser. Emote aloud during learning with autotutor: Applying the facial action coding system to cognitive–affective states during learning. *Cognition and Emotion*, 22(5):777–788, 2008.

[30] Mohammed Hoque and Rosalind W. Picard. Acted vs. natural frustration and delight: Many people smile in natural frustration. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 354–359, 2011.

[31] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1-2):160–187, 2003.

[35] Vitaliy Kolodyazhniy, Sylvia D. Kreibig, James J. Gross, Walton T. Roth, and Frank H. Wilhelm. An affective computing approach to physiological emotion specificity: Toward subject-independent and

[32] Nicu Sebe, Michael S Lew, Ira Cohen, Ashutosh Garg, and Thomas S Huang. Emotion recognition using a cauchy naive bayes classifier. In *Object recognition supported by user interaction for service robots*, volume 1, pages 17–20. IEEE, 2002.

[33] Elias Vyzas and Rosalind W Picard. Offline and online recognition of emotion expression from physiological data. In *Workshop on Emotion-Based Agent Architectures at the Third International Conference on Autonomous Agents*, volume Technical, 1999.

[34] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, 2001.
stimulus-independent classification of film-induced emotions. *Psychophysiology*, 48(7):908–922, jul 2011.

[36] Paul Ekman. *Emotions revealed: recognizing faces and feelings to improve communication and emotional life*. 2003.

[37] Dawn T Robinson, Jody Clay-Warner, Christopher D Moore, Tiffani Everett, Alexander Watts, Traci N Tucker, and Chi Thai. Toward an unobtrusive measure of emotion during interaction: Thermal imaging techniques. In *Biosociology and neurosociology*. Emerald Group Publishing Limited, 2012.

[38] Irving A Cruz-Albarran, Juan P Benitez-Rangel, Roque A Osornio-Rios, and Luis A Morales-Hernandez. Human emotions detection based on a smart-thermal system of thermographic images. *Infrared Physics & Technology*, 81:250–261, 2017.

[39] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[40] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning, part iii—cross-validation and hyperparameter tuning, 2016.

[41] Ebrahim Babaei, Namrata Srivastava, Joshua Newn, Qiushi Zhou, Tilman Dingler, and Eduardo Velloso. *Faces of Focus: A Study on the Facial Cues of Attentional States*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2020.

[42] 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.

[43] SK D'Mello, SD Craig, B Gholson, S Franklin, R Picard, and AC Graesser. Integrating affect sensors into an intelligent tutoring system. In *Affective interactions: The computer in the affective loop. Proceedings of the 2005 International Conference on Intelligent User Interfaces*, pages 7–13, 2004.

[44] Bethany McDaniel, Sidney D'Mello, Brandon King, Patrick Chipman, Kristy Tapp, and Art Graesser. Facial features for affective state detection in learning environments. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

[45] Colin Puri, Leslie Olson, Ioannis Pavlidis, James Levine, and Justin Starren. Stresscam: Non-contact measurement of users' emotional states through thermal imaging. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, page 1725–1728, New York, NY, USA, 2005. Association for Computing Machinery.

[46] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.

[47] Christiane Goulart, Carlos Valadão, Denis Delisle-Rodriguez, Eliete Caldeira, and Teodiano Bastos. Emotion analysis in children through facial emissivity of infrared thermal imaging. *PloS one*, 14(3):e0212928, 2019.