2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)

Multimodal Behavior Analysis and Impact of Culture on Affect

Tara Nourivandi
Computer Science and Engineering
University of South Florida
Tampa, USA
taranourivandi@usf.edu

Saandeep Aathreya
Computer Science and Engineering
University of South Florida
Tampa, USA
saandeepaath@usf.edu

Shaun Canavan
Computer Science and Engineering
University of South Florida
Tampa, USA
scanavan@usf.edu

***Abstract*—It has been shown that emotions are learned in a cultural way and expressions are often used to help convey these emotional states. Considering this, in this work, we investigate multimodal cultural behavior differences across 6 different cultures. More specifically, we investigate head pose, action units, and facial landmarks in British, Chinese, German, Greek, Hungarian, and Serbian cultures. Along with this, we also investigate the differences along valence and arousal dimensions for these cultures. To conduct this investigation, we evaluate the SEWA multimodal and multi-cultural dataset. We find varying differences exist that are impacted by culture, context, and modality. Based on these findings, we also perform context classification that takes into account these differences in culture. We show that incorporating culture into our pipeline improves classification performance.**

***Index Terms*—Affect, Culture, Multimodal**

I. Introduction

Emotion is a well-studied concept in psychology and human behavior research [1]–[3]. Along with this, it has also been shown that emotion is social and adaptive, where culture also impacts said emotions [4]. To communicate these emotions, humans often use many cues such as facial expressions, which are a powerful, natural, and universal tool for this communication [5]. Ekman stated that emotion is fundamentally generalized among all cultures [6], however, it has been shown that emotions are influenced both by biological and environmental factors [7]. A person encodes and decodes inner emotional states to and from facial expressions. This process is not culturally universal, meaning culture influences how individuals encode and decode inner emotion to and from facial expressions [8]. In recent decades, scientists (i.e., affective computing) have tried to map facial expressions and other physiological signals to emotional states of people [9]–[12]. These studies have applications in pain recognition [13], building sociable robots and socially aware systems [14], helping humans detect some mental disorders such as autism [15], and other applications to elevate quality of life and well-being. [16].

As these applications are more readily woven into the daily lives of people worldwide, we need to have fair and reliable systems for all individuals [17]. To this end, we need to minimize the impact of biases that make these systems work in favor of specific groups of people over others [18]. A major step in this direction is to identify biases and their roots. One group of bias in these applications rises from the inherent differences in human behaviors. It has been shown that it is difficult for current systems to generalize to every individual [19], [20]. This can be explained, in part, by cultural differences in subjects [21]. While cultural biases in affective computing have received attention in recent years, research has primarily focused on facial appearance variations across ethnicities rather than cultural background of subjects [22]–[24]. Considering this, the proposed work expands this understanding by demonstrating that cultural differences extend beyond facial features. We show that behavioral variations associated with culture also need careful consideration during intelligent system design and model training. Facial and head pose data are evaluated from 6 different cultures from the SEWA [25] dataset (British, Chinese, German, Greek, Hungarian, and Serbian). Facial landmarks, head movements, facial action units [26], and arousal and valence levels are evaluated. Transformer-based models are trained on different cultures to learn how culture impacts classification of context. More specifically, the contributions of our work are 3-fold and can be summarized by the following.

1) Multimodal analysis of cultural differences is presented. Facial landmarks, action units, and head pose are evaluated. Along with this, differences in valence and arousal are also investigated across cultures.
2) Impact of culture on classification of context is detailed. Three experiments are conducted to evaluate this: 1) Culture-independent classification; 2) Subject-independent classification and 3) Subject-independent classification within a culture. Details on these experiments can be found in Section V.
3) To aid with classification, we propose a new temporal sampling module to adaptively sample most informative frames. This sampling leverages the facial features utilized for analysis. We demonstrate benefits of this strategy over random temporal sampling.

The rest of this paper is structured as follows: Section II details related work in psychology and neuroscience, and affective computing. Section III gives background information on culturally diverse datasets. Section IV details the multimodal behavior analysis based on culture. Section V describes the experimental design and results for culture-based classification of context. Finally, Sections VI and VII give a conclusion and ethical impact statement, respectively.

979-8-3315-1643-7/24$31.00©2024 IEEE

II. Related Work

A. Cultural Research in Psychology and Neuroscience

Cultural neuroscience research is revealing growing evidence of brain activity patterns that are specific to different cultures [27]. Lim [7] showed that cultural differences exist in emotional arousal levels. More specifically Lim showed that in Western or individualist cultures, high-arousal emotions are more promoted and experienced than low-arousal emotions. In Eastern or collectivist cultures individuals prefer, value, and experience low-arousal emotions. Along with this, facial expressions are a common approach to explicitly show our inner emotion to other individuals. Ekman and Friesen proposed display rules as one aspect of production and interpretation of facial expressions [6], [28]. As an example, in some cultures showing displeasure in the workplace is not accepted. Matsumoto et al. [29] showed that these rules vary among different ethnic groups in the United States. In addition to display rules, different decoding rules influence the cross-culture variability of facial expression interpretation [30], [31]. For example, in some cultures one might avoid attributing negative emotions to other culture members to increase social harmony [32]. Daily et al. [33] have conducted two studies confirming that culture is a factor in how we receive and interpret facial expressions. First, they asked Japanese and American participants to interpret facial expressions and participants were better at classifying expressions within their own group. Second, they trained a model on Japanese and American images with posed expressions and the results show that the model is also better at classifying expressions when it is trained on the same group.

B. Cultural Research in Affective Computing

While less work investigates culture in affective computing, there are some encouraging works in this area. Han et. al [34] proposed an approach to mitigate forgetting in conventional machine learning algorithms. The proposed approach was tested on cross-culture audio-visual modalities. However, in this study, they have used a part of SEWA dataset (German subjects) [25] and the RECOLA dataset [35] in training and validation of their models. It is important to note that the data collection and cultures are not the same in these two datasets. Along with this, the focus is on the proposed algorithm rather than the impression of culture. Rudovic et al. [36] proposed a deep learning-based approach to automate the engagement estimation of children with autism from Asian and European cultural backgrounds. They proposed CultureNet that leverages multi-cultural data when presented. Performing within- and cross-culture evaluations in this study has shown that due to the large differences in the distribution of engagement levels in two cultures, the model cannot successfully generalize to the other culture. The authors name the underlying cultural differences, some bias in their data sample, differences in facial physiognomy, and dynamics of facial expression in cultures as possible source of the downgrade in cross-culture estimation performance. This work is the main motivation for our experimental design, as detailed in Section V. Similarly, we also conduct within-culture experiments with our proposed approach. We extend the state of the art by increasing the number of cultures to 6, and performing a culture independent experiment (cross-culture experiment). The proposed experimental design allows us to investigate the impact of multiple cultures in a cross-culture setting.

III. Culturally Diverse Datasets

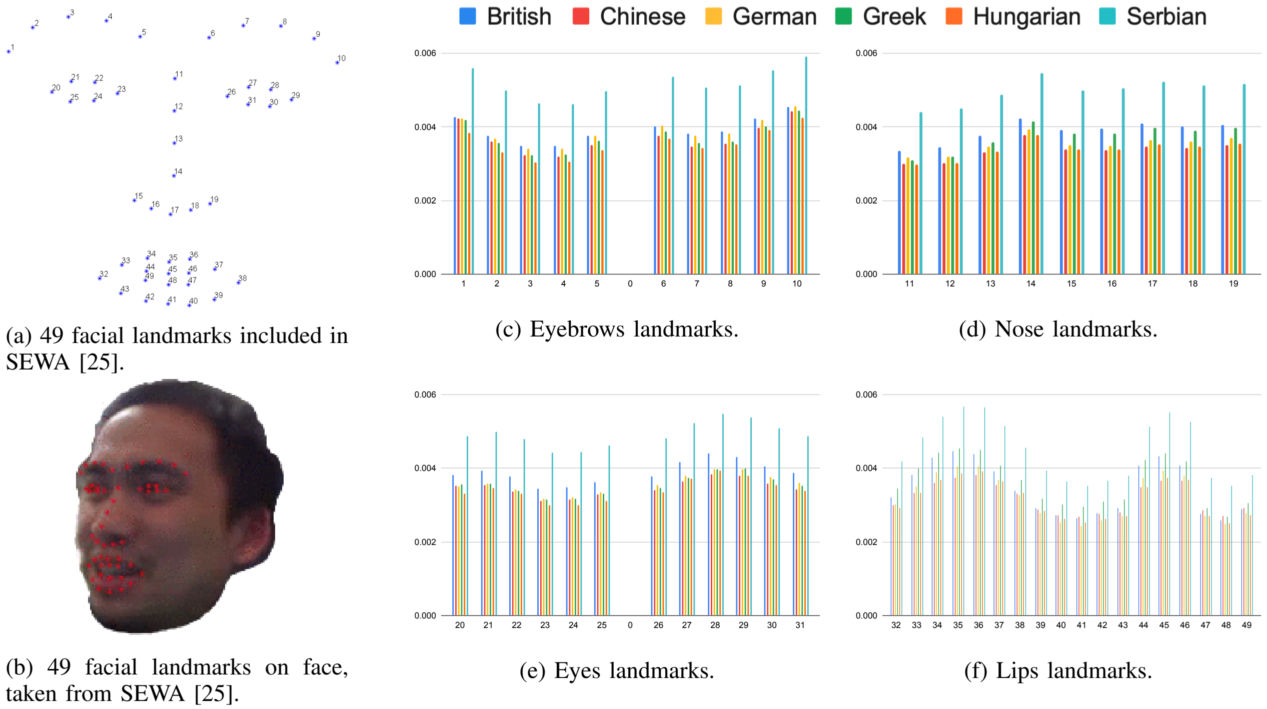
A. Single-Culture Datasets

In 2008, 12 hours of audio-visual data from German TV talk show "ldquoVera am Mittagrdquo" was collected in the VAM-faces dataset [37]. The videos are segmented into broadcasts, dialogue, and utterances. This dataset contains emotional, unscripted, and authentic discussions between talk show guests. Emotional labels, valence, activation, and dominance were annotated by human evaluators which are included in the dataset. RECOLA [35] is a multi-modal corpus of spontaneous, collaborative, and affective interactions of 46 French-speaking subjects. Data was collected during video conference sessions while subjects complete collaborative tasks. Audio, video, electrocardiogram, and electrodermal activity signals are provided from the sessions. Additionally, arousal, valence, and social behavior labels were annotated by 6 annotators, as well as self-reported measures and are included in the dataset. Although these datasets have merit, they are not a well-suited datasets for culture studies since they only contains subjects from French or German backgrounds. Considering our proposed study aims to investigate the impact of multiple cultures, these datasets were not appropriate for this investigation.

B. Multi-Culture Datasets

GENEVA dataset [38] was collected in 1997 and contains 112 audio files collected from the conversations of passengers in an international airport reporting lost luggage before and after interactions with an airline agent. Subjects speak in French, English, German/Northern Europe, Asian, and other languages. Type and intensity of the emotions felt by the subjects are reported as well. While this dataset contains multiple cultures, it only contains one modality (audio) and is relatively small with 112 audio files. Considering this, GENEVA was not an appropriate dataset for this investigation as multiple modalities are required along with multiple cultures.

SEWA [25] has diversity in culture and spoken language of the subjects including Chinese, English, German, Greek, Hungarian, and Serbian. Subjects complete identical tasks and converse in their native languages. This dataset contains data from 398 subjects (201 male and 197 female) ranging from 18 to over 60 years old. Each subject participates in five tasks: watching four advertisement videos designed to elicit disgust, pleasure, confusion, and interest. Following the fourth video, two participants from the same culture discuss the last advertisement for three minutes as the fifth task. The dataset contains audio, video, facial landmarks, action unit intensities, valance, arousal, and dialogue transcripts. Five annotators from the same culture as the subjects have annotated the valance and arousal levels from audio, video, and both audio and video of sessions. Due to it's multimodal and cross-cultural nature, SEWA is used for all analysis and experiments in this work.



(a) 49 facial landmarks included in SEWA [25].

(c) Eyebrows landmarks.

(d) Nose landmarks.

(b) 49 facial landmarks on face, taken from SEWA [25].

(e) Eyes landmarks.

(f) Lips landmarks.

Fig. 1: (a) Facial landmark template from SEWA [25]; (b) sample landmarks on subject from SEWA; (c) normalized facial landmark movement for eyebrows; (d) normalized facial landmark movement for nose; (e) normalized facial landmark movement for eyes; (f) normalized facial landmark movement for lips. In 1e (1c) the left 6 (5) columns correspond to left eye (eyebrow) and the 6 (5) columns correspond to right eye (eyebrow). It can be observed that while facial movements in all cultures often follow a similar trend, people from Serbian culture move their faces comparatively more than all other investigated cultures.

IV. Cultural Behavior Analysis

The proposed analysis on differences in cultural behavior is detailed here. For this investigation, Facial landmarks, action units, and head pose are evaluated. Along with this, valance and arousal are also investigated. Using these modalities, we look for differences in behaviors from different cultures including movement, expressions, and intensity of arousal.

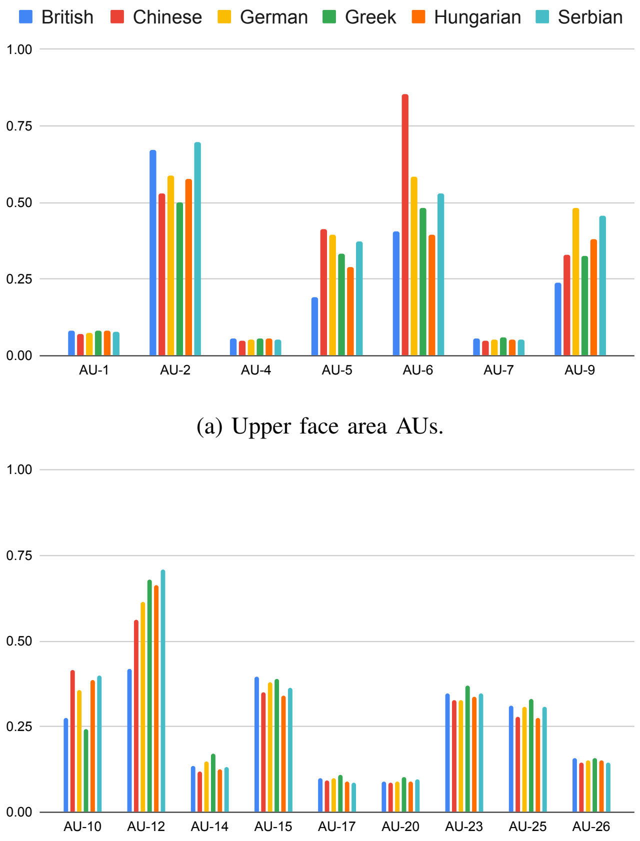
A. Facial Landmarks

For our analysis on facial landmarks, 49 points on the subject's face are tracked for each frame. These points detail where the eyes, eyebrows, nose, and lips are in each frame. We use the facial landmarks contained in SEWA dataset [25] for our investigation. A template of where each landmarks are can be seen in Fig. 1a, with an example frame from SEWA in Fig. 1b. For analysis of facial landmarks, we are interested in the overall movement on a frame-to-frame basis. To do this, the Euclidean distance of each (x, y) landmark is used to capture this movement. Then, each landmark movement is normalized over all data using min-max normalization as $l_{norm} = \frac{l-l_{min}}{l_{max}-l_{min}}$. Where l is the initial landmark movement value and l_{min} and l_{max} are the minimum and maximum Euclidean distance of the corresponding landmark movement across all data, respectively. The normalized landmark movement values are then averaged over all frames of all videos of each culture. For clarity, the facial landmarks are categorized into eyebrows, eyes, nose, and mouth (Figs. 1c - 1f).

As can be seen in Fig. 1, people from the Serbian culture have more facial landmark movement compared to all other investigated cultures. This can be seen for all parts of the face. Conversely, people from Chinese and Hungarian cultures, show less movements in their face compared to the other cultures. Across eyebrows, nose, and eyes it can be seen that there is a similar amount of movement. Based on the context in SEWA (i.e., watching ads or talking), this makes sense as these landmarks are largely on the upper half of the face. Conversely, there is more variation is facial movement for the lips across all cultures. This can be explained, in part, by the lips moving when people are talking. As the Serbian culture has larger movements, compared to other cultures, this suggests that subjects from this culture more often talked when discussing the ads. Conversely, Chinese and Hungarian subjects may have talked less during the discussion due to the smaller amount of movement.

B. Facial Action Units

A combination of Action Units (AUs) [3] have been used to detect facial expressions potentially linked to an inner emotional state [39]. It is important to note, however, that it has also been shown that emotion and facial expressions are not the same thing. Context, culture, and other factors largely influence emotion [21]. Here, we analyse AU intensities across cultures. As AU intensities are not provided in SEWA, we used the open source facial behavior toolkit, OpenFace [40], to extract them. In this investigation, intensities for AUs



(a) Upper face area AUs.

(b) Lower face area AUs.

Fig. 2: Average normalized action unit intensities across all cultures. (a) shows AUs from the upper part of the face, and (b) shows AUs from the lower part of the face.

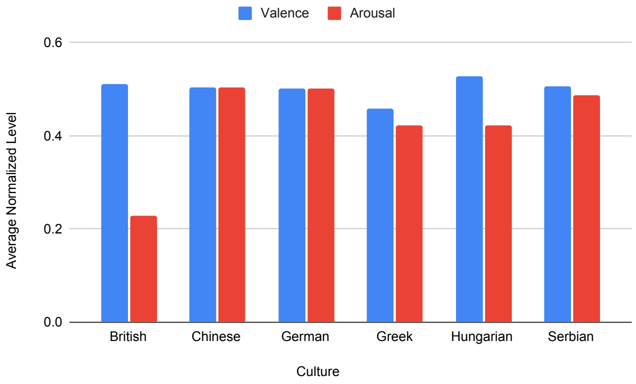


Fig. 3: Average normalized arousal and valence levels across all cultures cultures.

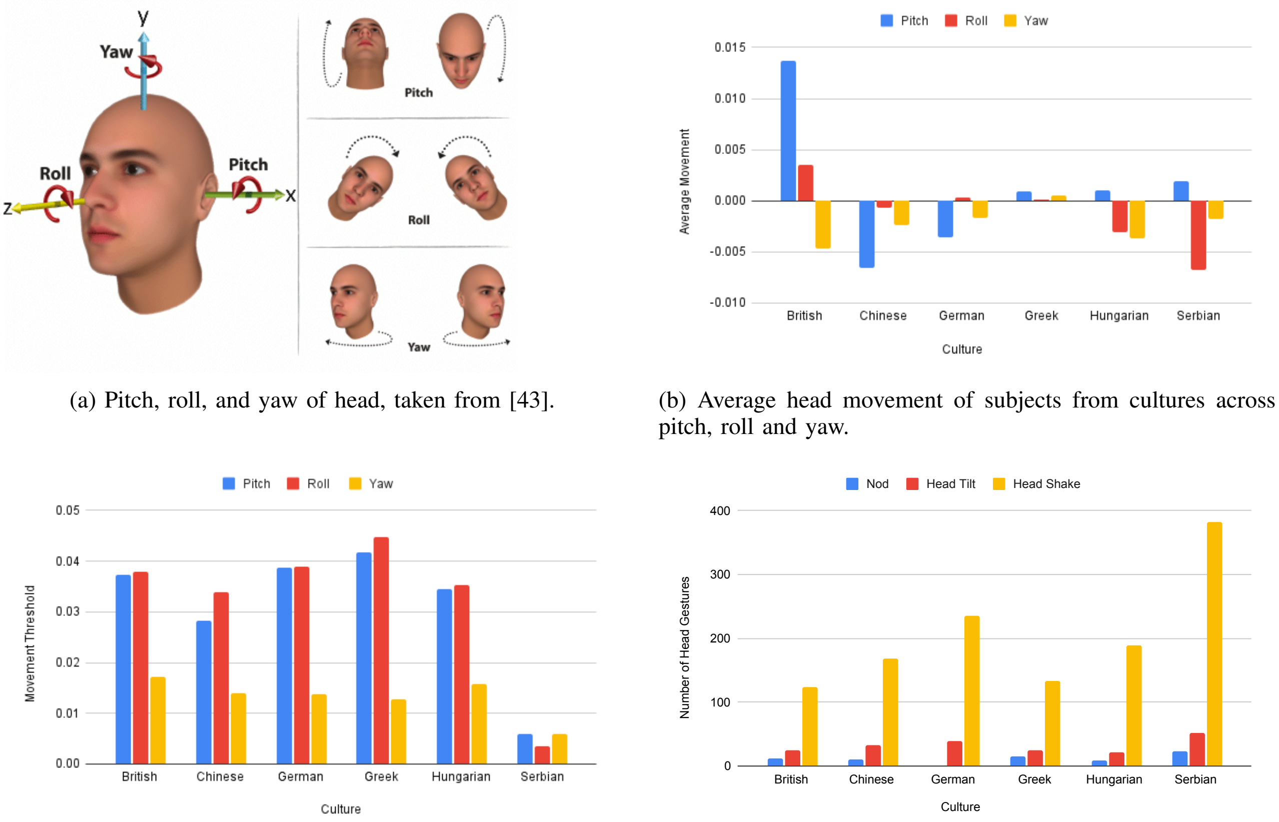
$\{1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 20, 23, 25, 26\}$ are tracked for each frame of a video. These AUs were selected as they are commonly used in the AU literature [39], [41].

Each AU intensity was normalized over all data, again using min-max normalization as done for the facial landmarks. For clarity, the AUs are categorized into the upper face area (eyes, eyebrows, and nose: AUs $\{1, 2, 4, 5, 6, 7, 9\}$) and lower face area (lips, cheeks, and chin: AUs $\{10, 12, 14, 15, 17, 20, 23, 25, 26\}$). The results of our comparison is presented in Fig. 2. There are some interesting findings that can be seen in this figure. First, we can see a relatively large spike in AU 6 (Fig. 2a) in the Chinese culture. This suggests that Chinese subjects often had intense smiles, as it has been shown that AU 6 may be an artifact of smile intensity [42]. Next, it can be seen that AU 6 and AU 12 (Fig. 2b) have a lower intensity in the British culture. These two AUs are often associated with a smile [9]. Conversely to the Chinese subjects, these results suggest that the British subjects smiled less compared to other cultures. British subjects also had lower intensities of AU 9, which is often associated with disgust [21]. In the SEWA dataset, disgust was one of the emotions that the videos were meant to elicit. This suggests British subjects may have experienced less disgust overall. Conversely German subjects had the higher intensity of AU 9, therefore suggesting they may have experienced more disgust. Lastly, in many cases Serbian subjects have a relatively high intensity of AUs. This is similar to the results from Fig. 1, where Serbian subjects had high facial landmark movement. This suggests that, generally, the Serbian subjects may have been more expressive (e.g., talked and smiled).

C. *Valence and Arousal*

One model of emotion suggests that emotions can be described along the dimensions of valence and arousal. Valence shows how negative or positive an experience is and arousal indicates how calming or exciting it is [44]. Here, we investigate the emotional differences (along valence and arousal dimensions) across different cultures. In SEWA, continuous values of valence and arousal for each subject have been manually annotated by 5 annotators, from the same culture as the subject, based on audio, video, and both audio and video data. The annotation was done in real time using a joystick. The minimum and maximum values for valence and arousal levels are -1000 and 1000 respectively. We again normalized the valence and arousal levels using min-max normalization. Since the joystick frequency and frame per second of the videos are not aligned, we don't have one value per frame for arousal or valence for each annotator. Additionally, not all annotators have annotated all timestamps. Thus, we have aggregated the values after normalizing them. More specifically, for each timestamp, if there is more than one value, we take the average of the values and if not, we have taken the single existing value. We calculated the average normalized valence and arousal levels for each culture.

It can be seen in Fig. 3 that many of the cultures had similar levels of valence and arousal, however, there are some notable differences. More specifically, British, Greek, and Hungarian cultures showed different levels. Each of these cultures have a relatively lower arousal level compared to other cultures, especially British. This may be explained, in part, by the annotators being from the same culture. British subjects may either show lower arousal levels or the British annotators interpreted expressions with lower intensity levels. This is supported by the finding that third party annotators often annotate lower intensity of expressions [45].



(a) Pitch, roll, and yaw of head, taken from [43].

(b) Average head movement of subjects from cultures across pitch, roll and yaw.

(c) Head movement thresholds for pitch, roll, and yaw. Lower values correspond to a smaller movement range.

(d) Head gestures count. Larger values corresponds to more head gestures overall.

Fig. 4: Head pose and movement across all cultures. (a) shows an example of roll, pitch, and yaw; (b) shows average head movement for roll, pitch, and yaw; (c) shows movement thresholds across roll, pitch, and yaw; and (d) shows total number of gestures made across all cultures.

D. *Head Pose Movement*

Here, we compare head movement (i.e., pose) across all cultures. In SEWA, head pose in each frame includes pitch, yaw, and roll, which are the movements of the head around the x-, y-, and z-axes (Fig. 4a), respectively. An analysis on the average pitch, yaw, and roll values across all frames for each culture has been conducted (Fig. 4b). It can be seen that, on average, British subjects tend to move their head down (positive pitch) and to the front more often than other cultures. Conversely, Chinese subjects move their head up (negative pitch) and to the back more. It can also be seen that British subjects often tilt their heads to the left (positive roll) considerably more than others while Serbian cultures tilt their head to the right (negative roll) more than other cultures. Interestingly, when comparing yaw, British people angle to the right (negative yaw) of the camera more than other cultures and Greek subjects are the only culture that look to the left (positive yaw) of the camera on average.

Next, we define three head gestures nod, head shake, and head tilt as a reciprocating movement of the subject's head in pitch, yaw, and roll within 10 frames. To filter head movements and select more meaningful ones, we define a threshold for pitch, yaw, and roll for each culture as $t_{mov.} = \frac{m_{max}-m_{min}}{n}$ in which m_{max} and m_{min} are the maximum and minimum of the corresponding movement across all videos within the culture, and n is the number of videos in the culture. If the head movement direction has changed once and the summation of movement (pitch/yaw/roll) in the last 10 frames surpasses the corresponding threshold, we count it as a meaningful head gesture. Looking at the movement thresholds (Fig. 4c), it can be seen that the Serbian culture has lower thresholds for all movements compared to other cultures. This means that the difference between maximum and minimum amount of movement is less. In other words, it suggests that the Serbian culture moves their heads in a more subtle way with less movement range. It can also be seen that roll thresholds are higher than pitch and yaw (except for Serbian), suggesting that people tend to move their head more freely along the x-axis (nods). It can also be seen that yaw values, for all cultures, are relatively low. This suggests that most subjects, across all cultures, generally focused in one direction, and had a relatively small range of movement across yaw. In other words, subjects may have focused in one general direction (e.g., at or near the screen/partner during discussion).

Finally, we counted the occurrence of each head gesture (nod, shake, and tilt) across all cultures (Fig. 4d). It can be seen that in all cultures that head shake is most occurring head gesture. Across all cultures, this difference is often at least an order of magnitude greater. For example, in the Hungarian culture there are 9 nods, and 189 shakes. An even larger difference is German, where there is one nod and 236 shakes. Additionally, Serbian subjects have the highest number of head gestures, with a total of 457 gestures. This suggests that although people from the Serbian culture move their head in a more subtle way (Fig. 4c), they do move their head more often than other cultures. Conversely, British subjects have the smallest number of head gestures with a total of 160 gestures. This suggests that while British subjects more often move their head down (Fig. 4b), they move their head less often.

V. Experimental Design and Results

A. Experiments

To further justify the role of culture in affect-related tasks, we train a ViViT model [46] to perform context classification using the face-aligned video data. We are motivated to classify context as it has been shown that the classification of context is an important application in affect-related research [47]. The experiments are designed in such a way to investigate whether the incorporation of culture has positive impact on classification results. More specifically, three context-based classification experiments are designed using the video data from SEWA, and all cultures.

- *Culture Independent (CI)*: Here, our model is trained on all videos from five cultures and tested on the videos of the sixth culture. The model is evaluated on 6 different test sets, one for each culture. In each iteration, all the participant videos belonging to 5 of 6 cultures are used as the training set and the videos from the sixth culture are the test set. This is repeated until videos of all 6 cultures fall under the test set once. The purpose of this experiment is to understand the model's capability in classifying the context on an unseen culture.
- *Subject Independent within a culture (SIc)*: Under SIc, we perform 10-fold cross-validation, where each fold has several subjects, but the training and testing testing sets are restricted to one specific culture. This is repeated for all the six cultures. The purpose of this experiment is to evaluate the effect of individual culture in subject independent mode. This experiment is directly motivated by the work from Rudovic et al. [36] (as detailed in Section II).
- *Subject Independent (SI)*: We perform 10-fold subject-independent cross validation, wherein, each fold has several participants randomly selected as the test set, irrespective of their cultural background. This is to gauge the model's ability in classifying on a new subject when culture is not considered.

B. Network Architecture

Fig. 5 shows the overall network architecture used for each experiment. Each full length face video is fed to a *temporal sampling* module to extract the most useful frames in the video. Next, the frames are pre-processed using RetinaFace [48] to crop out faces from the raw videos. Finally, the fixed length frames are fed to ViViT model to predict the context. We are motivated to use a ViViT as it has been shown to process temporal affect information effectively [49], [50].

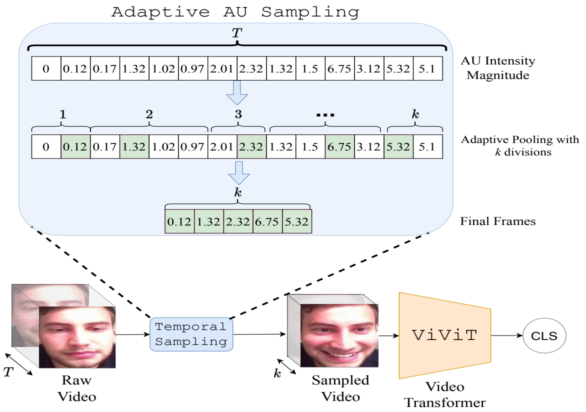


Fig. 5: Overall pipeline. Each face video is fed to the *temporal sampling* module which extracts the k most useful frames from a total of T frames. Next, the k frames are fed to a video transformer model (ViViT) that classifies the video into a specific context.

C. Temporal Sampling Module

Due to the large number of frames present in each video, we employ an adaptive frame sampling strategy based on the extracted action units. Randomly sampling a video with T frames to a fixed length k can potentially select frames with less useful information (for example, neutral expression). However, sampling the most expressive temporally consistent frames can benefit the model. To achieve this, we first compute the AU magnitude for the AU intensities for each frame in the video as $AU_{mag} = \sqrt{AU_1^2 + AU_2^2 + ... + AU_{26}^2}$. Next, to be temporally consistent, we perform an adaptive max-pooling on the T AU magnitudes to select k fixed indices with highest AU_{mag} within the region. The selected frames (the highlighted green cells in Fig. 5) are now the candidate frames to be processed by the ViViT model. This sampling strategy is different from naively selecting top k frames with highest AU_{mag}, since this can constrain the frames to a specific region, thus losing the temporal information. Note that AU_{mag} has been shown to inherently capture spatial expressiveness as shown by Uddin et al. [51]. In our experiments, we report values for both random and the proposed adaptive sampling strategy and show that the proposed strategy exhibits higher performance in all scenarios. We report accuracy and Matthew's correlation coefficient (MCC) scores.

D. Results

Culture Independent (CI). Table I reports the evaluation metrics for both uniform and adaptive sampling for culture-independent analysis. It can be seen from the table that culture plays an important role in affect, with the model performing the least when *Chinese* culture is used as test set with 39.7% and 0.25 as the accuracy and MCC, respectively. We observe that proposed frame sampling shows higher performance across all cultures compared to uniform sampling, effectively

Culture (Validation)	Sampling	Accuracy(%) ↑	MCC ↑
British	Uniform	40.00	0.25
	Adaptive	**42.70**	**0.30**
Chinese	Uniform	39.70	0.25
	Adaptive	**47.10**	**0.35**
German	Uniform	40.00	0.25
	Adaptive	**44.70**	**0.32**
Greek	Uniform	45.40	0.32
	Adaptive	**49.60**	**0.37**
Hungarian	Uniform	**45.10**	**0.32**
	Adaptive	**45.10**	**0.32**
Serbian	Uniform	41.90	0.28
	Adaptive	**44.40**	**0.31**
Average	Uniform	42.00	0.28
	Adaptive	**45.60**	**0.33**

TABLE I: Culture Independent (CI) cross-validation. Each row represents a culture evaluated as test set.

demonstrating our sampling approach. Overall, the model's poor performance on unseen cultures further highlights it's reliability on culture specific training. These results also show that the obtained results depend on which culture is left out. An interesting question is why this occurs, but is out of scope and left for future work.

Subject Independent Within a Culture (SIc). To extend our study on performance within a culture, we next report the results for subject-independent analysis within a culture (SIc) in Table II. Consequently, we observe an improved performance for both sampling strategies across all cultures when the model operates on only known cultures in the test set with model showing the best performance on the *Greek* culture. Furthermore, for Tables I and II, the performance improvement between uniform and adaptive sampling is highest for the *Chinese* culture with 7.4% for *CI* and 11.15% for *SIc*. The increase in performance can be partly attributed to the AU analysis in Fig. 2a which showed a distinct AU 6 pattern. More specifically, the AU-based sampling may have incorporated these distinct frames in the training and validation process. It is important to note that while the overall accuracies are generally low for both CI and SIc, regardless of sampling strategy, the *main contribution is the impact of culture*. Incorporating culture (SIc) gives an average improvement of 5.1% and 0.06 for accuracy and MCC, respectively. We hypothesize that similar experiments (i.e., culture-based with temporal sampling) with more advanced network architectures will result in overall higher accuracies and MCC scores.

Subject Independent (SI). Finally, we report the subject-independent (SI) performance of the model in Table III. We observe that on average, the model performs better than the *CI* setting and worse than the *SIc* setting. More specifically, the average SIc improves the SI results by 2.6% and 0.03 for accuracy and MCC, respectively. These results appear intuitive as the former operated on unseen cultures and the latter operated on only known cultures. Further suggesting the need to incorporate culture.

Culture (Validation)	Sampling	Accuracy(%) ↑	MCC ↑
British	Uniform	43.34	0.31
	Adaptive	**48.49**	**0.37**
Chinese	Uniform	39.70	0.26
	Adaptive	**50.85**	**0.40**
German	Uniform	47.42	0.36
	Adaptive	**51.50**	**0.41**
Greek	Uniform	51.00	0.41
	Adaptive	**56.00**	**0.47**
Hungarian	Uniform	46.87	0.35
	Adaptive	**50.10**	**0.39**
Serbian	Uniform	**47.27**	**0.35**
	Adaptive	**47.27**	**0.35**
Average	Uniform	45.93	0.34
	Adaptive	**50.70**	**0.39**

TABLE II: Subject Independent within a culture (SIc) cross-validation. Each row reports the average of 10-fold subject independent cross-validation for that culture.

	Sampling	Accuracy(%) ↑	MCC ↑
Average	Uniform	44.33	0.31
	Adaptive	**48.10**	**0.36**

TABLE III: Subject Independent (SI) cross-validation.

VI. Conclusion

We have shown that there are differences in facial landmark movements, action unit intensities, head pose, and valence and arousal across six cultures. In addition to culture-focused behavior analysis in SEWA, we have conducted experiments to assess the impact of culture in context classification. Using the videos in SEWA, we have done three experiments: Culture Independent (CI), Subject Independent Within a Culture (SIc), and Subject Independent (SI). The results of our experiments show that culture specific models yield better results for all investigated cultures. The results are encouraging, suggesting that when culture is considered, affect-related applications can show an improvement in performance.

While these results are encouraging, there are some limitations and future work that can be implemented. First, only one dataset (SEWA), was evaluated in our investigation. To the best of our knowledge, SEWA is the only multimodal, cross-cultural dataset that allows for this type of investigation. Considering this, collecting more multimodal datasets that focus on culture can be useful for the community at large. Second, we have only classified context in this work. Interesting future work is to classify arousal and valence, discrete expressions (e.g., happy or sad), or even the culture itself. Third, one network architecture was used for the experimental design. More networks need to be evaluated to learn which type is best for classification when culture is incorporated. Finally, it would be interesting to evaluate how different cultures annotate the valence and arousal levels of across cultures. This would require cross-culture annotators for all data.

VII. Ethical Impact Statement

Racial and cultural biases can be named as some roots of ethical concerns in affective computing. In this study we have investigated cultural behavioral differences, which may help to identify some sources of cultural biases present in systems today. Our approach has been to mitigate the limitations on generalizability across diverse cultures by formulating models that are sensitive to cultural specifics. We believe the data that we used in our experiments were collected in an ethically responsible way. Nevertheless, we acknowledge the persistence of certain ethical concerns. We have used machine learning tools to extract some features of data, such as facial action units, and cropping the subjects' face in videos and we recognize the potential biases within those tools including those related to age, gender, or race. One potential negative application of this work can be provoking existing cultural stereotypes or creating new ones. Along with this, anytime machine learning models are trained with human data, caution must be used when deploying these systems. In this work, we classified context using face videos, however, the work can be extended to other affect-related areas, such as expression. Considering this, a human should always remain in the loop if and when these models are deployed.

References

[1] L. F. Barrett and C. Westlin, "Navigating the science of emotion," in *Emotion measurement*. Elsevier, 2021, pp. 39–84.

[2] K. Hoemann, M. Gendron, A. N. Crittenden, S. M. Mangola, E. S. Endeko, È. Dussault, L. F. Barrett, and B. Mesquita, "What we can learn about emotion by talking with the hadza," *Perspectives on Psychological Science*, vol. 19, no. 1, pp. 173–200, 2024.

[3] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[4] S. Volynets, E. Glerean, J. K. Hietanen, R. Hari, and L. Nummenmaa, "Bodily maps of emotions are culturally universal." *Emotion*, vol. 20, no. 7, p. 1127, 2020.

[5] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[6] P. Ekman, "Universals and cultural differences in facial expressions of emotion." in *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[7] N. Lim, "Cultural differences in emotion: differences in emotional arousal level between the east and the west," *Integrative Medicine Research*, vol. 5, no. 2, pp. 105–109, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213422016300191

[8] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[9] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[10] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018.

[11] A. I. d. Paiva-Silva, M. K. Pontes, J. S. R. Aguiar, and W. C. de Souza, "How do we evaluate facial emotion recognition?" *Psychology & neuroscience*, vol. 9, no. 2, p. 153, 2016.

[12] D. Fabiano and S. Canavan, "Emotion recognition using fused physiological signals," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 42–48.

[13] G. Bargshady, J. Soar, X. Zhou, R. C. Deo, F. Whittaker, and H. Wang, "A joint deep neural network model for pain recognition from face," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2019, pp. 52–56.

[14] A. Ruiz-Garcia, N. Webb, V. Palade, M. Eastwood, and M. Elshaw, "Deep learning for real time facial expression recognition in social robots," in *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part V 25*. Springer, 2018, pp. 392–402.

[15] J. Awatramani and N. Hasteer, "Facial expression recognition using deep learning for children with autism spectrum disorder," in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, 2020, pp. 35–39.

[16] L. Chen, "Facial expression recognition with machine learning and assessment of distress in patients with cancer," *Number 1/January 2021*, vol. 48, no. 1, pp. 81–93, 2021.

[17] D. C. Ong, "An ethical framework for guiding the development of affectively-aware artificial intelligence," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[18] S. Yan, H.-T. Kao, K. Lerman, S. Narayanan, and E. Ferrara, "Mitigating the bias of heterogeneous human behavior in affective computing," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[19] S. Ramis, J. M. Buades, F. J. Perales, and C. Manresa-Yee, "A novel approach to cross dataset studies in facial expression recognition," *Multimedia Tools and Applications*, vol. 81, no. 27, pp. 39 507–39 544, 2022.

[20] B. Han, W.-H. Yun, J.-H. Yoo, and W. H. Kim, "Toward unbiased facial expression recognition in the wild via cross-dataset adaptation," *IEEE Access*, vol. 8, pp. 159 172–159 181, 2020.

[21] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[22] M. Sohail, G. Ali, J. Rashid, I. Ahmad, S. H. Almotiri, M. A. AlGhamdi, A. A. Nagra, and K. Masood, "Racial identity-aware facial expression recognition using deep convolutional neural networks," *Applied Sciences*, vol. 12, no. 1, p. 88, 2021.

[23] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 506–523.

[24] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 881–893, 2020.

[25] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 1022–1040, 2019.

[26] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[27] S. Han, "Understanding cultural differences in human behavior: a cultural neuroscience approach," *Current Opinion in Behavioral Sciences*, vol. 3, pp. 68–72, 2015, social behavior. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352154615000236

[28] W. V. Friesen, *Cultural differences in facial expressions in a social situation: An experimental test of the concept of display rules*. University of California, San Francisco, 1972.

[29] D. Matsumoto, "Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an american sample," *Motivation and emotion*, vol. 17, no. 2, pp. 107–123, 1993.

[30] Y. Huang, S. Tang, D. Helmeste, T. Shioiri, and T. Someya, "Differential judgement of static facial expressions of emotions in three cultures," *Psychiatry and clinical neurosciences*, vol. 55, no. 5, pp. 479–483, 2001.

[31] D. Matsumoto and P. Ekman, "American-japanese cultural differences in intensity ratings of facial expressions of emotion," *Motivation and emotion*, vol. 13, pp. 143–157, 1989.

[32] Z. Zhang, "Exploring interpersonal harmony: A cross-cultural comparison between eastern and western societies," *Arts, Culture and Language*, vol. 1, no. 4, 2023.

[33] M. N. Dailey, C. Joyce, M. J. Lyons, M. Kamachi, H. Ishi, J. Gyoba, and G. W. Cottrell, "Evidence and a computational explanation of cultural differences in facial expression recognition." *Emotion*, vol. 10, no. 6, p. 874, 2010.

[34] J. Han, Z. Zhang, M. Pantic, and B. Schuller, "Internet of emotional people: Towards continual affective computing cross cultures via audiovisual signals," *Future Generation Computer Systems*, vol. 114, pp. 294–306, 2021.

[35] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[36] O. Rudovic, Y. Utsumi, J. Lee, J. Hernandez, E. C. Ferrer, B. Schuller, and R. W. Picard, “Culturenet: a deep learning approach for engagement intensity estimation from face images of children with autism,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 339–346.

[37] M. Grimm, K. Kroschel, and S. Narayanan, “The vera am mittag german audio-visual emotional speech database,” in *2008 IEEE international conference on multimedia and expo*. IEEE, 2008, pp. 865–868.

[38] K. R. Scherer and G. Ceschi, “Lost luggage: a field study of emotion–antecedent appraisal,” *Motivation and emotion*, vol. 21, pp. 211–235, 1997.

[39] P. Yang, Q. Liu, and D. N. Metaxas, “Boosting coded dynamic features for facial action units and facial expression recognition,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–6.

[40] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “Openface: An open source facial behavior analysis toolkit,” in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.

[41] P. Khorrami, T. Paine, and T. Huang, “Do deep neural networks learn facial action units when doing expression recognition?” in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 19–27.

[42] J. M. Girard, G. Shandar, Z. Liu, J. F. Cohn, L. Yin, and L.-P. Morency, “Reconsidering the duchenne smile: indicator of positive emotion or artifact of smile intensity?” in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 594–599.

[43] A. Radmehr, M. Asgari, and M. T. Masouleh, “Experimental study on the imitation of the human head-and-eye pose using the 3-dof agile eye parallel robot with ros and mediapipe framework,” in *2021 9th RSI International Conference on Robotics and Mechatronics (ICRoM)*. IEEE, 2021, pp. 472–478.

[44] E. Kensinger, “Remembering emotional experiences: The contribution of valence and arousal,” *Reviews in the Neurosciences*, vol. 15, no. 4, pp. 241–252, 2004. [Online]. Available: https://doi.org/10.1515/REVNEURO.2004.15.4.241

[45] M. T. Uddin, L. Yin, and S. Canavan, “Spatio-temporal graph analytics on secondary affect data for improving trustworthy emotional ai,” *IEEE Transactions on Affective Computing*, 2023.

[46] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “Vivit: A video vision transformer,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[47] S. Hinduja, G. Kaur, and S. Canavan, “Investigation into recognizing context over time using physiological signals,” in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[48] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-stage dense face localisation in the wild,” *arXiv preprint arXiv:1905.00641*, 2019.

[49] S. Zhang, Y. Pan, and J. Z. Wang, “Learning emotion representations from verbal and nonverbal communication,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 993–19 004.

[50] Z. Ren, J. Ortega, Y. Wang, Z. Chen, Y. Guo, S. X. Yu, and D. Whitney, “Veatic: Video-based emotion and affect tracking in context dataset,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4467–4477.

[51] M. T. Uddin and S. Canavan, “Quantified facial expressiveness for affective behavior analytics,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 985–994.