Study of Interfaces for Time-Continuous Emotion Reporting and the Relationship Between Interface and Reported Emotion

Jason W. Woodworth* Christoph W. Borst†

University of Louisiana at Lafayette

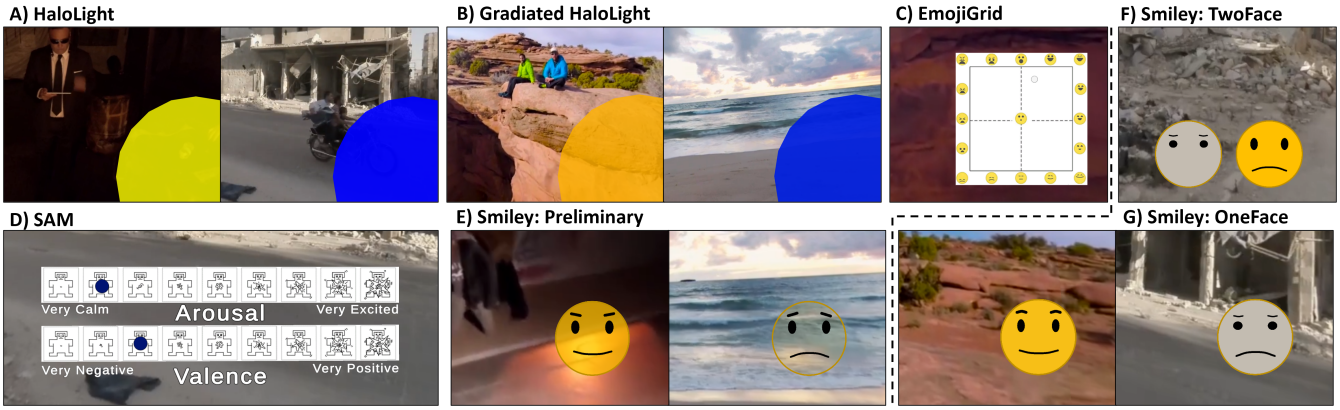

Figure 1: The interfaces compared in these studies with example 360° videos. Visuals show either a happy (Arousal=8, Valence=5) or sad (Arousal=2, Valence=3) report. Images are cropped/zoomed to increase interface detail due to the large field of view in VR.

Abstract

This paper presents interfaces for reporting emotion in real-time during VR stimuli. Self-reported emotional responses are critical for developing emotion recognition systems. Such responses can vary throughout a stimulus such as 360° video, but most interfaces for reporting emotion are designed to be used after the experience. This reduces the entire experience to a single data point and raises concerns about validity when multiple emotions can be elicited across the stimulus. We introduce and compare user interfaces that allow for real-time emotion reporting throughout the length of the stimulus. Each interface varies on how emotion is physically input by the user and displayed back to them for confirmation. A preliminary study compared five such interfaces, gathering initial impressions, comparing control schemes, and rating intuitiveness. A primary study considered four refined interface designs and compared reporting precision and subjective opinions. Results suggest that a single interface face icon responding to arousal and valence reports and a gradiating color wheel are intuitive, precise, and unobtrusive. More broadly, results indicate the type of rating interface has a significant effect on the given ratings.

Keywords: Affective Computing, Emotion Rating Interfaces

1 Introduction

Automated emotion recognition hinges on quality data that maps physiological and behavioral patterns to labeled emotions. This labeling is often done post-hoc, allowing the user to report how the stimulus made them feel after the experience. Methods and interfaces for users to perform this style of reporting have been widely investigated, ranging from the lengthy Semantic Differential scale [20] to the simplified Self-Assessment Manikin (SAM) [5] to pictorial options like AffectButton [6].

The use of these post-hoc styles of self-assessment comes with notable drawbacks. It ignores the fact that emotions vary on a moment-to-moment basis and a single stimulus can elicit multiple emotions over time [9–11, 18, 35]. For example, a narrative film clip may aim to elicit sadness in the first half, then aim to relieve tension with amusement in the second half. Reducing this experience to a single self-reported emotion value at the end then introduces questions about which part of the experience the report reflects and what physiological data should be associated with it.

Some have attempted to overcome these drawbacks by associating multiple self-reported labels with a single stimulus. One way to do this is through cued recall [4, 7, 18], in which the subject experiences the full stimulus, then is shown individual clips from their experience and asked to rate how they felt at that moment. This technique has been shown to not introduce the same memory bias as free recall, which is associated with subjects only reporting the most intense feelings [7], and is thus considered an effective strategy. However, the nature of the technique incurs additional time requirements. This can become problematic if the proctor wishes to collect a large amount of data, as long experiment times can lead to boredom that can negatively affect emotion-related research [2].

Another strategy is to have subjects report their emotions in a time-continuous manner as they are experiencing the stimulus. For example, while a subject is watching a video they may turn a dial [18], move a joystick [34], or push a throttle [9] to indicate a position on a dimension-based emotion model to indicate what they feel in the moment. This naturally reduces time, captures the full range of the emotional experience elicited by the stimulus, and has the added benefit of producing fine-grained data for use in emotion recognition models. While this removes the dangers of memory bias associated with cued recall, it introduces the concern that subjects may not understand their emotions while they experience them and the act of monitoring emotion may introduce an amount of cognitive load that distracts from the experience. These concerns appear to be assuaged by reports that time-continuous ratings match up with

*e-mail: jason.woodworth1@louisiana.edu
†e-mail: cwborst@gmail.com

This is an author-formatted version. Original publication: J. W. Woodworth and C. W. Borst, "Study of Interfaces for Time-Continuous Emotion Reporting and the Relationship Between Interface and Reported Emotion," 2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Bellevue, WA, USA, 2024

cued-recall [18] and end-of-task assessments [35].

Concerns about increased cognitive load can be further mitigated by giving the user visual feedback on what emotion they are reporting to the system. While these have been explored more outside the immersive VR context, the all-encompassing nature of video within VR changes the requirements for these visuals. Continuous rating interfaces within VR have begun to be explored, with HaloLight and DotSize being prominent examples [35]. However, the design process for these interfaces was minimal, being based on input from a small number of experts and facing no formal comparisons to other designs. In fact, to our best knowledge, no formal comparison of any kind has been done on VR interfaces for time-continuous emotion rating. We consider the need to explore a larger design space, including different control schemes and interface visuals.

In this work, we introduce several novel interface designs and continuous adaptations of existing methods. As an exploratory work, we consider the following broad research questions: What design elements of VR interfaces for time-continuous emotion reporting are most effective in terms of reporting speed, precision across multiple reports, and user preference? What is the effect of the type of rated emotion on reporting speed and precision? How does the rated emotion and reporting interface interact to affect speed, precision, and the reported values themselves?

To answer these, we first conducted an exploratory preliminary study with a small number of subjects to determine the most viable designs for further study and what control schemes should be used for them. A primary study then goes into further detail on selected interfaces, comparing their speed, precision, and intuitiveness across various kinds of emotion. Results suggest that users find a face icon with expression changed by arousal / valence input to be more intuitive than color-based visuals, but overall may prefer the color-based visual with interpolated colors between diagonal extremes of the arousal / valence spectrum. Results also point to an interaction between interface and reported emotion, suggesting a need for deeper exploration of the impacts of emotion reporting interfaces. Given this, we consider our contributions to be the following:

- Introduction and evaluation of two well-performing continuous emotion rating interfaces: Smiley and Gradiated HaloLight.
- Insight into key interface design considerations and guidance for future designs.
- Evidence for the interaction between interface design and different types of emotion, showing that interfaces can have a real effect on how we rate different emotions.

2 Background and Related Works

2.1 Models of Emotion

To allow users to report emotion, we must first define a model for them to report within. Emotion is typically modeled as discrete or dimensional [28]. Discrete models look to give names to a set of universal emotions, such as Ekman's seminal model of six basic emotions [22] or Plutchik's Wheel [23, 24], and tend to map well onto our everyday linguistic understanding of the world. They also, however, can be difficult to analyze quantitatively and may leave out emotions that are harder to express in words.

Dimensional models account for a lack of specificity in language by identifying a set of orthogonal dimensions that explain most of the variance in emotion, then considering all emotions as having a specific value along those axes. Several such models have been developed; among the most frequently used is Russell's arousal-valence model [25, 26] with emotions ranging on the arousal axis from calm to excited and on the valence axis from unpleasant to pleasant. Despite its simplicity, this model is shown to explain much of the variance in emotion and has been used in many works within [8, 15–17, 29, 31] and outside of [1, 14, 18, 21] a VR context.

In our context, a model of emotion is only useful if users can accurately and precisely note their felt emotions within it. Moreover, they must be able to do so quickly and with minimal cognitive overhead to allow for real-time continuous reporting. Thus, we consider the arousal-valence model gives a good tradeoff between perceptual simplicity and descriptive depth.

2.2 Self-Reported Emotion

In early emotion reporting work, a series of semantic differential scales were often used [20] measuring reactions by ratings on bipolar scales with two opposing adjectives. Ratings on these scales could then be factored down into individual values on the pleasure, arousal, and dominance axes. The Self-Assessment Manikin (SAM) [5] was later created to simplify this process by letting the user pick a value on each axis directly, aided by icons showing what a placement along that axis might represent. SAM, or some version of it, is likely the most commonly used rating interface as it is simple, straightforward, and easy to port to digital [3] and even VR interfaces [15].

Other interfaces for rating along these dimensions have been explored. For example, the AffectButton [6] interprets dimensional values based on mouse position on a 2D square and tweens a face graphic to approximate an appropriate emotion for that response. The EmojiGrid [30] allows the user to select a point on a 2D grid representing the arousal-valence space with emojis lining the edges that correspond to what extreme emotion they are expressing. These, and those used in many other works [13, 29, 36], are designed to be used post-hoc.

Other works have provided an interface for the user to report their emotion continuously throughout the experience. For example, an early work gave the user a custom dial that would control input along one dimension while the user watched a video [18]. Similarly, CARMA [10] allowed the user to use a mouse or arrow keys to manipulate a slider representing one dimension placed next to the video. DARMA [11], its followup, gave the user a joystick to move about the 2D space, again visually placed next to the video. Fayn et al. [9] introduced a novel throttle input device that could separate dimensions across different throttles, avoiding user conflation between the dimensions, with either slider or graph based imagery again placed next to the video.

However, such interfaces were designed for use on a desktop, with all the peripherals and design affordances that entails. Using VR changes the typically available input devices, and more importantly, using 360° video changes the available visual space to place interface visuals. To get around this, some VR works still use traditional video clips, just rendered within a designed virtual environment, that still allows them to place interface visuals next to the video [12]. The visually all-encompassing nature of 360° video, however, requires interface visuals to be overlaid atop the video. Recognizing this, Xue et al. [34, 35] prototyped several interfaces that would be visually simple and minimally invasive to reduce cognitive load. However, to our knowledge, no other works currently explore the design space of continuous rating interfaces to address the specific challenges brought up by VR or perform robust comparisons between different designs. We attempt to address this gap with this work.

3 Preliminary Interface Design

All designs (seen in Figure 1) visualize emotion based on a 2D input for an arousal-valence model of emotion. The input used Vive controller trackpads. Primary design considerations were that an interface should be intuitive, to minimize required training or reminders, and minimally invasive, to support focus on the content.

We considered four control schemes. Two one-handed schemes used horizontal and vertical movement on one trackpad, with the horizontal component being arousal (AV) or valence (VA). This allowed a user to easily specify quadrants by diagonal moves from trackpad center. Two two-handed schemes assigned arousal to left

	Int	Eff	MI	LI	Best	CS	Pro	Con
HL	0	0	1	1	0	VA	Simple	Colors not intuitive
GHL	0	2	2	1	2	VA	Color mixing helped precision	More colors is overwhelming
EG	1	0	1	0	0	VA	Labeling is intuitive	Invasive, too much to see
SAM	1	1	3	0	1	THX	Easy and precise after training	Requires training and visually large
Smiley	5	4	0	5	4	THY	Intuitive and non-invasive	Face doesn't capture all emotions

Table 1: Summary of preliminary results, indicating the number of times an interface was rated most intuitive (Int), most effective for watching videos (Eff), most invasive / distracting (MI), least invasive / distracting (LI), best overall, which control scheme was most preferred, and the most commonly given pros and cons.

and valence to right hands, with values changed with either vertical (THY) or horizontal (THX) movement. This supported mental separation of the two dimensions. Interface visuals were placed in the bottom of the user's vision, to be as unobtrusive as possible, and at a distance of 5.5 meters from the user to match the typically distant imagery in 360° videos.

HaloLight (HL): This method from prior work [34, 35] changed color of a circle to generally indicate the reported emotion. Quadrants of the arousal / valence spectrum were assigned certain colors, with red representing high arousal / low valence (stress), blue low arousal / low valence (sad), green low arousal / high valence (relaxed), and yellow high arousal / high valence (excited). Circle opacity represented emotion "intensity," or distance from neutral.

Gradiated HaloLight (GHL): One concern of HaloLight is that the use of one color per quadrant does not give feedback for variations within quadrants, e.g., tension and anger would look the same. Gradiated HaloLight addressed this by mixing the two colors of the two quadrants closest to the user's input, e.g., inputting high arousal and neutral valence would color the circle orange (mixing yellow and red). This gave finer-grained feedback showing different emotions within one quadrant.

EmojiGrid (EG): A continuous adaptation of [30], this interface presented a 2D arousal-valence grid with emojis representing emotions around its edges. A moving dot showed the user's reported arousal (vertically) and valence (horizontally) showing the user which emoji they are closest to reporting.

SAM: A continuous adaptation of a classic self-assessment manikin [5]. Nine manikins anchor numbers 1-9 on the arousal and valence scales, with expressions based on arousal or pleasure level. Moving dots were added to both scales to show current ratings.

Smiley: A novel approach with a dynamic face visual. Valence controlled a mouth curve, with the amount of smiling proportional to valence (neutral input gives a straight line, and lower values gives a frown). Arousal was mapped to either *opacity*, with low arousal making the face closer to transparent, or *eyebrow* tilt, with low arousal turning the eyebrows outward and high arousal turning them inward. We intend the face to give an intuitive visual of valence and arousal, making emotion easier to report.

3.1 Preliminary Study Design

We gathered opinions of the five interface designs to guide future studies on continuous rating. Seven subjects (age: 20-27, mean=24.1, sd=3, gender: 4 male, 3 female) participated, using a Vive Pro Eye headset and Vive controllers. The study consisted of 5 phases and was approved by the university IRB.

Initial Impressions: Subjects first used interfaces in an empty VR world with instructions to play with the trackpad to see how it affected the visuals, then to give their initial impressions of what those visuals meant within the emotion-reporting context. If subjects did not offer opinions, the proctor asked questions like "what emotion do you associate the possible colors with?" or "how do you interpret the interpolation between colors?" to guide responses. After the subject provided enough feedback, the proctor advanced them to the next interface. This phase let us gather initial observations, but also provided context for the training phase.

Training: We considered that subjects would have difficulty reporting in the arousal / valence model in real-time without fairly significant training on the model itself. To mitigate these learning effects, subjects exited VR and were given a 10-15 minute training session on the arousal-valence dimensions and how to give input. Training was done through a PowerPoint presentation; after the proctor explained each dimension, the subject was shown the SAM scales asked to verbally give a number from 1-9 on each scale for the emotion terms angry, elated, content, sad, relaxed, frustrated, depressed, excited, bored, and sleepy. This set was chosen for showcasing terms in each quadrant, as well as some that have were likely to be neutral in one dimension (e.g. "sleepy" may have a neutral valence). If subjects gave an unexpected response (e.g. a high valence for the "depressed") they were given further clarification on the related dimensions. At the end, they were given a physical demonstration of reporting with the controller, and told to imagine that the position of their thumb on the trackpad corresponded with the 1-9 scale they had just used, and that the visuals should somehow reflect this input.

Control Scheme: After training, subjects reentered VR and practiced giving ratings by reading an emotion word and using each interface to report that emotion (similar to the training session). At the same time, they tried each interface with each of the four control scheme, with the control scheme changed by the proctor upon request. Helper images explaining the current control scheme and what the interface visuals meant were shown beside the emotion word. After deciding they had enough practice, they picked the control scheme they preferred for each interface. The preferred control scheme for each interface was carried into the next phase.

Video: Subjects watched 5 60-second 360° videos using the interfaces to continuously rate their emotions. The 5 videos included "Zombie Apocalypse Horror," "Walk the Tight Rope," "War Zone," "Speed Flying," and "Malaekahana Sunrise" from [15], with timestamps taken from [35]. Each video was paired with a different interface at random, and was shown in a random order.

Questionnaire: Finally, subjects exited VR and answered a questionnaire about their experience with the interfaces. Questions included "Which interface felt the most intuitive and why?" "Which interface did you feel you could use most effectively while watching a video?" "Which interface felt the most and least invasive / distracting?" and "For each interface, what did you like and dislike?" They also chose a best interface according to their overall preference.

3.2 Preliminary Results

Table 1 summarizes subjective results. We see a trend of subjects finding the Smiley interface intuitive, minimally invasive, and effective for watching videos. Comments noted that the face's mouth was a clear indication of valence, and that subjects could easily find

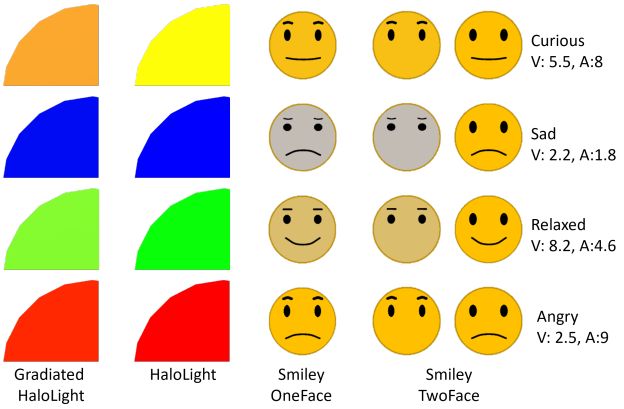

Figure 2: Examples of the 4 interfaces used in the primary study as rendered with certain arousal / valence values. Note that the further into a quadrant the report is, the closer in color HL and GHL are.

most expressions that matched their emotions. Subjects were split on if arousal should be represented by opacity (4) or eyebrows (3). The eyebrows made more sense as an emotional indicator, but could be misinterpreted in too many ways. The opacity was more abstract, but simpler to interpret once subjects knew what it meant. Most (5) subjects preferred two-handed input, saying it was easier to focus on each dimension individually and avoid accidental changes.

Results on HL and GHL suggest that the colors are not universal. During initial impressions, only 2 subjects correctly interpreted the emotions represented by all 4 colors. Those who liked them generally preferred GHL for its increased precision; the 4 colors made subjects feel that they were choosing from only 4 options. Feedback on SAM showed it was intuitive and precise, but only after training, and was large and distracting during video. Feedback on EG showed that the faces were intuitive and subjects liked seeing their input directly mapped to a grid, but showing all of the faces at once was too distracting during video. These results were used to improve designs and continue comparison in the primary study.

4 Primary Interface Design

Given our design considerations for a good continuous rating interface, we moved forward to a more formal study with those interfaces rated the most intuitive and least invasive in the preliminary study. Smiley was chosen for its consistently high ratings. HL was included because it is the most relevant method from the modern literature. GHL was chosen for being generally better rated than HL, and its inclusion allowed a direct comparison with HL. While EG and SAM were considered somewhat intuitive, they were also considered visually invasive and generally unimpressive, and were therefore removed for the primary study.

The designs of HL and GHL were kept consistent in the primary study, and both used the VA control scheme. After incorporating subject feedback, new Smiley designs were tested with a small pilot group, resulting in slight design changes and a new variation as described below. Both used the THY control scheme due to majority preference in the preliminary study. The 4 studied interfaces are shown in Figure 2 with example arousal / valence ratings

Smiley, OneFace (SOF): One primary criticism of the Smiley interface was the vagueness of the arousal representation through opacity and eyebrow movement. In reconsidering research on facial expression dynamics and their link to arousal [6, 22, 30, 33], we adjusted the eyebrow behavior to change scale and arc shape instead of just tilt. Specifically, as arousal is raised, the eyebrow is arched upward (midpoint is higher than endpoints), it becomes larger and is raised higher on the face, and the eyes are narrowed and grow taller. As arousal is lowered, the eyebrow is arched downward (midpoint is lower than endpoints), it becomes smaller and is lowered closer to the eyes, and the eyes widen and shorten. The intent is to make high arousal look surprised or alert and low arousal look bored or tired.

In addition to changing eyebrow representation, we also considered replacing opacity adjustment with saturation. In dark scenes, fully reducing the opacity of the face might render the dark mouth and eyebrow features hard to see, limiting the interface's usefulness and possibly artificially incentivizing higher arousal reports to keep the face visible. This problem is avoided by reducing saturation instead of opacity, with low arousal turning the face from a bright yellow to a duller blue-gray color. In a small informal pilot with 6 participants, feedback indicated that they preferred the saturation change to opacity, and preferred a combination of both saturation and eyebrow effects to only one effect.

Smiley, TwoFace (STF): While the SAM interface had drawbacks in being too visually invasive, users did seem to like having the two dimensions visually separated. We considered that this separation could be replicated on a visually smaller interface by showing two Smileys side-by-side with valence only affecting the right face and arousal affecting the left. Valence and arousal features are removed for the face that doesn't show them. While other interfaces were shown in the bottom right corner of the user's vision, TwoFace was shown in the bottom center of the user's vision to ensure both faces were equally visible.

4.1 Primary Study Design

Our goals for an expanded study were primarily to gather further initial opinions on the meaning of interface visuals, compare the reporting precision of interfaces, compare interfaces when reporting different types of emotions, and collect subjective opinions on the effectiveness of the interfaces for continuous rating during videos. To those ends, we followed a similar study procedure to the preliminary study (Section 3.1) with changes identified below.

4.1.1 Participants and Apparatus

26 undergraduate students (age: 18-32, mean=20.5, sd=3.17, gender: 15 male, 9 female, 2 non-binary) were recruited to participate, 15 being from computer science, 8 from electrical engineering, and 3 from informatics. VR experience varied, with 9 claiming no VR experience, 9 claiming minimal to infrequent use, and 8 claiming frequent use or headset ownership. Subjects answered 5 questions from the Immersive Tendencies Questionnaire [27] , scoring an average 24.1 out of a possible 35. Subjects used an HTC Vive Pro Eye headset and controllers to view the scene and interact. Arousal / valence values were captured at 50 Hz based on thumb input on the controller trackpads according to the control scheme. To make it easier for subjects to report min and max arousal / valence values, we constricted the radius of input to 85% of the trackpad's natural radius. Any input given outside of that was considered reported as the max for that dimension.

All experiment software and interfaces were made in Unity 2019.4.11. Experiments were performed on an Alienware Aurora R13 with an Intel i7-12700KF CPU and GeForce RTX 3080 graphics card. Study methodology was approved by the university IRB.

4.1.2 Study Procedure

The primary study followed a similar 5 phase procedure to the preliminary study. The **Initial Impressions** and **Training** phases were kept identical aside from the use of different interfaces. The third through last phase are described below.

Label Reporting: Similar to the **Control Scheme** phase, subjects practiced giving ratings by reporting an emotion labeled on a screen in the VR environment. To support measuring precision, subjects rated a set of the same 8 words 4 times, with the order of

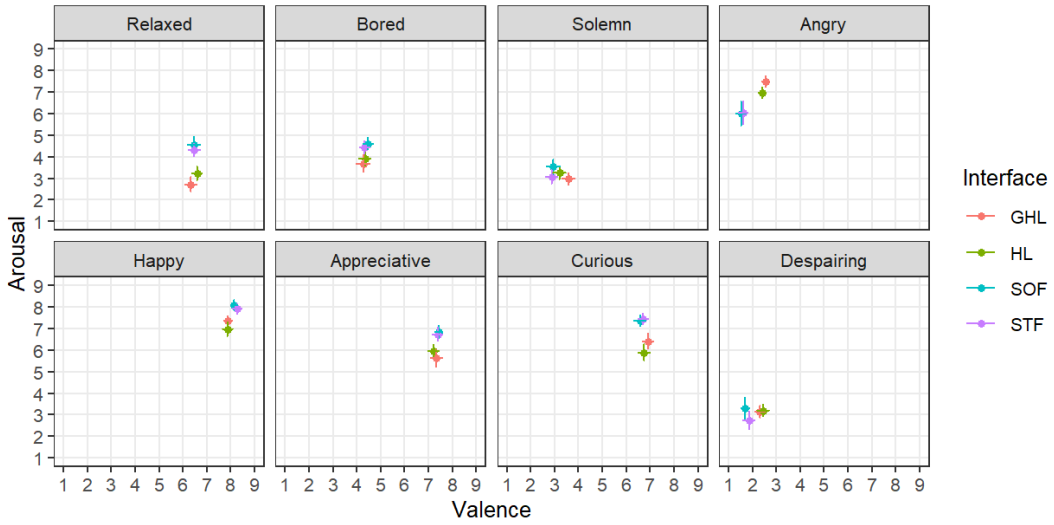

Figure 3: Mean arousal and valence values for each emotion word using each interface with standard error bars for both dimensions.

the words randomly shuffled each time. To avoid confusion, we ensured the first word of a new set was not the same as the last word of the previous set when shuffling. Considering subjects would need practice when using a new interface, the first set showed a helper graphic for the current interface and control scheme and was not counted during analysis.

The 8 chosen words were "Angry," "Curious," "Happy," "Appreciative," "Relaxed," "Solemn," "Bored," and "Despairing," taken from seminal work by Mehrabian [19] in which he provided a large list of words with average arousal and valence ratings given by a population of several hundred students. Four words were chosen for their average ratings falling cleanly into one of the four quadrants, while the other four were chosen chosen for having a roughly neutral average in one dimension. We considered using words with a neutral value in one dimension may produce different results between HL and others, since we speculate that the design of HL promotes simply picking 1 of 4 quadrants.

Video: Similar to the **Video** phase from the preliminary study, subjects watched 4 60-second 360° videos using the interfaces to continuously rate their emotions. The 4 videos included were "Zombie Apocalypse Horror," "Walk the Tight Rope," "War Zone," and "Malaekahana Sunrise" ("Speed Flying" was removed due to this study using 1 fewer interface). Again, each video was randomly paired with a different interface and shown in a random order.

Questionnaire: Before exiting VR, subjects ranked the 4 interfaces by their overall preference using a drag-and-drop ray-based interface. After confirming their ranking with the proctor, they exited VR and answered the same questionnaire as in the preliminary study, minus choosing the best interface.

4.1.3 Study Metrics

The primary independent variable of interest is the interface. In the **Label Reporting** phase, the emotion label being rated was considered an independent variable in order to check for interactions between interface and emotion.

From the **Label Reporting** phase, we first measure the given arousal and valence ratings for differences in reporting behaviors. We then also extract a *precision* and *time-to-report* dependent variable. We consider that the interfaces should be able to be used quickly in order for it to be used in real-time, thus we consider a low time-to-report to be imperative. We also consider that a user should be able to report a single emotion with relative precision over multiple reportings, i.e., if a user wishes to report "happy," their arousal / valence values should be similar each time. To that end, we measure precision as the statistical variance in arousal and valence reports for the last 3 times they report a single emotion, with a lower variance considered more precise.

From the **Questionnaire** phase, we extract a general *ranking*, assigning values 1 through 4 to each interface (1 being best) based on their position in the ranking. We also extract which ratings are considered most *intuitive*, most *effective*, *most invasive*, and *least invasive* from their associated questionnaire items.

Table 2: Two-way ANOVA results, including the corrected degrees of freedom, F and p values, and Cohen's f effect size.

Metric	Variable	DoF	F	p	f
Arousal	Interface	(1.46, 35.15)	4.77	.023	.45
	Interaction	(6.32, 151.58)	4.62	<.001	.44
Valence	Interface	(2.05, 49.1)	3.62	.033	.39
	Interaction	(8.3, 199.21)	2.59	.01	.33
Precision: Arousal	Interface	(2.28, 54.8)	4.23	.016	.42
	Emotion	(2.6, 62.37)	1.5	.226	NA
Precision: Valence	Interface	(1.95, 46.79)	.89	.413	NA
	Emotion	(4.29, 102.9)	.51	.739	NA
Report Time	Interface	(2.23, 53.51)	8.32	<.001	.59
	Emotion	(4.57, 109.71)	8.95	<.001	.61
	Interaction	(9.36, 224.61)	4.42	<.001	.4

4.2 Primary Results

We analyzed precision and time-to-click metrics with two-way repeated measures ANOVA tests using interface and emotion label as within-subject independent variables, followed up with pairwise t tests when appropriate, and used Cohen's f and d to measure effect size. For clarity, ANOVA results are listed in Table 2 and significant pairs are listed in Table 3. We analyzed subjective rankings using a Friedman test and followup Wilcoxon signed rank tests. Finally, we compared counts for other questionnaire metrics with exact multinomial tests and followup binomial tests. In all cases, Holm corrections were applied to p-values to correct for family-wise error.

4.2.1 Label Reporting

Arousal and valence reports from the Label Reporting phase are summarized in Figure 3 for each interface and emotion label. A

Table 3: Significant diffs. between interfaces for indicated metrics.

Metric	Result	t	p	d
Arousal	GHL ≠ SOF	3.45	.003	.244
	GHL ≠ STF	2.53	.037	.179
	HL ≠ SOF	3.73	.001	.264
	HL ≠ STF	2.74	.027	.194
Arousal Precision	HL > STF	2.77	.03	
	SOF > STF	3.21	.009	
Report Time	HL < GHL	7.86	.001	.32
	HL < SOF	6.25	.001	.26
	HL < STF	5.75	.001	.23
	GHL > SOF	2.73	.019	.112
	GHL > STF	2.56	.021	.104

two-way repeated-measures ANOVA measuring effect of interface and emotion label on arousal showed a main effect of interface and a significant interaction between interface and emotion (see Table 2). A similar ANOVA measuring effect on valence showed a main effect of interface and an interaction between interface and emotion.

Pairwise followup t tests between just the 4 interfaces show that HL and GHL arousal reportings both differed from SOF and STF. Followup tests did not show any difference between interfaces for valence reportings after Holm corrections, though there was a possible trend between GHL and SOF ($t = 2.59, p = .062$). This reflects a trend seen in the 2D plot: visually similar interfaces (GHL / HL and SOF / STF) are slightly clustered together for arousal ratings.

Investigating the interaction between interface and emotion label, we visually see different patterns between the interfaces for different words, with some words showing tighter clusters and some showing opposite patterns of which group of interfaces has a higher or lower arousal. Performing pairwise comparisons of arousal reportings between interfaces for each emotion label, we see that for "Angry," GHL ($\bar{x} = 7.49$) incurred a higher reporting than SOF ($\bar{x} = 6, t = 2.65, p = .05, d = .531$) and STF ($\bar{x} = 6.03, t = 2.7, p = .05, d = .54$), while for "Relaxed," GHL ($\bar{x} = 2.72$) incurred a lower reporting than SOF ($\bar{x} = 4.54, t = .3.96, p = .003, d = .792$) and STF($\bar{x} = 4.31, t = 4.07, p = .003, d = .815$). A similar trend of lower arousal for GHL and HL is seen with "Curious." When looking at valence, we only see a difference for "Angry," where GHL ($\bar{x} = 2.56$) and HL ($\bar{x} = 2.42$) incurred a higher valence reporting than SOF ($\bar{x} = 1.53, t = 4.68, p < .001, d = .937$ for GHL, $t = 4.45, p < .001, d = .89$ for HL) and STF ($\bar{x} = 1.61, t = 4.47, p < .001, d = .894$ for GHL, $t = 4.17, p = .001, d = .833$ for HL). This interaction effect suggests that the difference between interfaces is not constant, and could not be fixed with a simple offset. In other words, it suggests that the interface had a real effect on how the subject perceived the emotion they were reporting.

Precision As previously mentioned, we measure precision as the statistical variance in the 3 reportings a subject gave per interface per word. Results are summarized in Figure 4. A two-way repeated-measures ANOVA on arousal reports showed a significant main effect for interface but not emotion label. A similar ANOVA on valence reports showed no significant main effects for either.

Followup pairwise comparisons between interfaces for arousal ratings showed STF incurred a higher arousal reporting variance than HL and SOF. While much of the variance is rather low (medians for all interfaces are below 0.5), we also see a strong positive skew in arousal reporting. This may suggest a possible confusion or carelessness during reporting that is exacerbated by using STF.

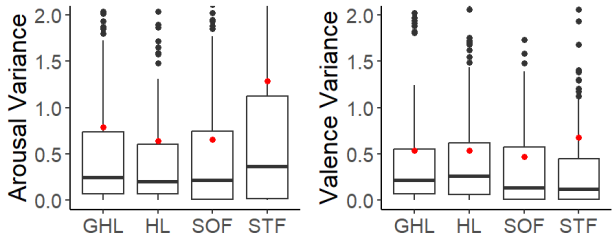

Figure 4: Variance within individual subjects for their 3 reportings given per interface per word during the Label Reporting phase. Red dots denote means.

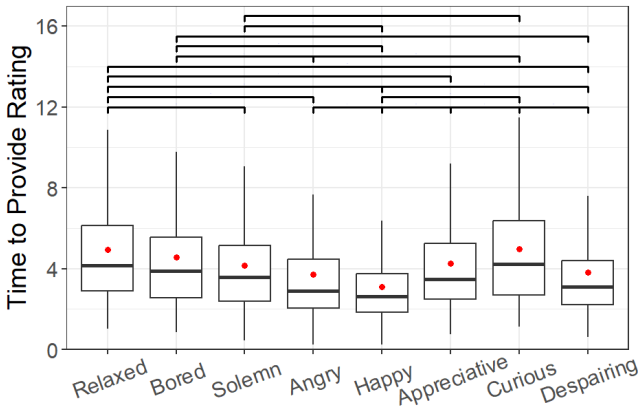

Figure 5: Boxplots summarizing time-to-report for each emotion label with interfaces collapsed. Connecting brackets indicate significantly different pairs. Note outliers are removed for bracket readability.

Time to Report We consider the time taken to provide a reporting for an emotion label a general measure of the speed at which a user can rate an emotion as they feel it. As these interfaces are meant to be used in a real-time context, speed is important to ensure the reports are current with measured physiology. Results for time-to-report are summarized in Figure 6.

A two-way repeated measures ANOVA on time-to-click showed a significant main effect for interface and emotion label, as well as an interaction between the two variables. Looking only at differences between interfaces, with emotion labels collapsed, followup pairwise comparisons show that HL ($\bar{x} = 3.42s$) was faster than GHL ($\bar{x} = 4.72s$), SOF ($\bar{x} = 4.3s$), and STF ($\bar{x} = 4.31s$). GHL was also shown to be slower than SOF and STF. This implies the partial ranking HL is faster than SOF and STF which are faster than GHL.

Figure 5 more directly investigates the times to rate different emotion labels. As shown, many differences are found. For example, "Relaxed" ($\bar{x} = 4.94s$) and "Curious" ($\bar{x} = 4.99s$) took longer to report than 5 other emotion labels each. "Happy" ($\bar{x} = 3.1$), on the other hand, was faster to report than all other emotion labels, beating "Relaxed" and "Curious" by almost 2 full seconds. This may suggest that some emotion words are more familiar and can be quickly mentally mapped to an arousal / valence value, while others may take more thought.

To investigate interactions between interface and emotion label on time-to-report, we performed pairwise followups between each interface for data from each emotion label. Patterns remain similar for most comparisons, with HL generally taking less time than others and GHL generally taking more. GHL took more time than all other interfaces for the words "Curious" and "Despairing," but took less time than SOF and STF for the word "Angry." For the words "Bored" and "Appreciative," no significant differences are found, and the means across interfaces are closer together. This

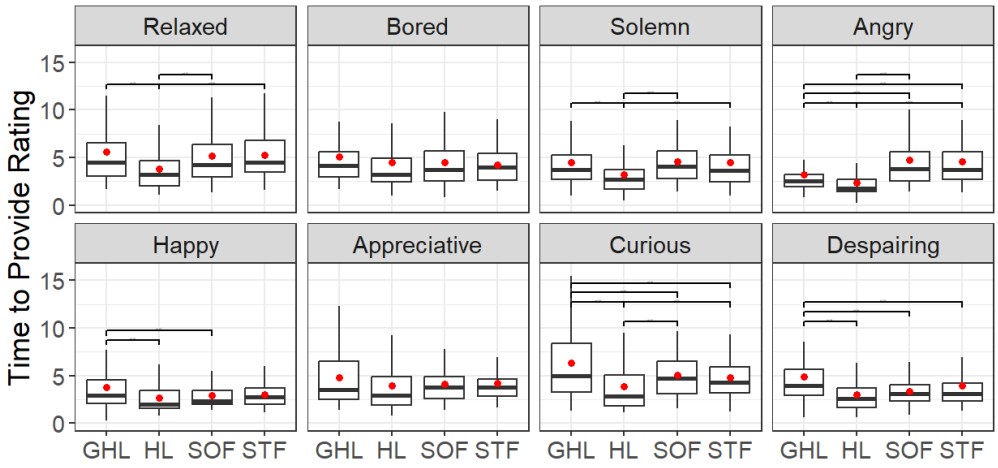

Figure 6: Boxplots summarizing time-to-report in seconds during the Label Reporting phase for each combination of emotion label and interface. Red dots denote means. Connecting brackets indicate significant pairs. Note outliers are removed for bracket readability.

interaction suggests that certain emotions, like "Angry," may be easier to report using GHL than other emotions, and more generally, that certain interfaces can vary in difficulty across emotions.

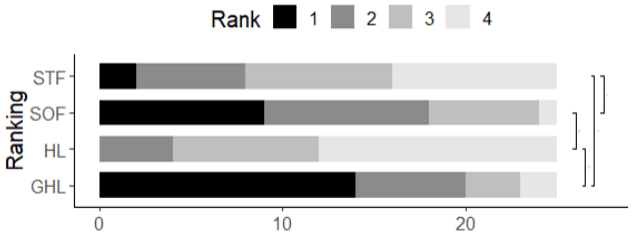

Figure 7: Ranking for preferred emotion rating interface. Bar color indicates how many subjects gave the interface the indicated ranking. A ranking of 1 implies best. Brackets denote significant differences.

4.2.2 Subjective Results

Ranking results from the questionnaire phase are shown in Figure 7. A Friedman test showed a moderate difference between interface rankings $[\chi^2(3) = 27.768, p < .001, W = .37]$. Pairwise followup comparisons showed GHL (median rank 1) ranked higher than STF (median 3, $p = .009$) and HL (median 4, $p < .001$), and SOF (median 2) also ranked higher than STF ($p = .014$) and HL ($p < .001$). This supports a clear partial ranking of GHL and SOF > STF and HL.

Results from the rest of the questionnaire are summarized in Table 4, showing the number of times an interface was chosen for each question. Exact multinomial tests were performed to detect departure from a random uniform selection of interfaces for each question. Results for *most intuitive* showed unequal proportions $[\chi^2(2) = 21.48, p < .001]$, with followup binomial tests suggesting HL was picked fewer times than GHL ($p < .001$) and SOF ($p = .01$) and GHL was picked more times than STF ($p = .017$). This supports a partial ranking for perceived intuitiveness of GHL / SOF >STF / HL. Results for *most effective* also showed unequal proportions $[\chi^2(2) = 17.5, p < .001]$, with followups suggesting HL was picked fewer times than GHL ($p = .005$) and SOF ($p = .003$). Results for *most invasive* again showed unequal proportions $[\chi^2(2) = 6.8, p = .033]$, with followups not showing significant differences between pairs after corrections. Finally, results for *least invasive* showed unequal proportions $[\chi^2(2) = 20.76, p < .001]$, with followups suggesting SOF was chosen more times than HL ($p = .004$) and STF ($p < .001$).

Below we summarize subjective feedback, supported by direct quotes when appropriate, and give a count for *italicized words* representing the frequency of the associated sentiment in the exit questionnaire. HaloLight was liked for its *simplicity* (N=7), with 4 distinct colors ensuring they knew the general emotion they were reporting ("Quadrants make it easy to know where you are" - Subject 19). However, this simplicity also made it more *vague* (N=14) than others, with less feeling of control or precision likely making it less intuitive and effective ("Can't go in between colors when the feeling is in between" - S. 7). Gradiated HaloLight was generally considered the better version ("Just a better version of HL" - S. 10), with the gradient giving increased precision and more colors that could represent a *wider range* (N=12) of emotions. However, the small changes in color after small input changes led to difficulty distinguishing those smaller differences, and the wide range of colors overall *overwhelmed* (N=6) some ("Had to put too much thought in to distinguish collors" - S. 9). Despite these concerns, it was rated as highly intuitive and effective.

Feedback for the OneFace Smiley suggested subjects liked the added nuance of facial features, and the natural mapping of emotional qualities onto dynamic facial features was more *intuitive* (N=8) for some than color association ("I can match it to my own face and not have to think about color" - S. 16). The main complaint was that the simplicity of the face its possible expressions did not express the *full range* (N=8) of emotions. "Simpler emotions [(e.g. happy)] are easy to associate with facial expressions, but more complex emotions [(e.g. appreciative)] feel more abstracted" - S. 17.

Feedback for the TwoFace Smiley suggested subjects did not like the *separation* (N=16) of the two dimensions. While this was perceived as a positive quality of the SAM interface in the preliminary study, it appears this did not translate well onto a face visual ("Kind of hard to fuse the two faces" - S. 1, "Seemed better to just have one face" - S. 7). As such, it was not perceived as very intuitive or effective, and received the highest selection for most invasive.

5 Discussion

Our primary goal in this work was to explore the design space of continuous emotion rating interfaces and compare examples to discover useful design elements. The HaloLight interface, taken from prior work, was the simplest design we tested, associating the 4 quadrants of the arousal / valence spectrum with 4 colors and adjusting color based on the user's input. Because this design does not give feedback for changes within quadrants beyond a general "intensity," we expected it would be quicker to use but result in less precision when trying to report the same emotion multiple times.

Analysis showed that HaloLight was faster to use, with small-to-moderate effect sizes and an average difference of about 1 full

Table 4: Summary of questionnaire results, indicating the number of times an interface was chosen as the most intuitive (Int), most effective for watching videos (Eff), most invasive / distracting (MI), and least invasive / distracting (LI), and the most commonly given pros and cons.

	Int	Eff	MI	LI	Pros	Cons
GHL	14	11	8	6	Most precise, wider variety, gradients more accurately describe emotions	Hard to distinguish between small gradients, easy to lose position, should be a limit on colors
HL	0	0	7	2	Minimal options make it simple and quick	Vague, hard to report emotions outside quadrants, harder to control
SOF	10	12	1	17	Simple and intuitive, visual is relatable for most emotions, nuanced	Facial features don't express the full range of emotions, processing facial expression took mental effort
STF	2	3	10	1	Two faces helps distinguish dimensions	Having to fuse both faces took mental effort, not intuitive to think of emotion dimensions separately

second, but did not show significant differences in precision, with mean and median values similar to the others. This indicates a usefulness in situations where very fast reporting is of utmost importance. However, subjects did not favor this technique, with it receiving the lowest median ranking, no subjects naming it most intuitive or effective, and many citing general vagueness or lack of confidence using it. This, then, may suggest there is a more appropriate approach for building user confidence and favorability.

Gradiated HaloLight was designed to mitigate concerns about HaloLight's precision by allowing the colors to gradiate across different inputs. While this did not appear to significantly improve precision, it does appear that it raised user confidence; it had the highest median subjective rank, significantly outranked HaloLight, and was rated as more intuitive and effective. However, this improved confidence appears to come at a time cost, with Gradiated HaloLight being slower for emotions like "Happy," "Curious," and "Despairing." Based on subject feedback, we expect this is because of difficulty distinguishing between small gradients in color. Based on final positions in Figure 3, we consider it likely that subjects quickly moved to a similar area as with HaloLight, then spent extra time performing corrective motion [32] to refine the position based on the extra feedback information. This may warrant further investigation into the motion performed during rating.

Smiley was designed to provide a more intuitive interface exploiting emotional associations with facial expressions instead of colors. The OneFace version appears to have achieved this, being frequently picked as the most intuitive and least invasive, and having the highest median rank among the interfaces. During the initial observations phase, all subjects immediately grasped that adjusting the curve of the smile meant adjusting some kind of happiness value, while not all subjects immediately grasped the meaning of each color for HaloLight. Mean and median precision values and rating times were also generally on par with those of Gradiated HaloLight.

The TwoFace version of Smiley was designed to emphasize the separation of arousal and valence dimensions, since this was pointed out as a positive aspect of the SAM interface in the preliminary study. However, this was not appreciated by subjects and, in combination with confusing arousal representation, may have hampered precision for the arousal dimension. In all comparisons with statistical significance, the OneFace Smiley outperformed the TwoFace version, suggesting OneFace should be used instead moving forward.

Results on the differences between emotion labels may reveal more general information about the process of rating emotion with these kinds of interfaces. As seen in Figure 5, certain emotions take less time overall to provide a rating for, e.g., "Happy" is rated almost 2 seconds faster than "Relaxed" or "Curious." This suggests that some elicited emotions may take more mental effort and time to adjust reporting, which may need to be taken into account when associating physiological data with a continuous self-report. Given that HaloLight performed faster than all other interfaces for both of these words (see Figure 6), it again may be the preferred interface when such emotions are being elicited and rating time is considered paramount. However, if emotions like "Happy" or "Bored" are being elicited, an interface like Smiley may be used with similar results.

6 Conclusion

This work explored designs for time-continuous emotion rating interfaces. A preliminary study introduced 2 novel techniques, the Smiley and Gradiated HaloLight, and roughly compared 5 interfaces to guide selection for a larger study. The larger study introduced the TwoFace Smiley and compared the 2 Smiley and 2 HaloLight interfaces. Studies compared interfaces by their time and precision when rating certain emotion words and by subjective opinions.

Results indicate that the OneFace Smiley and Gradiated HaloLight were significantly preferred to the HaloLight from prior work, being ranked higher and named more intuitive and effective for use in real-time. HaloLight was shown to incur lower reporting times for most emotion labels, suggesting a possible advantage in real-time context. Results on the differences between different emotion labels give further insight into the nature of reporting emotion. Certain emotion words took less time to rate than others, possibly suggesting more mental load to describe certain emotions. Certain emotions labels were also reported with different values when using different interfaces, with HaloLight and Smiley variants clustered closer together. Interactions between interface and emotion labels suggest a real difference in how emotions are reported on different interfaces and not one that can be calibrated for.

When considering applications for these interfaces, we imagine two main types. First, an application which wants to quickly understand a user's response (e.g. market testing) could consider the Smiley for its intuitive and minimally invasive design that is understandable with minimal training. Second, research applications aiming to train AI models to recognize a wide range of emotions could consider the Gradiated HaloLight for its perceived wider range and intuitiveness after training. However, researchers collecting time-series data should anticipate the types of emotions elicited and understand that reports for more "complex" emotions may be made later than reports for "simpler" ones.

The exploratory nature of this study created several limitations. Because we wanted to assess our designs' impact for knowledgeable users, subjects received thorough training before assessment, and thus our results may not apply to an untrained user. Our focus on isolating repeated and precise reports meant we had to limit subjects' experience with videos to only 4 brief clips, thus limiting broader implications for other natural stimuli these interfaces are meant to be used with. Finally, because we wanted to compare with existing interfaces, our designs were mostly inspired by 2D imagery, but 3D indicators or interactables may be desirable in VR.

Future works could address these limitations with more targeted studies. Considering the remaining ambiguity in the Smiley's arousal representation, and other existing struggles to represent arousal [6, 30], more investigation could should be done to find the best facial representation for the dimension. Future designs could also make more use of the 3D nature of VR, expanding the range of input or visualization. Most importantly, future work should more directly investigate the interfaces' usability in natural video stimuli.

References

[1] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6:209–222, 7 2015. doi: 10.1109/TAFFC.2015.2392932 2

[2] F. Bevilacqua, H. Engström, and P. Backlund. Game-calibrated and user-tailored remote detection of stress and boredom in games. *Sensors (Switzerland)*, 19(13):1–43, 2019. doi: 10.3390/s19132877 1

[3] P. Bota, J. Brito, A. Fred, P. Cesar, and H. Silva. A real-world dataset of group emotion experiences based on physiological data. *Scientific Data*, 11(1):1–17, 2024. 2

[4] G. H. Bower. Affect and Cognition. *Philosophical Transactions of the Royal Society; Series B*, 302(1110):387–402, 1983. 1

[5] M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25:49–59, 1994. 1, 2, 3

[6] J. Broekens and W. P. Brinkman. Affectbutton: A method for reliable and valid affective self-report. *International Journal of Human Computer Studies*, 71:641–667, 2013. doi: 10.1016/j.ijhcs.2013.02.003 1, 2, 4, 8

[7] A. Bruun, E. L.-C. Law, M. Heintz, and P. S. Eriksen. Asserting real-time emotions through cued-recall: Is it valid? In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, article no. 37, 10 pages. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2971485.2971516 1

[8] L. B. Cadet and H. Chainay. How preadolescents and adults remember and experience virtual reality: The role of avatar incarnation, emotion, and sense of presence. *International Journal of Child-Computer Interaction*, 29:100299, 2021. doi: 10.1016/j.ijcci.2021.100299 2

[9] K. Fayn, S. Willemsen, R. Muralikrishnan, B. Castaño Manias, W. Menninghaus, and W. Schlotz. Full throttle: Demonstrating the speed, accuracy, and validity of a new method for continuous two-dimensional self-report and annotation. *Behavior research methods*, 54:1–15, 2022. 1, 2

[10] J. M. Girard. Carma: Software for continuous affect rating and media annotation. *Journal of open research software*, 2(1), 2014. 1, 2

[11] J. M. Girard and A. G. C. Wright. Darma: Software for dual axis rating and media annotation. *Behavior research methods*, 50:902–909, 2018. 1, 2

[12] M. Gnacek, L. Quintero, I. Mavridou, E. Balaguer-Ballester, T. Kostoulas, C. Nduka, and E. Seiss. Avdos-vr: Affective video database with physiological signals and continuous ratings collected remotely in vr. *Scientific Data*, 11(132), 2024. 2

[13] Q. Guimard, F. Robert, C. Bauce, A. Ducreux, L. Sassatelli, H.-Y. Wu, M. Winckler, and A. Gros. Pem360: a dataset of 360° videos with continuous physiological measurements, subjective emotional ratings and motion traces. In *Proceedings of the 13th ACM Multimedia Systems Conference*, MMSys '22, 7 pages, p. 252–258. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3524273.3532895 2

[14] S. Koelstra, C. Mühl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3:18–31, 2012. doi: 10.1109/T-AFFC.2011.15 2

[15] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams. A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Frontiers in psychology*, 8:2116, 2017. 2, 3

[16] D. Liao, L. Shu, G. Liang, Y. Li, Y. Zhang, W. Zhang, and X. Xu. Design and Evaluation of Affective Virtual Reality System Based on Multimodal Physiological Signals and Self-Assessment Manikin. *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, 4(3):216–224, 2020. doi: 10.1109/JERM.2019.2948767 2

[17] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, and G. Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8(1):1–15, 2018. doi: 10.1038/s41598-018-32063-4 2

[18] I. B. Mauss, L. McCarter, R. W. Levenson, F. H. Wilhelm, and J. J. Gross. The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, 5(2):175–190, 2005. doi: 10.1037/1528-3542.5.2.175 1, 2

[19] A. Mehrabian. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Oelgeschlager, Gunn & Hain, Cambridge, 1980. 5

[20] A. Mehrabian and J. A. Russell. *An approach to environmental psychology.* the MIT Press, 1974. 1, 2

[21] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):1–16, 2020. doi: 10.1038/s41597-020-00630-y 2

[22] P.Ekman and W.V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. 2, 4

[23] R. Plutchik. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pp. 3–33. Academic Press, 1980. doi: 10.1016/B978-0-12-558701-3.50007-7 2

[24] R. Plutchik. The nature of emotions. *American Scientist*, 89(4):344–350, 2001. 2

[25] J. A. Russell. Evidence of convergent validity on the dimensions of affect. *Journal of Personality and Social Psychology*, 36(10):1152–1168, 1978. doi: 10.1037/0022-3514.36.10.1152 2

[26] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980. 2

[27] S. Rózsa, R. Hargitai, A. Láng, A. Osváth, E. Hupuczi, I. Tamás, and J. Kállai. Measuring immersion, involvement, and attention focusing tendencies in the mediated environment: The applicability of the immersive tendencies questionnaire. *Frontiers in Psychology*, 13, 2022. doi: 10.3389/fpsyg.2022.931955 4

[28] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. A review of emotion recognition using physiological signals. *Sensors (Switzerland)*, 18(7), 2018. doi: 10.3390/s18072074 2

[29] N. S. Suhaimi, C. T. B. Yuan, J. Teo, and J. Mountstephens. Modeling the affective space of 360 virtual reality videos based on arousal and valence for wearable eeg-based vr emotion classification. In *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*, pp. 167–172, 2018. doi: 10.1109/CSPA.2018.8368706 2

[30] A. Toet, D. Kaneko, S. Ushiama, S. Hoving, I. de Kruijf, A.-M. Brouwer, V. Kallen, and J. B. F. van Erp. Emojigrid: A 2d pictorial scale for the assessment of food elicited emotions. *Frontiers in Psychology*, 9, 2018. 2, 3, 4, 8

[31] J. W. Woodworth and C. W. Borst. Design and validation of a library of active affective tasks for emotion elicitation in vr. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 398–407, 2024. doi: 10.1109/VR58804.2024.00061 2

[32] R. S. Woodworth. Accuracy of voluntary movement. *The Psychological Review: Monograph Supplements*, 3(3):i, 1899. 8

[33] Y. Wu, Y. Wang, S. Jung, S. Hoermann, and R. W. Lindeman. Using a fully expressive avatar to collaborate in virtual reality: Evaluation of task performance, presence, and attraction. *Frontiers in Virtual Reality*, 2, 2021. doi: 10.3389/frvir.2021.641296 4

[34] T. Xue, A. E. Ali, and T. Zhang. Rcea-360vr: Real-time, continuous emotion annotation in 360? vr videos for collecting precise viewport-dependent ground truth labels. Association for Computing Machinery, 5 2021. 1, 2, 3

[35] T. Xue, A. E. Ali, T. Zhang, G. Ding, and P. Cesar. Ceap-360vr: A continuous physiological and behavioral emotion annotation dataset for 360 vr videos. *IEEE Transactions on Multimedia*, 2021. 1, 2, 3

[36] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar. Weakly-supervised learning for fine-grained emotion recognition using physiological signals. *IEEE Transactions on Affective Computing*, 14(3):2304–2322, 2023. doi: 10.1109/TAFFC.2022.3158234 2