Quantified Facial Expressiveness for Affective Behavior Analytics

Md Taufeeq Uddin
University of South Florida, Tampa, FL, US
mdtaufeeq@usf.edu

Shaun Canavan
University of South Florida, Tampa, FL, US
scanavan@usf.edu

Abstract

The quantified measurement of facial expressiveness is crucial to analyze human affective behavior at scale. Unfortunately, methods for expressiveness quantification at the video frame-level are largely unexplored, unlike the study of discrete expression. In this work, we propose an algorithm that quantifies facial expressiveness using a bounded, continuous expressiveness score using multimodal facial features, such as action units (AUs), landmarks, head pose, and gaze. The proposed algorithm more heavily weights AUs with high intensities and large temporal changes. The proposed algorithm can compute the expressiveness in terms of discrete expression, and can be used to perform tasks including facial behavior tracking and subjectivity quantification in context. Our results on benchmark datasets show the proposed algorithm is effective in terms of capturing temporal changes and expressiveness, measuring subjective differences in context, and extracting useful insight.

1. Introduction

Affective data analytics can be a powerful tool to explore expressions within context to discover underlying patterns and relationships between expressions and other variables of interest (e.g., EEG data [28]). It can be especially useful since there are two opposing theories about emotional expressions [45], namely the classical view of emotion, and the theory of constructed emotion. The classical view of emotion states that emotions are universal among humans, whereas the theory of constructed emotion states that emotions come from the complex dynamics of humans and context [5]. It has also been shown that expressiveness is subjective and happens at different frequencies and intensities [51]. Tools for analyzing expressions allow for insight into affective data and how it relates to each opposing theory.

While expressiveness has been extensively studied in psychology [1, 15, 20, 43], fewer works appear in affective computing. With the increase in large-scale emotion-based datasets [14, 54], the current manual approach to annotating expressiveness [35] is not scalable. An automated approach is needed to objectively, and quickly analyze facial images to facilitate further advances in affective computing, especially as the need for data grows with deep learning approaches to expression [50] and emotion [40] recognition.

The difficulties with manual annotation and the importance of emotional expressiveness [12] motivates us to quantify facial expressiveness within context (i.e. external stimuli). This can be useful for more objective scientific studies with affective data, as well as quantitatively evaluating the differences in expressiveness between people. As more context-aware affect models [31] are developed, a better understanding of context can also be useful. Considering this, we propose to analyze expressiveness as it is related to the context (i.e., external stimuli are used to elicit expressions, which occur at different intensities). We investigate two publicly available datasets, namely DISFA [39] and BioVid Pain [49] datasets. We find that context influences expressiveness and there is a subjective difference in the intensity and frequency of said expressiveness. The main contributions of this work are detailed below.

1. A quantified approach to the analysis of facial expressiveness is proposed (Fig. 1). It is bounded by a lower and upper limit of expressiveness, which allows us to more objectively compare different data.
2. Detailed analysis of the relationship between context and expressiveness is given on two publicly available datasets. Our results suggest that different context can impact the overall expressiveness of subjects. A Granger causality-based hypothesis between facial expressiveness and temporal context is also tested.
3. The subjective differences in expressiveness are demonstrated using the proposed, bounded, quantitative approach. We show that given the same context (i.e. the subjects are introduced to the same external stimuli), different subjects have different intensity and frequency of expressiveness.
4. We demonstrate how the proposed algorithm can be used to analyze, summarize, and interpret human affective behavior exploiting affect videos and relevant information to augment affective computer vision.

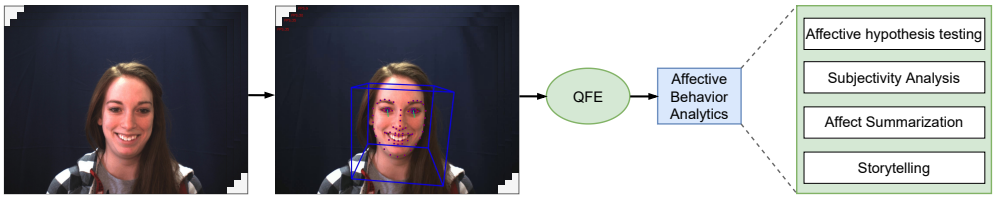


Figure 1: Workflow of the proposed method. Given affect videos, multimodal facial features are extracted to use as input to the proposed QFE algorithm. Then, the computed expressiveness score is used along with other modalities such as context to perform affective behavior analytics to perform varied tasks. These tasks include, but are not limited to, enhancing emotional signal analysis, affective hypotheses testing, subjective difference analysis, summarizing affect data, and tell engaging stories.

2. Related Works

Many works in psychology have studied different types of expressiveness including personal [51], family [24], and nonverbal [20]. Ogren and Johnson [42] found that the expressiveness of the primary caregiver of children strongly relates to their understanding of emotion. Ludwisowski et al. [38] investigated the relationship of gender and expressiveness, with a specific focus on how it can explain different gender interests. They found females were more expressive than males, which had a chain effect that impacted artistic interests. Self-report is generally considered accurate, however, subjects may not be truthful on them [21]. Although psychologists rely on self-report [4], having an automatic approach to analyze the emotional expressiveness of a subject would offer a fast, objective alternative.

Over the last few decades, researchers studied affect in numerous ways such as categorical (happy, sad) and dimensional (valence, arousal) [22]. Normally, data were annotated by subjects (self-report) or an observer. One of the limitations of the dimensional approach is that it provides comparatively generic information such as unpleasant to pleasant, which is often used in sentiment modeling [8]. On the other hand, categorical models use classes such as happy, surprise, or sad [37]. Also, many studies focus on the presence or absence of each class, not including the intensity of the expression [33]. That being said, there are recent works focused on these limitations. For example, Lin et al. [35, 36], and Lei et al. [32] measured the facial expressiveness at the video sequence level using human annotators. Although these results are encouraging, it does have the limitation that subjective human ratings need to be collected, which is time-consuming, and lack of expressiveness details. Note that the expressiveness is not uniform throughout the sequence. As pointed out by Gunes et al. [23], a single label (annotation) may not capture the complexity of expressions. Hence, we need methods that can measure expressiveness at a granular level in multiple dimensions (expressions are likely to be mixed [11] such as joy, happiness, celebration).

Uddin and Canavan [48] proposed TED to quantify facial expressiveness. While encouraging, there are some limitations to this approach that motivate our current work into quantified facial expressiveness. Their proposed approach can't measure dynamic changes (e.g. gaze) properly, and the quantification is unbounded and biased towards the number of action units, which makes the comparison of different expressions data infeasible. Our proposed algorithm extends state of the art by removing the need for human annotations and providing a fast and objective measure of affective expressiveness that can be used to compare multiple datasets. The proposed approach can be used on various types of expressiveness including but not limited to mixed, complex, and simultaneous.

3. Quantified Facial Expressiveness

3.1. Quantified Facial Expressiveness Algorithm

There are two major components of facial expressiveness: spatial (static) and temporal (dynamic). Spatial expressiveness is observed in a static video frame in a given moment in time. This expressiveness can be captured from the intensities of facial AUs given that AUs have well-defined meaning based on the classical view of emotion [17, 29]. AUs are also associated with individual expressiveness, personality, stimuli, and self-report [16]. Here, we compute a spatial, continuous expressiveness score for a given frame bounded by a lower and upper limit by

$$\sigma = \frac{\lambda}{n[\exp(1) - 1]} \sum_{i=1}^{n} \left[\exp\left(\frac{x_i}{x_{max}}\right) - 1 \right] \qquad (1)$$

where x_i is a vector containing the intensities of AUs of interest, n is the length of the vector, and x_{max} is the maximum possible intensity of the AUs. By convention, AUs are coded in between $[0, 5]$ where 0 indicates absence of the AU, and 5 indicate maximum activated AU intensity. The motivation behind Eqn. 1 is to more heavily weight the active AUs, while bounding the spatial expressiveness score in between $[0, \lambda]$, where λ is a constant multiplier. It is important to note that Eqn. 1 can only capture the spatial (static) expressiveness. Hence, to capture temporal expressiveness, other essential modalities such as facial landmarks, head pose, and eye gaze are exploited. These modalities don't have intensity as AUs do and are generally represented by coordinates in $2D/3D$ space, or orientation and rotation. Considering this, we track temporal expressiveness from

these modalities using the following set of equations. First, we measure the relative change by computing the velocity for consecutive frames (Eqn. 2).

$$\Delta v = \frac{\Delta x}{\Delta t} \tag{2}$$

Where Δx and Δt represent the change in corresponding values between two frames, and interval between the frames, respectively. As we want to more heavily weight the location where major change happens, and approximate the information in the neighboring frames, we approximate the temporal expressiveness for each modality using the Taylor series approximation of the following exponential function: $e^y - 1$, in our case, $y = \Delta v$ [1]. Note that $e^{\Delta v} - 1$ is bounded by $[0, 1.718]$ when $0 \leq \Delta v \leq 1$, which is useful to get a lower and upper bounded temporal expressiveness score. Hence, for a given modality, for each pair of points, we approximate the temporal expressiveness using Eqn. 3, where n and m are the length of facial feature vector, and the order to which the approximation is performed, respectively.

$$t_{exp} = \sum_{j=1}^{n} [\exp(\Delta v) - 1] = \sum_{j=1}^{n} \sum_{m=1}^{\infty} \frac{\Delta v^m}{m!} \tag{3}$$

It is then scaled to $[0, 1]$ (Eqn. 4), in which $\Delta_{max} = 1$ given the feature vectors are scaled between $[0, 1]$.

$$\delta = \frac{t_{\exp}}{n * [\exp(\Delta_{max}) - 1]} \tag{4}$$

From Eqns. 1 and 4, spatial expressiveness, σ, and temporal expressiveness, δ, are in between $[0, \lambda]$, and $[0, 1]$.

3.1.1 Combining spatial and temporal expressiveness

Approach 1**.** We hypothesize that σ is the main source of expressiveness following literature of the classical view of emotion [16], and δ is the auxiliary source of expressiveness. Hence, to obtain the quantified facial expressiveness (**QFE**) score (τ) treating σ as the essential source of expressiveness, we combine σ and δ by

$$\tau = \sigma * \Big[1 + \frac{1}{n_{mod}} \sum_{k=1}^{n_{mod}} \lambda_k \delta_k \Big]. \tag{5}$$

Here, λ_k represents the weight parameter for a given temporal modality and n_{mod} represents the number of modalities, which are needed to compute the weighted mean of the temporal modalities. Hence, τ represents the QFE score for a given face for a given moment in time. Notice that depending on the λ_k, we can have τ bounded in between $[0, n_b\lambda]$, where n_b is a scalar. For instance, from Eqn. 5, if we set $\lambda_k = 1$, then the QFE score is $0 \leq \tau \leq 2\lambda$.

Approach 2. σ and δ can be combined using the weighted combination with an additional adjustment term as offset, i.e. $\tau_{wc} = w_i\sigma + w_{i+1}\delta + \epsilon$, where w_i and w_{i+1} are the weights and ϵ is an adjustment term. There could be scenarios where this formulation could be relevant: i) both spatial and temporal modalities are equally important; ii) temporal expressiveness is more crucial than spatial expressiveness. For instance, in the case of student engagement, autism spectrum disorder, or driver behavior studies, eye gaze could be more relevant than other modalities including AUs [19, 47]. In these scenarios, τ_{wc} maybe more effective and can be computed putting more weight on the gaze.

Approach 3**.** Instead of using domain knowledge (approach 1) or manually weighing the modalities (approach 2), the expressiveness score τ can be estimated using a linear generative model with Gaussian latent variables [10]. Here, we feed all modalities to the generative model as factors to compute latent facial expressiveness variable (τ_{fa}).

We refer the reader to Fig. 5 for the QFE score distribution across these 3 approaches.

3.2. Affective Behavior Analytics

The quantification of facial expressiveness is essential as τ provides detailed information captured from both the spatial and temporal expressiveness. Using τ can help perform affective science and emotion AI research at scale, given that it has the potential to enhance data collection and hypothesis testing on large-scale datasets, while incorporating context. This has the potential to augment affective behavior analytics. To demonstrate use cases of the QFE in emotion research and affective computer vision, in this section, we describe two important affective computing tasks.

Granger Causality between Temporal Context and Affective Facial Expressions. In emotion research and affective computer vision, stimuli or context are used to elicit facial expressions on subjects. One natural question that arises is that for a given stimulus, *can we measure whether the stimulus elicited the facial expression?* In this work, we use the QFE score τ and temporal context to test the relationship between stimuli and facial expressions. We formulate the problem as follows: assuming τ and the stimulus are temporal variables, we use the Granger causality [44] test to evaluate whether stimulus Granger-causes the expressiveness. We represent this as $GC(c) \longrightarrow \tau$ where c, and GC are the stimulus, and Granger causality, respectively. We can say c Granger-causes τ if the historic values of stimulus c can predict the future value of the τ. If we find significant evidence of c Granger-causing τ, then we can conclude that stimulus is able to elicit facial expressiveness. This along with the proposed τ has the potential to explore the relationship between context and expressiveness at scale to augment emotion research, and evaluate users' responses to multimedia content. See Sec. 4 for more details.

Quantifying Subjectivity in Context. Affect is highly subjective [27, 30, 6] due to factors including, but not limited to, personality, gender, and culture. To develop auto-

[1] $e^y - 1 = e^{\Delta v} - 1 = \sum_{m=1}^{\infty} \frac{\Delta v^m}{m!} = \Delta v + \frac{\Delta v^2}{2!} + \frac{\Delta v^3}{3!} + \dots$

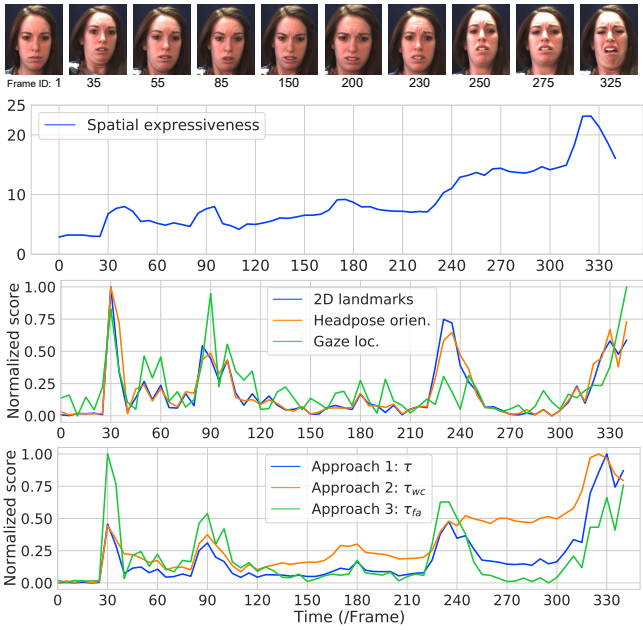


Figure 2: Sample sequence with computed quantified expressiveness scores. Top to bottom: frames from the sequence, the magnitude of the spatial expressiveness σ, temporal expressiveness: 2D landmarks, headpose orientation, gaze location and QFE scores: τ, τ_{wc}, and τ_{fa}. For visualization purposes, temporal expressiveness, and τ, τ_{wc}, and τ_{fa} are normalized in between $[0, 1]$. (Best viewed in color).

mated affect perception models, a sound understanding and quantitative analysis of subjectivity of facial expressiveness is required. In this work, we demonstrate how quantified facial expressiveness can be exploited to quantify the subjective difference among people. To do so, we first compute the QFE score τ for each subject in a given context, and then, we perform several statistical measurements to quantify the difference. See Sec. 4 for more details.

4. Experiments and Analysis

As the goal of this work is to quantify facial expressiveness in a given moment in time and demonstrate use cases, we first computed the QFE score using the DISFA and BioVid pain datasets. Then, we experimented with two downstream tasks, namely Granger causality analysis among modalities (e.g. context, QFE score, ground truth), and subjectivity quantification. These are two important tasks that are essential in emotion research and applied affective computer vision given the constructed theory on emotion and its relationship to context [45, 7] and face as sensing, as well as human subjectivity [27, 30].

Data preparation. To track and extract the facial features such as landmarks (LM), head pose (HP), eye gaze (G), and facial AU intensities, we used OpenFace [2], which is a publicly available facial behavior analysis tool. Since LM, HP, and G represent spatial and depth information of the face and do not have expression intensity in the same way we have for AUs, we normalized LM, HP, and G in between $[0, 1]$ using min-max normalization [25].

Table 1: Descriptive summary of facial expressiveness scores τ for given pain level on BioVid pain dataset. Here, $25\%P.$ and $75\%P.$ denote 25^{th} and 75^{th} percentiles. We can see that with the change in the pain level (PL), τ is not changing much. Especially PL $1, 2$, and 3, which are very similar in terms of expressiveness. This questions the efficacy of heat as the stimuli to elicit facial pain expressions. The τ summary also partially explain the failure to pass the GC test in Table 3. This summary is also aligned with the findings of Werner et al. [52] as they pointed out the weak and low facial pain response for PL 1 and 2.

PL	Mean	SD	Min.	25% P.	Med.	75% P.	Max.
1	8.2	8.2	0	2.8	5.8	10.8	139.8
2	8.3	8.4	0	2.8	5.8	11.1	144
3	8.6	9.2	0	2.8	5.7	11.4	137.3
4	10.0	10.9	0	3.2	7.0	13.0	146.3

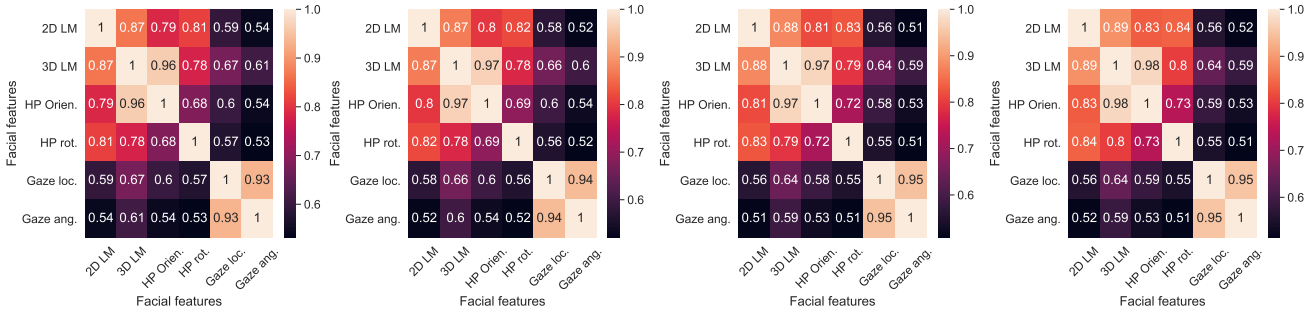
4.1. Datasets

DISFA dataset [39] is a publicly available spontaneous facial expression dataset which contains 27 subjects (12 females, and 15 males) aged in between $[18, 50]$ and ethnically 3 Asian, 1 Black, 21 Caucasian, and 2 Hispanic subjects. The dataset contains frontal face images and action unit (AU) [16] annotations at frame level, by expert annotators. The dataset contains 27 videos comprising $130,000$ images. A video comprising of 9 segments with different types of content was used as the stimuli to elicit the expression. The stimuli video and corresponding frontal face videos of the subjects are each 242 seconds long.

BioVid pain dataset [49] contains 90 subjects performing pain and other expressions. The dataset is balanced in terms of gender. There are three age groups in the dataset in the age range of $[18-35], [36-50], [51-65]$ years old. In this work, we used the raw data which is part C in the data portion and contains 87 subjects. Continuous heat (temperature) was used as a stimulus to elicit pain expression, which was self-calibrated by the subjects. The length of each session is approximately 25 minutes, and in total, in Part C, is comprised of approximately 3.26 million images. Aside from the frontal face videos and temperature, the dataset contains pain labels in between $[0, 4]$, where 0 means no pain, and 4 means maximum pain.

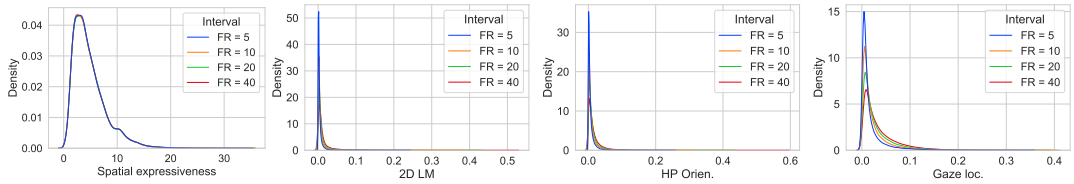
4.2. Quantified Facial Expressiveness

To compute the quantified facial expressiveness (QFE) score, we computed the spatial expressiveness σ using Eqn. 1. Note that σ can be computed using all available AUs, or a subset of AUs depending on the context and task. In this work, we computed the **overall** spatial (static) expressiveness of the human face for a given video frame using AUs:



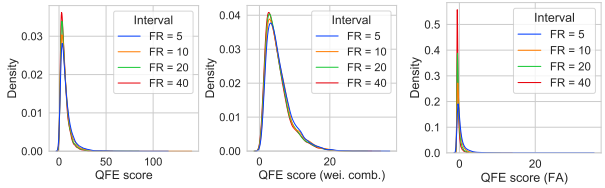
(a) FR = 5 (b) FR = 10 (c) FR = 20 (d) FR = 40

Figure 3: Association among candidate temporal facial features. (Best viewed in color and zoomed in).



(a) σ (b) δ_{lm} (c) δ_{hp} (d) δ_g

Figure 4: Spatial and temporal expressiveness distribution (Best viewed in color and zoomed in).



(a) Approach 1 (b) Approach 2 (c) Approach 3

Figure 5: QFE score distribution from the 3 approaches.

$1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28$, and 45 on both datasets. We also computed spatial expressiveness in the context of **pain** using pain-related AUs [53]: $4, 6, 9, 10, 25$ on BioVid. We set $\lambda = 100$ to have σ in between $[0, 100]$. We then computed the temporal expressiveness using landmarks, head pose, and eye gaze using Eqn. 3, where $\infty = 20$. We found little change with $\infty > 20$.

Ablation study on frame rate, and temporal modalities (features). This ablation study is performed by sampling data from both DISFA and BioVid pain datasets. We performed an ablation study on impact of frame rate (FR) (interval at which we picked two frames to compute the difference (i.e. Δ_x in Eqn. 2)), and facial features used to capture the temporal expressiveness. More precisely, we experimented with $2D$ and $3D$ landmarks (**LM**), orientation and rotation of headpose (**HP**), and location and angle of eye gaze (**G**), while setting FR to $5, 10, 20$, and 40, respectively. We measured the association among the modalities using Spearman's rank correlation coefficient (**SRCC**) [9], and the obtained results are reported in Fig. 3. We found that 2D LM, 3D LM, and HP orientation and rotation are moderately positive to strongly correlated. We also found that gaze location and angle are strongly correlated.

Figure 6: Example video segments from DISFA showing subjective differences with same context. Top to bottom: peak frames from stimulus video, frames from subjects SN001, SN002, SN003, and QFE scores τ computed for each subject, respectively. (Best viewed in color).

To keep a balance between modalities and to reduce the redundant computation, we selected $2D$ LM, HP orientation, and gaze location to capture the temporal expressiveness in our algorithm. Fig. 4 depicts the distribution (Estimated using kernel density estimation (KDE) method) of σ, 2D LM, HP orientation, and G location setting FR to $5, 10, 20$, and 40 in which we can observe that aside from the distribution of gaze, the shape of distribution did not change much with the change in FR. Hence, in our downstream tasks, we set FR to 5 to compute the temporal expressiveness.

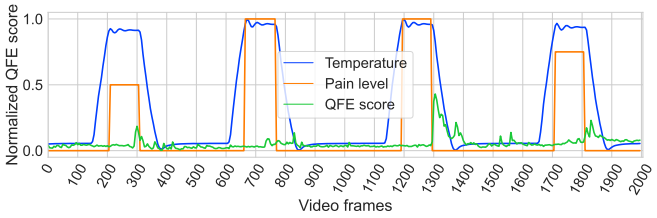


Figure 7: Association among temperature, QFE score (τ_{pain}), and pain level (BioVid). For visualization purposes, temperature, τ_{pain}, and pain level are normalized to $[0, 1]$. We excluded face images from the visualization following data usage policy of bioVid pain dataset [49].

The expressiveness score τ is computed using the three approaches. In approach 1, we computed τ assuming σ is the major source of expressiveness, and δ as the minor source of expressiveness; hence, we set λ_k in Eqn. 5 to $100, 100$, and 50 for computing δ using 2D LM (δ_{lm}), using HP orientation (δ_{hp}), and using G location (δ_g), respectively. In approach 2, we computed the weighted combination of σ, δ_{lm}, δ_{hp}, and δ_g in which we set each weight $w = 100$, and $\epsilon = 0$. Finally, in approach 3, we used a latent variable model to estimate the τ feeding all available facial features. In case of approach 3, we performed Bartlett's test to evaluate the factorability of the input facial features, and Kaiser-Meyer-Olkin (KMO) test to evaluate suitability of data for factor analysis on the input features. The data passed the factorability test with $p - value < 0.001$, mean (standard deviation (**SD**)) $KMO = 0.727 \pm 0.02$, and $p - value < 0.001$, mean (SD) $KMO = 0.744 \pm 0.02$ across four FR settings for DISFA dataset and BioVid pain dataset, respectively. An example on computed expressiveness score is shown in Fig. 2 in which we observe that the proposed method can capture the facial expressiveness, as all three approaches were able to measure the overall expressiveness of the face. In our next set of experiments, we used the τ computed using **approach** 1 since we focus on affective experience/responses for those tasks.

Evaluation via human annotators: To measure the correctness of the QFE algorithm, we collected ratings from three annotators (2 males, 1 female), that were given instructions on rating beforehand. We used a questionnaire with three questions. $Q1$: 'Did the algorithm capture the expressivity? (response: yes/no)'; $Q2$: Rate the expressiveness score computed by the algorithm in between $[1, 5]$, where $1 =$ poor, $2 =$ weak, $3 =$ marginal, $4 =$ very good, $5 =$ excellent. We also asked the raters to provide their confidence on assessment ($Q3$), in between $[0, 100]$, (uncertain to certain). We collected ratings for DISFA and BioVid datasets. In the case of DISFA, the entire sequence (242 seconds) was rated. In the case of BioVid, we selected 200 random samples (5 seconds long) sequences so that we can observe how QFE performed for both short and long sequences. The obtained rating summary is highlighted in Table 2. It can also be seen in Fig. 8 that most subjects responded to the stimuli, however, some subjects did not respond to 'surprise', 'disgust', and 'fear'. Finally, *even though most subjects responded to stimuli, the level and duration of responsiveness (facial expressiveness τ) were variable and diverse (Sec. 4.3.2).* These results are encouraging, as they show that QFE algorithm captured the expressiveness since the average assessments for both datasets falling in-between 'very good' and 'excellent.

Table 2: QFE Algorithm evaluation via human annotators.

Datasets	Q1	Q2	Confidence
DISFA	1.0	4.6± 0.46	96.7 ± 4.1
BioVid	0.995 ± 0.07	4.58 ± 0.58	96.4 ± 4.1

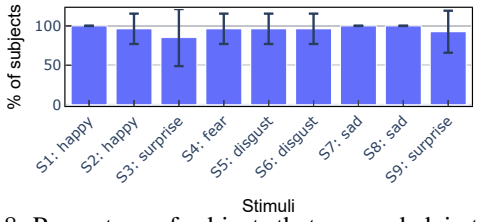


Figure 8: Percentage of subjects that responded, in terms of facial expressivity, to each stimulus in DISFA. Here, human annotators observed both stimuli video and QFE scores in parallel to identify whether a given stimulus video segment caused facial expressiveness on subject.

Qualitative comparison with related work. Note that previous works mostly focused on sequence-level expressiveness and relied on subjective opinions from annotators and/or coders. In contrast, this work measures the expressiveness at video frame level using domain knowledge from affective computing. Also, to the best of our knowledge, there are no public visual affect datasets that were annotated at the video frame level for expressiveness. Considering this, a quantitative comparison with previous works is infeasible. Here, however, we discuss a qualitative comparison of the work from Uddin and Canavan [48], which also computed an expressiveness score at video frame level. While this work is similar to ours, there are some differences including no bounds, the results are skewed towards the total number of AUs, and it lacks the ability to compare among different categories of expressions (e.g. happy, pain). The proposed algorithm addresses these limitations by providing an unbiased (toward # of AUs) lower and upper bounded expressiveness score. This is essential for a measurement scale, so we can perceive the relative importance of the expressiveness of a given frame.

4.3. Affective Behavior Analytics

4.3.1 Granger causality between context and ground truth, and between context and QFE score

We hypothesize that context will elicit facial expression since during data collection, in the BioVid pain dataset, temperature was used to elicit pain experience. We formu-

Table 3: Percentage (%) of video segments for which temperature Granger-caused ground-truth pain level (PL), and facial pain expressiveness (τ_{pain}). We set the significance level α to 0.05. **PVSP** = percentage of video segment passed.

Lag		**GC (temperature) $\longrightarrow$ PL**					**GC (temperature) $\longrightarrow$ τ_{pain}**				
Time (sec.)	# of frames	PVSP LR test χ^2	PVSP params F test	PVSP SSR χ^2	PVSP SSR F test F	ALL	PVSP LR test χ^2	PVSP params F test	PVSP SSR χ^2	PVSP SSR F test F	ALL
1	5	50.8	47.8	52.8	47.8	47.8	11.0	10.0	11.0	10.0	10.0
2	10	87.7	82.8	89.3	82.8	82.8	16.0	14.0	17.0	14.0	14.0
5	25	100.0	100.0	100.0	100.0	100.0	33.0	21.0	39.0	21.0	21.0
7	35	100.0	100.0	100.0	100.0	100.0	46.0	23.0	57.0	23.0	23.0
10	50	100.0	99.4	100.0	99.4	99.4	67.0	15.0	80.0	15.0	15.0

late this as a Granger causality (GC) test in which we use the temperature to test whether temperature Granger causes facial expressiveness (τ). We also tested the hypothesis that temperature Granger-causes the ground truth pain level. Note that temperature, pain level (**PL**), and QFE score τ for pain expression are modeled as temporal variables (Fig. 7).

Data preparation. To test the hypotheses, we extracted one-minute-long video segments from the BioVid pain dataset, resulting in 1740 video segments from 87 subjects. Then, we computed the QFE score for pain expression (τ_{pain}) using the AUs that are associated with the pain expression (AUs: $4, 6, 9, 10, 25$) [53]. For each temporal variable, to make the variable stationery, we computed the difference between the consecutive values, and then, we performed an Augmented Dickey-Fuller (AD-Fuller) test [55] to check whether the temporal variables are stationary or not, and found that all three variables passed the test. Then, we performed: $GC(temperature) \longrightarrow PL$ (i.e. temperature Granger-causes ground truth pain level); $GC(temperature) \longrightarrow \tau_{pain}$ (i.e. temperature Granger-causes facial pain expressiveness). We also performed an ablation study on the temporal history of the temperature using a lag ranged in between $[1, 10]$ seconds with an interval of 1 second. For fair evaluation, we performed four different statistical tests: likelihood-ratio (**LR**) χ^2 test, residual sum of squares (**SSR**) based χ^2 test, parameters (**params**) F test, and SSR based F test. We reported the percentage of video segments that passed each test separately and the percentage of segments that passed all four tests (**ALL**). Here, we set the significance level α to 0.05.

Table 3 highlights the percentages of the video segments that passed the tests. We can infer from Table 3 that in case of ALL, temperature Granger-caused the ground truth pain level (PL) in between $[47\%, 100\%]$ of the video segments, while temperature Granger-caused the facial expressiveness in between $[10\%, 23\%]$ of the video segments. A lag of temperature in between $[5, 10]$ seconds was useful to predict the pain level and facial pain experience, which indicates applying heat for a long time may induce a painful expression. As the temperature was self-calibrated by subjects for their own pain tolerance level, the strong predictive power of temperature towards PL is reasonable.

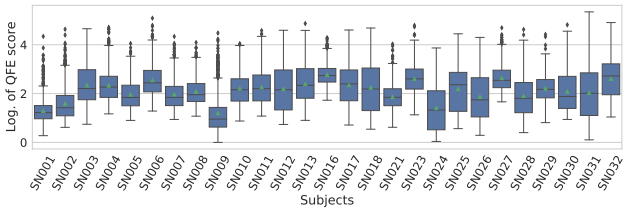
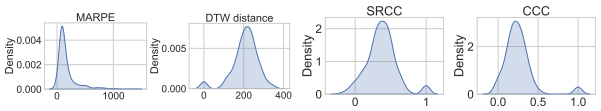


Figure 9: Facial expressiveness distribution across subjects.

Per our hypothesis, we should observe high facial pain expressiveness when the temperature is high, however, based on the summary of τ_{pain} in Table 1 and the highlighted results in Table 3, we did not observe that in the collected affect. In Fig. 7, we can see there is a strong relationship between pain level and temperature. However, even though we expect to observe facial pain expression with a change of temperature, we rarely observed that in this sequence. Considering this, analyzing expression on this dataset may not be reliable as it will give less insight into the pain level. Our results suggest that analyzing temperature could be a better solution towards perceiving pain (i.e. context is needed). Alternatively, this could be explained by inappropriate affect [26], where the subject's expression does not match the scenario. In our experiments, this would mean the subjects felt pain due to the temperature, however, they did not show a painful facial expression which can be validated by the construction theory of emotion [6].

4.3.2 Subjective Difference Analysis

We also conducted experiments to quantify subjectivity in terms of expression in context using DISFA. An example subjective difference is shown in Fig. 6 from which we can observe that SN002 is more expressive than SN001, and SN003 is more expressive than SN002. The subject-specific distribution of the natural logarithm of the computed overall expressiveness score τ on DISFA is shown in Fig. 9. Based on the individual expressiveness distribution, we can say that people are quite different in terms of expressing themselves even though the context was the same.



(a) MARPE (b) DTW dist. (c) SRCC (d) CCC

Figure 10: Distribution of subjective differences.

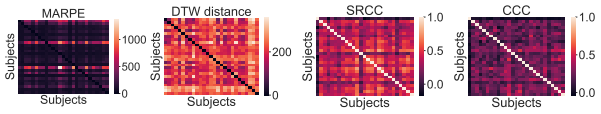
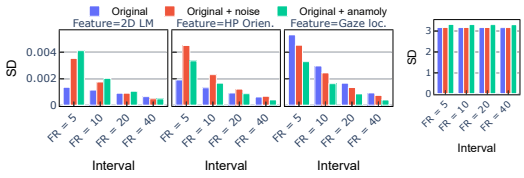


Figure 11: Heatmap representation of the quantified subjective differences by cross-referencing subjects in DISFA dataset. Here, *higher MARPE, higher DTW distance, lower SRCC, lower CCC indicate higher subjective difference*.

We measured subjective differences using four metrics: mean absolute-relative percentage error (**MARPE**), dynamic time warping (**DTW**) distance [41], SRCC, and concordence correlation coefficient (**CCC**) [34]. MARPE is defined as $MARPE = \frac{1}{n_f}\sum_{i=1}^{n_f}\left|\frac{x-y}{x}\right| * 100$, where x and y are QFE scores computed from two subjects, and n_f is the number of frames. We computed MARPE, DTW distance, SRCC, and CCC across all subjects for all combinations. Then, using the KDE [46] method, we estimated the distribution of the quantified subjectivity for each metric (Fig. 10). We also computed the mean and standard deviation (SD) of each metric, and obtained MARPE $= 122.4\% \pm 130\%$, DTW distance $= 93.13 \pm 35.24$, SRCC $= 0.43 \pm 0.22$, and CCC $= 0.35 \pm 0.2$. Here, high SD indicates high subject variability in terms of expressiveness.

In Fig. 11, a cross-reference among subjects in terms of MARPE, DTW distance, SRCC, and CCC is shown. From the MARPE, we can infer that some subjects were more expressive than others, with a high margin. From SRCC and CCC, we can say that some subjects had high similarity of expressiveness compared to the others. Notice that CCC score is comparatively lower which can be explained, in part, as CCC looks for consistency in addition to the similarity in temporal sequences. It is important to evaluate these different metrics, as they conveyed different information (see Fig. 11). Based on our observation, subjects were not only different in terms of expressiveness but there were also differences in terms of lag and delay. To be precise, some subjects begin their expressions earlier, and had a longer duration, while others did not. For example, subject SN001, in DISFA (Fig. 6), was noticeably different from the rest of the subjects. When frame IDs were in between $[1000, 1100]$, subject SN002 was likely to be surprised and shocked while subject SN003 was likely to be confused, and subject SN001 was likely to be neutral in terms of expressions. Hence, we measured statistical significance along with the SRCC, and we found that SN001 was correlated with negative to random chance to subjects SN006, SN023, SN021, SN007, SN009, SN010, SN011, SN024, SN012,



(a) δ (b) σ

Figure 12: Simulating influence of noise and anomaly incorporated from automated feature extraction models. (SD = standard deviation).

SN025, SN026, SN027, SN028, SN030, and SN031. In addition, SN029 was significantly different from the SN005 and SN016. The rest of the subjects (25) showed moderately positive to strong similarity ($p - value < 0.05$), indicating that context (stimuli) was effective at inducing affective facial expressions (Figs. 8 and 10c). To the best of our knowledge, this is the first work to show these findings. We encourage the use of them as a baseline for the quantification of subjectivity of facial expressions in DISFA.

5. Discussion, Limitations and Future Work

An interesting direction, of this work, is the incorporation of dimensional models of expressions (e.g., valence, arousal) [3, 11]. This could give us a better view of expressiveness. Along with this, while OpenFace was used for feature extraction, the proposed approach is not limited to this, as other methods can be used such as AFAR [18]. An investigation into which automated tool is best could be beneficial to the field, as automated prediction of features could introduce noise and anomalies into QFE. To evaluate this, we also simulated the influence of the presence of noise and anomalies on δ and σ. More formally, $D_{ns} = D * (1 + ns); ns \longleftarrow random(0, 0.05)$; where D is the original data, and ns is the noise generated from normal distribution, and D_{ns} is generated noisy data (original + noise). To simulate anomalies, $D_{0.02a} = D_{0.02} * a; a \longleftarrow random(0, 2)$; we made only 2% of the sample anomalous, and replaced those 2% original samples with $D_{0.02a}$ to get anomalous data (original + anomaly) D_a. As can be seen in Fig. 12, noise and anomalies negatively influence the QFE scores. A possible way to mitigate this limitation is preprocessing the extracted features before computing QFE. For instance, in case of σ, the influence of outliers can be reduced using the fact that $0 \leq AU_{intensities} \leq 5$. To handle noise and outliers, we can leverage the confidence of the feature extraction models to deal with poorly extracted features. For instance, OpenFace outputs face detection probabilities, which can be used to pre-process data and mitigate noise and anomalies. Along with this, anomaly detection techniques can be used [13].

Acknowledgment

We thank Nasimul Hasan and Liza Jivnani for providing human ratings. We also thank reviewers for their valuable feedback.

References

[1] Jack Balswick and Christine Proctor Avertt. Differences in expressiveness: Gender, interpersonal orientation, and perceived parental expressiveness as contributing factors. *Journal of Marriage and the Family*, pages 121–127, 1977.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[3] Lisa Feldman Barrett. Discrete emotions or dimensions? the role of valence focus and arousal focus. *Cognition & Emotion*, 12(4):579–599, 1998.

[4] Lisa Feldman Barrett. Feelings or words? understanding the content in self-report ratings of experienced emotion. *Journal of personality and social psychology*, 87(2):266, 2004.

[5] Lisa Feldman Barrett. Functionalism cannot save the classical view of emotion. *Social Cognitive and Affective Neuroscience*, 12(1):34–36, 2017.

[6] Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[7] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019.

[8] Valentin Barriere and Alexandra Balahur. Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation. *arXiv preprint arXiv:2010.03486*, 2020.

[9] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[10] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[11] Christopher Blank, Shaila Zaman, Amanveer Wesley, Panagiotis Tsiamyrtzis, Dennis R Da Cunha Silva, Ricardo Gutierrez-Osuna, Gloria Mark, and Ioannis Pavlidis. Emotional footprints of email interruptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[12] Jude Cassidy, Ross D Parke, Laura Butkovsky, and Julia M Braungart. Family-peer connections: The roles of emotional expressiveness within the family and children's understanding of emotions. *Child development*, 63(3):603–618, 1992.

[13] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[14] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4dfab: A large scale 4d data for facial expression analysis and biometric apps. In *CVPR*, 2018.

[15] Julie C Dunsmore and Amy G Halberstadt. How does family emotional expressiveness affect children's schemas? 1997.

[16] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[17] 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.

[18] Itir Onal Ertugrul, László A Jeni, Wanqiao Ding, and Jeffrey F Cohn. Afar: a deep learning based tool for automated facial affect recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–1. IEEE, 2019.

[19] Diego Fabiano, Shaun Canavan, Heather Agazzi, Saurabh Hinduja, and Dmitry Goldgof. Gaze-based classification of autism spectrum disorder. *Pattern Recognition Letters*, 135:204–212, 2020.

[20] Howard S Friedman, Louise M Prince, Ronald E Riggio, and M Robin DiMatteo. Understanding and assessing nonverbal expressiveness: The affective communication test. *Journal of personality and social psychology*, 39(2), 1980.

[21] Carolina Fuentes, Valeria Herskovic, Iyubanit Rodríguez, Carmen Gerea, Maira Marques, and Pedro O Rossel. A systematic literature review about technologies for self-reporting emotional information. *Journal of Ambient Intelligence and Humanized Computing*, 8(4):593–606, 2017.

[22] Didier Grandjean, David Sander, and Klaus R Scherer. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition*, 17(2):484–495, 2008.

[23] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 827–834. IEEE, 2011.

[24] Amy G Halberstadt, Valerie W Crisp, and Kimberly L Eaton. Family expressiveness: A retrospective and new directions for research. 1999.

[25] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, 5(4):83–124, 2011.

[26] Arthur Harris and Maryse Metcalfe. Inappropriate affect. *Journal of neurology, neurosurgery, and psychiatry*, 19(4):308, 1956.

[27] Jeff Kochan. Subjectivity and emotion in scientific research. *Studies in History and Philosophy of Science Part A*, 44(3):354–362, 2013.

[28] 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.

[29] Ann M Kring et al. The facial exp coding system (faces): Dev, validation, and utility. *Psych assessment*, 19(2):210, 2007.

[30] Joseph E LeDoux and Stefan G Hofmann. The subjective experience of emotion: a fearful view. *Current Opinion in Behavioral Sciences*, 19:67–72, 2018.

[31] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition net-

works. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10143–10152, 2019.

[32] Su Lei, Kalin Stefanov, and Jonathan Gratch. Emotion or expressivity? an automated analysis of nonverbal perception in a social dilemma. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 770–777. IEEE Computer Society, 2020.

[33] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020.

[34] Jason JZ Liao and Jerry W Lewis. A note on concordance correlation coefficient. *PDA journal of pharmaceutical science and tech*, 54(1):23–26, 2000.

[35] Victoria Lin, Jeffrey M Girard, and Louis-Philippe Morency. Context-dependent models for predicting and characterizing facial expressiveness. *arXiv preprint arXiv:1912.04523*, 2019.

[36] Victoria Lin, Jeffrey M Girard, Michael A Sayette, and Louis-Philippe Morency. Toward multimodal modeling of emotional expressiveness. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 548–557, 2020.

[37] Ping Liu, Yuewei Lin, Zibo Meng, Lu Lu, Weihong Deng, Joey Tianyi Zhou, and Yi Yang. Point adversarial self-mining: A simple method for facial expression recognition. *IEEE Transactions on Cybernetics*, 2021.

[38] Wyndolyn MA Ludwikowski, Patrick I Armstrong, and Daniel G Lannin. Explaining gender differences in interests: The roles of instrumentality and expressiveness. *Journal of Career Assessment*, 26(2):240–257, 2018.

[39] 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.

[40] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243, 2020.

[41] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[42] Marissa Ogren and Scott P Johnson. Primary caregiver emotion expressiveness relates to toddler emotion understand. *Infant Behavior and Dev*, 2021.

[43] William Roberts and Janet Strayer. Empathy, emotional expressiveness, and prosocial behavior. *Child development*, 67(2):449–470, 1996.

[44] Anil Seth. Granger causality. *Scholarpedia*, 2(7):1667, 2007.

[45] Erika H Siegel, Molly K Sands, Wim Van den Noortgate, Paul Condon, Yale Chang, Jennifer Dy, Karen S Quigley, and Lisa Feldman Barrett. Emotion fingerprints or emotion populations? a meta-analytic investigation of autonomic features of emotion categories. *Psychological bulletin*, 144(4):343, 2018.

[46] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[47] Shivam Srivastava, Saandeep Aathreya SIdhapur Lakshminarayan, Saurabh Hinduja, Sk Rahatul Jannat, Hamza Elhamdadi, and Shaun Canavan. Recognizing emotion in the wild using multimodal data. In *ICMI*, pages 849–857, 2020.

[48] Md Taufeeq Uddin and Shaun Canavan. Quantified facial temporal-expressiveness dynamics for affect analysis. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3955–3962. IEEE, 2021.

[49] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *CYBCO*, pages 128–131. IEEE, 2013.

[50] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[51] Alan S Waterman. Personal expressiveness: Philosophical and psychological foundations. *The Journal of Mind and Behavior*, pages 47–73, 1990.

[52] Philipp Werner, Ayoub Al-Hamadi, and Steffen Walter. Analysis of facial expressiveness during experimentally induced heat pain. In *ACIIW*. IEEE, 2017.

[53] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind Picard. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, 2019.

[54] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, pages 3438–3446, 2016.

[55] W Zhu. Augmented dickey-fuller test. `http://www.ams.sunysb.edu/~zhu/ams586/UnitRoot_ADF.pdf`, 2021.