RankTrace: Relative and Unbounded Affect Annotation

Phil Lopes
Computer Vision and Multimedia Lab
University of Geneva
louis.p.lopes@um.edu.mt

Georgios N. Yannakakis
Institute of Digital Games
University of Malta
georgios.yannakakis@um.edu.mt

Antonios Liapis
Institute of Digital Games
University of Malta
antonios.liapis@um.edu.mt

***Abstract*—How should annotation data be processed so that it can best characterize the ground truth of affect? This paper attempts to address this critical question by testing various methods of processing annotation data on their ability to capture phasic elements of skin conductance. Towards this goal the paper introduces a new affect annotation tool, *RankTrace*, that allows for the annotation of affect in a continuous yet unbounded fashion. *RankTrace* is tested on first-person annotation lines (traces) of tension elicited from a horror video game. The key findings of the paper suggest that the *relative* processing of traces via their mean gradient yields the best and most robust predictors of phasic manifestations of skin conductance.**

I. Introduction

Emotion annotation is the most laborious yet, arguably, the most critical task within the affective computing field. The areas of affect modeling, affect expression and affect-driven adaptation are all dependent on appropriate labels of affect. Regardless of the task at hand, special care must be taken on how measurable estimates of affect—such as labels or values—are collected and analyzed. The *validity* and *reliability* of such estimates is naturally questioned given the numerous factors contributing to deviations between an annotator's label and the actual underlying affective state. These factors include the annotator's experience, the affect representation chosen, the design of the annotation tool, person-dependent annotation delays, and the annotation analysis followed [1], [2].

When it comes to analyzing the obtained labels, the dominant practice in continuous affect annotation is to use the data in an *absolute* and direct manner. For example, averaging annotation values (across specified time windows) as obtained from continuous annotation tools such as FeelTrace [3] or Gtrace [4] is a common practice in the literature [5]–[7]. Since data is typically treated in an absolute fashion, it becomes a necessity to constrain the annotator within certain bounds (e.g. arousal values lie within [0,1]) so that the statistical analysis becomes feasible. After all, a common scale is required if the provided annotation data is analyzed as interval or absolute values. We argue that these practices are *detrimental* to the collection and analysis of annotation signals as they factor in a multitude of subjective annotation biases that yield both constant (lack of validity) and variable (lack of reliability) deviations from the underlying ground truth [8], [9].

We are instead inspired by the principle of *ordinal*, or *relative*, annotation as followed by an increasing number of studies in affective computing [2], [8]–[10] and theories of psychology such as the *adaptation level theory* [11]. The latter suggests that humans cannot maintain a constant value about subjective notions; instead, their preferences are made on a pairwise comparison basis using an internal ordinal scale [12], [13]. The relative nature of emotions, and naturally their annotation, is also supported by relative judgment models [14], [15] suggesting that our experience with stimuli gradually creates our internal context, or else *anchors* [16], against which we rank any forthcoming stimulus or perceived experience. Finally, we are motivated from several studies in affective computing [8] that have already showcased the advantages of relative annotation and processing for higher inter-rater reliability in video annotation [2], and affect model generality for sound [17], music and speech [18].

In this paper we introduce a continuous affect annotation tool that—similarly to Gtrace [4]—is built on the principles of single-dimension annotation. The proposed *RankTrace* tool, however, a) does not constrain the user within bounds and b) inspired by [19] it uses a wheel-like hardware as a more natural means of user interfacing with continuous annotation tasks. Using this tool 14 players of the *Sonancia* horror game [20] annotate their tension levels by watching their video-captured playthroughs; during the playthrough their skin conductance (SC) was measured via a bracelet. We consider SC as the ground truth against which we test annotation data. For that purpose we extract two *phasic* and *tonic* features via continuous decomposition analysis [21] of the SC signal. We focus on phasic activations as they are associated to manifestations caused by external stimuli [21], [22].

Using the obtained annotation and ground truth data, we test the hypothesis that treating continuous annotations in a relative fashion is a beneficial approach for approximating the ground truth. We thus compare two approaches of analyzing continuous annotations which evaluate the *absolute* values (mean and integral within a time window) against two *relative* approaches based on changes in the signal (amplitude and average gradient). Results from rank correlation analysis show that annotation features which assess relative annotation changes within the window are better and more robust linear predictors of the phasic driver of SC. Our findings validate our hypothesis and suggest that treating a continuous annotation signal in a relative fashion, e.g. via the its gradient, yields

linear predictors of higher predictive capacity.

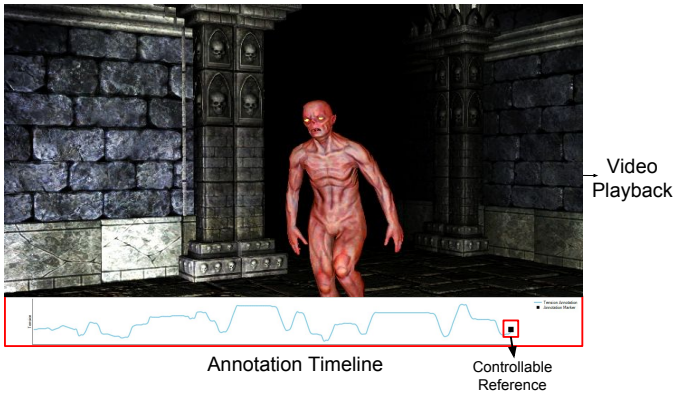
This paper is novel in two critical ways. First, it introduces a new annotation tool, *RankTrace*, that allows for efficient first- or third-person continuous annotation. The tool is largely inspired by [19] but it is enhanced for affect annotation by offering flexible annotation beyond predetermined bounds. Further, the paper offers a new approach for the analysis of continuous annotation—may it be video, sound, speech or gameplay annotation—that is based on the *relative* rather than the *absolute* processing of the obtained data.

II. Emotion Annotation

Manually annotating emotion is a challenging task, the complexity of which depends on both the annotators involved and the annotation protocol designed. Both the *validity* and the *reliability* of the provided annotations need to be questioned since the first is associated with the degree to which the annotation task per se captures the underlying affect whereas the latter is associated with the degree to which the obtained data is consistent. While annotation reliability—in particular inter-rater reliability—has been the focus of several studies in affective computing [2], this paper does not examine the impact of *RankTrace* on reliability. Given that our chosen domain is games, we wish to obtain annotations about gameplay experience directly from the player. The result is that each game session is annotated by a single annotator (i.e. the player) in a first-person fashion. Instead, this paper focuses on the validity of the provided annotations and tests the degree to which different methods for processing continuous annotations yield predictors of the underlying ground truth as manifested by the phasic drivers of skin conductance.

Representing changes of affect over time as a continuous function has been among the dominant annotation practices within affective computing over the last 20 years. Continuous annotation is advantageous compared to discrete states, as labels used in discrete states have fuzzy boundaries that lead to inter-rater disagreements [23]. The dominant approach in continuous annotation uses the arousal-valence circumplex model of affect [24]. Several tools allow the continuous labeling of affective dimensions: free software such as FeelTrace [3] and its variant GTrace [4] are popular for real-time video annotation, but other options for annotating video [25] and music [26] exist. Such continuous annotation tools require substantial cognitive effort from users and may lead to low inter-rater agreement and generally unreliable annotations [27], [28].

To counter some of the subjective biases of continuous annotations, converting the raw values of annotated affective dimensions into ranks can already lead to a higher inter-rater agreement. *AffectRank* [2] has demonstrated that rank-based video annotations leads to significantly higher inter-rater reliability compared to FeelTrace. This paper, however, does not perform inter-rater comparisons, as gameplay videos are annotated exclusively by the players that produced them in a first-person manner.



(a) Annotation with *RankTrace*: participants watch a recorded playthrough of *Sonancia* (top) while annotating. The entire annotation trace is shown (bottom) for the participant's own reference, acting as the anchor of their annotation.

(b) The *Griffin PowerMate* wheel interface is used along with *Ranktrace*.

Fig. 1. The *RankTrace* tool (Fig. 1a) allows participants to annotate their emotional experience using the *PowerMate* controller (Fig. 1b) in real-time, while watching a video of their playthrough.

III. The *RankTrace* Annotation Tool

Inspired by [2], [19] we developed *RankTrace*, a new annotation tool for the purpose of reliably approximating the ground truth of affect via relative continuous annotation. The core idea behind the *RankTrace* tool was introduced by [19]: the tool allows participants to watch the recorded playthrough of a play session and annotate in real-time the perceived intensity of a single emotional dimension (see Fig. 1a). *RankTrace* provides 4 annotation samples per second and its interfacing is similar to GTrace [4]. The annotation process in *RankTrace*, nevertheless, is controlled through a "wheel-like" hardware (see Fig. 1b), allowing participants to meticulously increase or decrease emotional intensity by turning the wheel, similarly to how volume is controlled on a stereo system. Unlike the tool presented in [19] and other continuous annotation tools such as FeelTrace [3] or Gtrace [4], annotation in *RankTrace* is *unbounded*: participants can continuously increase or decrease the intensity to their desire without constraining themselves to an absolute scale. This design decision is built on the anchor [16] and adaptation level [11] theories by which affect is a temporal notion based on earlier experience, baselines or current context that is best expressed in relative terms [2]. Several participants during the piloting of a bounded *RankTrace* version expressed the will to further increase (or decrease) the annotation value beyond its limits, confirming our hypothesis. With our unbounded approach, instead, participants may work with a broader range of emotional intensity; as broad as they may wish.

For the purposes of this study and the aims of the horror game genre, *RankTrace* is used for the annotation of *tension*.

While tension and arousal have been used interchangeably in some affective models we follow the model of Schimmack and Grob [29] that represents affective states via the dimensions of tension, arousal and valence.

IV. Testbed Game and Protocol

All first-person annotation experiments presented in this paper are performed on the *Sonancia* [17], [20] game generation system. According to our protocol, human participants play a horror game level, then annotate the video captured during their playthrough. Game events and skin conductance are logged while participants play as discussed in Section V. This section describes the game and the experimental protocol.

A. *The* Sonancia *Horror Game*

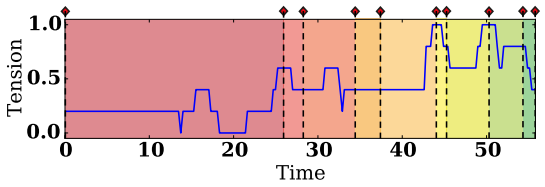
Sonancia is a generative system which creates playable levels for a horror game. In this game, players explore a haunted manor (i.e. the level) in order to find an objective located within one of its many rooms. Reaching and activating the objective ends the level. The level consists of different rooms, separated by walls and connected by doors. Rooms can contain monsters, light sources and the objective; each room also has its own background soundtrack, allocated through a sonification process. Players are unarmed and must avoid direct confrontation with hostile monsters; monsters act as an instigator of tension and fear. For the interested reader, the level design algorithms use a tension model as specified by the designer [20] or based on crowdsourced models of tension [17] for placing sounds in rooms. This allows *Sonancia* to create new levels with unique sound combinations. In this study, however, two pre-generated levels are used by all players.

B. *Experimental Protocol*

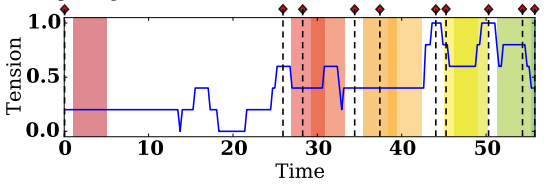
To test *RankTrace*, an experimental protocol was designed to allow participants to first play *Sonancia* and then annotate their playthroughs. Each participant was first introduced to the experiment and then answered some demographical questions e.g. about age and gender. Afterwards, the physiological devices were synchronized to the game and the participant began playing a level of *Sonancia*; once they completed the level, they were asked to watch a video capture of their most recent playthrough while annotating tension via the process described in Section III. This was repeated three times, as each player played three variants of the same level using three different sonification methods. Special care was taken so that the distribution of the two levels, and the order of variants played was fair among participants to avoid biases in the data.

V. Data Collection and Feature Extraction

The data obtained from the devised experiment includes in-game events and skin conductance signals collected while the game was played, and self-reported continuous annotations of perceived tension via *RankTrace*, collected during a playback video of the player's game. A total of 41 annotated playthroughs were collected from 14 different participants (9 male; 5 female). This section covers how this data was processed for analysis.



(a) Splitting the annotation trace into continuous windows.



(b) Splitting the annotation trace into reactive windows.

Fig. 2. Indicative splitting of the normalized annotation trace of tension (blue line) into windows of different colored backgrounds. The events that trigger frame changes are room changes (rhombi and dotted lines). For reactive windows, the window starts 1 sec after the event and ends after 5 sec.

A. *Annotations*

As mentioned earlier, once each level variant was complete, participants were tasked to annotate data using *RankTrace*. The output of *RankTrace* is a continuous representation (trace) of perceived tension, during a segment of game-play (see Fig. 1). Following the literature [2]–[4], we split the signal into several *time windows*, or frames, from which we extract statistical features. A frame is a subset of the trace capturing the perceived emotion elicited through a gameplay event. Fig. 2 showcases the two window framing methods employed in this paper. The first method is referred to as a *continuous window*, where signal parsing starts immediately after a gameplay event occurs and continues until another event occurs. In this paper, a triggering event is when a player enters a room (coinciding with a change in the background music). The second method is referred to as a *reactive window*, according to which parsing starts 1 second after a game event occurs and ends 5 seconds after. This methodology builds on the assumption that most participant annotations are in reaction to occurring (tense) events—or else have a phasic nature. It also takes into consideration the annotator's potential time lag [1], [3]. As reactive time windows have a constant duration, windows may overlap if players swiftly change rooms.

Once either framing technique is applied, statistical features are extracted from each available window. In this paper we extract and compare four *annotation metrics* (see Fig. 3): a) the mean value (μA); b) the area of a window, i.e. its composite trapezoidal integral ($\int A$), normalized to the window's duration; c) the amplitude (max-min difference) ($\hat{A}$); d) the average gradient (ΔA). The first two features extract annotation values in an *absolute* fashion. The latter two features are independent of the absolute value of the annotation but instead measure *relative* changes within the window.

B. *Skin Conductance*

As the current study focuses on horror games which target negative emotions such as stress and fear, skin conductance

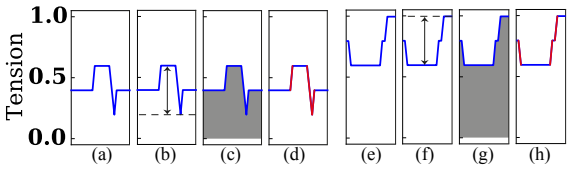


Fig. 3. Processing of two indicative time windows from Fig. 2a. The μA is 0.45 in 3a and 0.7 in 3e, as the average value of the approximately 30 data points in these windows. Calculation of $\hat{A}$ is shown in 3b and 3f based on the maximum minus the minimum values in that window (0.4 in both cases). The integral is calculated based on the area under the trace in Fig. 3c (0.11) and 3g (0.17), normalized to the window's duration. The average gradient calculates the difference of adjacent data points, which is non-zero for the red parts of Fig. 3d and 3h; note that ΔA is 0 for 3d as there are equal positive and negative gradients which cancel each other out.

(SC) can be safely considered as a reliable manifestation and ground truth for these specific affective states [30]–[32]. Moreover, SC monitoring devices are particularly easy to setup and non-intrusive for participants. For the above reasons SC was used as ground truth for validating annotated experience (tension) during play. SC is monitored through *Empatica*'s E4 device [33], which consists of a bracelet-like apparatus akin to a wristwatch connected to the computer via bluetooth. SC in E4 is measured in μS (micro Siemens), sampled at 4Hz: high μS values indicate high arousal (i.e. high conductance), while low μS values indicate low arousal (i.e. low conductance).

Skin conductance signals are characterized by two different types of activity, *tonic* and *phasic*. Tonic SC refers to the phenomenon of slow changing variation of the signal through time, considered to be the level of SC in the absence of external events or stimuli. Phasic skin activity, instead, is the abrupt increase of SC levels occurring within short-term event intervals. These typically occur after an environmental event or stimulus [21]. In this paper we use the Continuous Decomposition Analysis (CDA) approach [21] to decompose our SC signals into continuous tonic and phasic activity. We can derive an estimate of tonic activity by sampling the signal at defined intervals, presuming the SC signal is stable. Phasic activity can then be extracted by simply subtracting the tonic activity, resulting in what is called a *phasic driver* expressed in μS. The phasic driver consists of a baseline corrected measure, capable of capturing the affect of a given stimulus. The stimulus-response window for SC typically ranges around intervals of $[1, 3]$ to $[1, 5]$ seconds after a stimulus event [21]. For the purposes of this paper, SC statistical features are extracted within a defined response window of $[1, 4]$ seconds after the occurrence of a stimulus event. Inspired by [21], [22], [32] we extracted two SC features that are considered appropriate manifestations of stress, tension or arousal and can form reliable ground truths for them: a) the mean phasic driver within the response window (μP_d); and b) the integral of the phasic driver within the response window ($\int P_d$). We explicitly do not investigate the tonic component in this paper as the emphasis is on game event-based manifestations of tension and stress; thus the phasic component, by nature, defines a more accurate approximation of the underlying ground truth.

To reduce the noise of the raw SC signals, a Gaussian smoothing function is applied on each SC signal before applying CDA. Only valid SC signals, which presented a stable continuous signal, were taken into consideration in this paper. After pruning, 40 game sessions were considered, with an average duration of 95.74 seconds each ($\sigma = 44.11$).

VI. Results

This section explores the impact of the four different ways of extracting information from *RankTrace* annotations—as described in Section III—on predicting the phasic components of SC. For all the experiments presented below we follow the relative analysis approach proposed in [10] and we derive a ranked order of time windows with respect to each of the four annotation metrics described in Section V-A. We do that in two different degrees of *annotator memory*: a) *all windows* assumes that the annotator maintains her anchors throughout the experiment and, thus, considers all possible window pairs for deriving a global rank per trace; b) *adjacent window* assumes that the memory of the annotator is limited to one time window and thus considers only adjacent windows to derive the global rank of the annotation metrics.

In the following sections we explore the predictive capacity of the annotation metrics with respect to three dimensions: a) the two different windowing methods (continuous vs. reactive); b) the two degrees of annotator memory (all windows vs. adjacent window), and c) the impact of min-max normalization of the raw annotation values versus using the raw traces. In this paper we rely only on linear predictive capacity (i.e. correlations) and we do not explore non-linear machine learning processes which are left for future analysis.

A. Raw Data

We first investigate the predictive capacity of all annotation metrics when annotation data is not normalized, to establish a baseline of the most challenging scenario in terms of the relativity of data. Unprocessed annotation data likely maintain all the subjective reporting biases we already mentioned; we wish to explore how much those biases may affect correlation between our annotation metrics and our ground truth.

Table Ia shows the rank correlations between all video annotation metrics (on raw data) and the phasic driver features considered. At first observation the ΔA metric appears to be the best and more robust predictor. It not only yields the highest correlation values but it also manages to provide significant effects with both SC phasic features across both window types (continuous and reactive) when ranking all windows. It also yields significant positive correlations with μP_d when ranking only adjacent windows. Beyond ΔA we observe that the annotation amplitude is also a good predictor of phasic elements of the SC but only when we assume that the annotator maintains reference points across all annotation windows (i.e. ranking all windows). In this extreme case of treating annotation data in an absolute fashion, findings suggest that a relative measure (ΔA) can be a reliable linear

TABLE I

RANK CORRELATION OF ANNOTATION VALUES AND THE PHASIC DRIVER FEATURES OF SC, COMPUTED ACROSS WINDOW TYPES (CONTINUOUS VS. REACTIVE) AND ANNOTATOR MEMORY (ALL WINDOWS VS. ADJACENT WINDOWS). SIGNIFICANT VALUES ARE IN BOLD [0.05 (*) AND 0.01 (**)].

		All Windows		Adjacent Windows	
	Metrics	μP_d	$\int P_d$	μP_d	$\int P_d$
Continuous	μA	0.006	0.007	-0.049	-0.046
	$\int A$	0.007	0.009	-0.048	-0.057
	$\hat{A}$	**0.057***	**0.045***	0.101	0.041
	ΔA	**0.101****	**0.082****	**0.170***	0.053
Reactive	μA	0.040	0.027	-0.032	-0.039
	$\int A$	0.038	0.026	-0.07	-0.05
	$\hat{A}$	0.035	0.032	0.032	-0.003
	ΔA	**0.116****	**0.094****	**0.165***	0.110

(a) Raw Data

		All Windows		Adjacent Windows	
	Metrics	μP_d	$\int P_d$	μP_d	$\int P_d$
Continuous	μA	0.008	0.007	0.06	-0.074
	$\int A$	0.029	0.013	-0.015	-0.072
	$\hat{A}$	**0.057***	**0.045***	0.027	0.013
	ΔA	**0.135****	**0.106****	**0.172***	0.054
Reactive	μA	0.037	0.032	-0.057	-0.045
	$\int A$	**0.048***	0.037	0.008	-0.028
	$\hat{A}$	0.035	0.032	0.032	-0.004
	ΔA	**0.117****	**0.091****	**0.186***	0.105

(b) Normalized Data

predictor of the ground truth which is robust regardless of windowing methods and degrees of annotator memory.

B. Normalized Data

In the second round of experiments we normalize the annotation data to $[0, 1]$ using min-max normalization (based on the bounds of each individual annotation trace) and repeat the above process. Table Ib shows the corresponding rank correlation values. Once again, the predictive capacity of ΔA is directly observable, as it yields the highest correlation values across window types and degrees of annotation memory. It is the only annotation metric that manages to be a significant predictor of the phasic driver of SC regardless of the conditions it is tested on. Similarly to the earlier experiment, the amplitude of the annotation trace is highly correlated with both phasic driver features only when we consider rankings of all windows split based on the continuous window method.

VII. DISCUSSION

The introduced *RankTrace* annotation method was tested in a horror game for the first-person annotation of tension in a player's recorded playthrough. On the hardware level, the wheel-like sensor allowed for intuitive and low-fatigue annotation. On the software level, the lack of bounds allowed users to annotate their perceived tension without having to anticipate whether a future experience will be more tense than the current one. The unbounded signal can be used as is or normalized post-hoc without affecting the annotator's experience. Our experiments explored ways of processing the traces, both in terms of the type of time windows used and in terms of the memory we consider for the annotator. Regardless of the experimental setup, it was revealed that the average gradient of the annotation is the most efficient and robust predictor of the hypothesized ground truth (i.e. the phasic activation of SC) among the four metrics tested. The other relative metric, which disregards the actual annotation values but only assesses their amplitude, also fared better than two absolute metrics (i.e. mean and integral of the trace). While the integral of the trace reached a significant correlation in one test, statistics based on absolute values overall performed worse, especially when the trace was split into continuous windows. Assuming limited annotator memory, ranking between adjacent windows yielded higher correlations in general. Our correlation analysis showed that ΔA has considerable potential, but combining it with other annotation metrics (e.g. $\hat{A}$), game events (e.g. monster attacks) and the current game state (e.g. room illumination) in a non-linear fashion via preference learning [34] is expected to result in more accurate models of the ground truth.

This initial study explored the impact of *RankTrace* on players' annotation traces, focusing solely on the capacity of the annotation data at predicting the ground truth. While our initial findings point towards the use of relative measures of annotation data and the informal user feedback about the interface was positive, the full potential of *RankTrace* and its hardware as an annotation protocol needs to be further tested across several factors. For instance, a future user study needs to compare bounded and unbounded versions of *RankTrace* to potentially reveal the positive effects of unbounded annotation. Another user study could compare versions of standard mouse-based interfaces (e.g. Gtrace) versus *RankTrace*'s wheel-like interface via post-use questionnaires, interviews and annotation trace analysis. We expect motion artifacts from hand movement to be present in both mouse and wheel interfaces; their degree might vary, however. Regardless of the annotation interface used, frequency-based signal filtering can reduce any unnecessary motion artifacts. Future experiments could also assess whether the relative characterization of annotations via *RankTrace* results in higher degrees of inter-rater or intra-rater agreement; to test this, each recorded playthrough could be evaluated by several annotators (inter-rater agreement) or the same annotator could annotate the same gameplay experience several times (intra-rater agreement). Further, the tool needs to be thoroughly tested using other signals as ground truth (beyond skin conductance), in other games and game genres and in other emotion annotation tasks beyond game experience. To promote the use *RankTrace*, an accessible version of the tool is available online[1]. Finally, we note that the gender of participants is not balanced in this study (9 males; 5 females) and future studies should investigate whether the gender of annotators may have an impact on our core findings. Comparing annotations from different genders showed similar trends, with stronger correlations between the SC ground truth and annotations from female participants; however, this could be due to more data points originating from male participants.

[1] http://www.autogamedesign.eu/software

VIII. Conclusions

This paper introduced the *RankTrace* tool which allows for a continuous yet unbounded and relative annotation of affect. The interface promotes relative-based annotation as it relies on a wheel-like hardware; its software allows the annotator to observe her own annotation trace over time without being bounded by any limits. We tested the tool in a horror game and asked for first-person annotations of tension. Participants first played the game, then observed their video-recorded playthroughs while annotating the level of tension; their skin conductance was recorded during play. Important game events were used for splitting the annotation traces into time windows and deriving the phasic driver of skin conductance via CDA [21]. Next, we derived two intensity-based (absolute) and two relative metrics from the annotation traces as we wanted to explore the ability of different methods to capture the underlying ground truth. Our results reveal one core pattern: the relative metric of average gradient of the annotation traces is the most consistent and robust predictor of our ground truth (i.e. the phasic driver of skin conductance). The average gradient manages to predict the SC phasic response best regardless of whether we normalize the annotations or not, whether we consider full versus limited annotator memory, or whether we split the entire trace into frames or consider only time windows based on particular events. This paper adds to the increasing number of studies demonstrating the benefits of relative annotation for modeling affect more reliably [9].

References

[1] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proceedings of the Conference on Automatic Face and Gesture Recognition*, 2013.

[2] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *International Conference on Affective Computing and Intelligent Interaction*, 2015.

[3] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *Proceedings of the ISCA Tutorial and Research Workshop on Speech and Emotion*, 2000.

[4] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with emotionml," in *Proceedings of the Conference on Affective Computing and Intelligent Interaction*, 2013.

[5] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, 2013.

[6] P. M. Müller, S. Amin, P. Verma, M. Andriluka, and A. Bulling, "Emotion recognition from embedded bodily expressions and speech during dyadic interactions," in *International Conference on Affective Computing and Intelligent Interaction*, 2015.

[7] Y. Zhang, L. Zhang, S. C. Neoh, K. Mistry, and M. A. Hossain, "Intelligent affect regression for bodily expressions using hybrid particle swarm optimization and adaptive ensembles," *Expert Systems with Applications*, vol. 42, no. 22, 2015.

[8] G. N. Yannakakis and H. P. Martinez, "Ratings are Overrated!" *Frontiers on Human-Media Interaction*, 2015.

[9] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions," in *International Conference on Affective Computing and Intelligent Interaction*, 2017.

[10] G. N. Yannakakis and J. Hallam, "Rating vs. preference: a comparative study of self-reporting," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 437–446.

[11] H. Helson, "Adaptation-level theory," *American Journal of Psychology*, 1947.

[12] N. Stewart, N. Chater, and G. D. Brown, "Decision by sampling," *Cognitive psychology*, vol. 53, no. 1, 2006.

[13] G. A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956.

[14] D. Laming, "The relativity of absolute judgements," *British Journal of Mathematical and Statistical Psychology*, vol. 37, no. 2, 1984.

[15] N. Stewart, G. D. Brown, and N. Chater, "Absolute identification by relative judgment." *Psychological review*, vol. 112, no. 4, 2005.

[16] B. Seymour and S. M. McClure, "Anchors, scales and the relative coding of value in the brain," *Current opinion in neurobiology*, vol. 18, no. 2, 2008.

[17] P. Lopes, A. Liapis, and G. N. Yannakakis, "Modelling affect for horror soundscapes," in *Transactions of Affective Computing*, 2017, accepted for publication.

[18] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the International Conference on Multimodal interfaces*. ACM, 2004.

[19] A. Clerico, C. Chamberland, M. Parent, P.-E. Michon, S. Tremblay, T. Falk, J.-C. Gagnon, and P. Jackson, "Biometrics and classifier fusion to predict the fun-factor in video gaming," in *Proceedings of the Computational Intelligence and Games conference*. IEEE, 2016.

[20] P. Lopes, A. Liapis, and G. N. Yannakakis, "Targeting horror via level and soundscape generation," in *Proceedings of the AAAI Artificial Intelligence for Interactive Digital Entertainment Conference*, 2015.

[21] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of neuroscience methods*, vol. 190, no. 1, 2010.

[22] D. R. Bach and K. J. Friston, "Model-based analysis of skin conductance responses: Towards causal models in psychophysiology," *Psychophysiology*, vol. 50, no. 1, 2013.

[23] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1, 2003.

[24] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, 1980.

[25] D. S. Messinger, T. D. Cassel, S. I. Acosta, Z. Ambadar, and J. F. Cohn, "Infant smiling dynamics and perceived positive emotion," *Journal of Nonverbal Behavior*, vol. 32, no. 3, 2008.

[26] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmüller, "Emujoy: Software for continuous measurement of perceived emotions in music," *Behavior Research Methods*, vol. 39, no. 2, 2007.

[27] L. Devillers, R. Cowie, J. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie, "Real life emotions in french and english tv video clips: an integrated annotation protocol combining continuous and discrete approaches," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2006.

[28] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Proceedings of the Conference on Acoustics, Speech and Signal Processing*, 2011.

[29] U. Schimmack and A. Grob, "Dimensional models of core affect: A quantitative comparison by means of structural equation modeling," *European Journal of Personality*, vol. 14, no. 4, 2000.

[30] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, 2005.

[31] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011.

[32] C. Holmgård, G. N. Yannakakis, H. P. Martínez, K.-I. Karstoft, and H. S. Andersen, "Multimodal PTSD characterization via the startlemart game," *Journal on Multimodal User Interfaces*, vol. 9, no. 1, 2015.

[33] M. Garbarino, M. Lai, D. Bender, R. W. Picard, and S. Tognetti, "Empatica E3A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," in *Proceedings of the International Conference on Wireless Mobile Communication and Healthcare*. IEEE, 2014.

[34] G. N. Yannakakis, "Preference learning for affective modeling," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2009.