Grounding Truth via Ordinal Annotation

Georgios N. Yannakakis and Héctor P. Martínez
Institute of Digital Games, University of Malta, Msida, 2080, Malta
Email: {georgios.yannakakis, hector.p.martinez}@um.ed.mt

***Abstract*—The question of how to best annotate affect within available content has been a milestone challenge for affective computing. Appropriate methods and tools addressing that question can provide better estimations of the ground truth which, in turn, may lead to more efficient affect detection and more reliable models of affect. This paper introduces a rank-based real-time annotation tool, we name *AffectRank*, and compares it against the popular rating-based real-time FeelTrace tool through a proof-of-concept video annotation experiment. Results obtained suggest that the rank-based (ordinal) annotation approach proposed yields significantly higher inter-rater reliability and, thereby, approximation of the underlying ground truth. The key findings of the paper demonstrate that the current dominant practice in continuous affect annotation via rating-based labeling is detrimental to advancements in the field of affective computing.**

***Keywords*—affect annotation; ranking; FeelTrace; AffectRank; inter-rater agreement; Krippendorff's alpha**

I. Introduction

Affect annotation is a laborious and challenging process of utmost importance for affective computing as it provides an estimation of the *ground truth* of highly subjective constructs such as emotional states. The *accuracy* of that estimation is regularly questioned as there are numerous factors contributing to a deviation between a user's label and the actual underlying phenomenon investigated (e.g. an affective state). These factors include, but not limited to, the annotator's motivation and experience, the emotion representation chosen (e.g. continuous vs. discrete), the annotation tool and the interface provided, and person-dependent annotation delays [1].

In this paper we explore the design and use of annotation tools and interfaces towards more reliable affect annotation which brings us closer to the ground truth of emotion. We are motivated by earlier studies in subjective assessment comparisons between ratings and ranks showcasing the supremacy of the latter for obtaining first-person annotations of lower inconsistency and order effects [2], [3]. We are also driven by observations of recent studies in third person video annotation indicating that "...humans are better at rating emotions in *relative* rather than absolute terms." [1]. Grounded in the aforementioned earlier evidence and observations we have designed a rank-based real-time annotation tool we name *AffectRank* that can be used for the annotation of any type of content including images, video, text or speech. In this initial study we explore the use and efficiency of the tool for video affect annotation. While annotation efficiency depends on a number of criteria such as usability and validity [4] in this paper we primarily focus on inter-rater reliability.

Motivated by the supreme properties of rank-based annotation in dissimilar studies within affective computing [2], [3], [5], [6], [1], [7] the key hypothesis that we attempt to validate in this paper is as follows: *Rank-based annotation yields higher inter-rater reliability than rating-based annotation*. We test this hypothesis in a proof-of-concept experiment composed of five videos from two different datasets and four annotators that use both the FeelTrace [4] continuous annotation tool and the proposed *AffectRank* discrete rank-based annotation tool on the arousal-valence 2D plane. The core results obtained validate our hypothesis: *AffectRank* provides annotations that are significantly more reliable (with respect to inter-rater agreement) than the annotations obtained from FeelTrace.

This paper is novel in several ways. First, it introduces a rank-based (ordinal) annotation tool that is of generic use across dissimilar emotive content (videos, images, sounds, text etc.). Second, it proposes a generic methodology for comparing different types of emotive annotations such as ratings and ranks. Finally, it offers a first thorough comparison between dissimilar video annotation tools and, as a result, it challenges directly the dominant practice of continuous rating-based emotion annotation.

II. Affect Annotation: Background

Manually annotating emotion is a challenge in its own right both with respect to the human annotators involved and the annotation protocol chosen. On one hand, the human annotators need to be skilled enough to be able to approximate the perceived affect well and, therefore, eliminate subjective biases introduced to the annotation data. On the other hand, there are many open questions left for the designer of the annotation study when it comes to the annotation tools and protocols used. Will the person experiencing the emotion (first person) or others (third-person) do the labeling? How well trained (or experienced) should the annotators be and how will the training be done? Will the labeling of emotion involve states (discrete representation) or does it involve the use of emotion intensity or affect dimensions (continuous representation)? When it comes to time, should it be done in real-time or offline, in discrete time periods or continuously? Should the annotators be asked to *rate* the affect in an absolute fashion or, instead, *rank* it in a relative fashion? Answers to the above questions yield different data annotation protocols and, inevitably, data quality, validity and reliability.

Representing both time and emotion as a continuous function has been one of the dominant annotation practices within

affective computing over the last 15 years. Continuous labeling *with respect to emotion* appears to be advantageous compared to discrete states labeling for several reasons. The states that occur in naturalistic data hardly fit word labels or linguistic expressions with fuzzy boundaries. Further, when states are used it is not trivial to capture variations in emotion intensity and, as a result, earlier studies have shown that inter-rater agreement tends to be rather low [8]. The dominant approach in continuous annotation is the use of Russell's two-dimensional (arousal-valence) circumplex model of affect [9]. Valence refers to how pleasurable (positive) or unpleasurable (negative) the emotion is whereas arousal refers to how intense (active) or lethargic (inactive) that emotion is.

Continuous labeling *with respect to time* has been popularized due to the existence of tools such as FeelTrace (and its variant GTrace [10]) which is a freely available software that allows real-time emotional annotation of video content [4], the continuous measurement system [11] which has also been used for annotating videos, and EmuJoy [12] which is designed for the annotation of music content. The real-time continuous annotation process, however, appears to require a higher amount of cognitive load compared to e.g. offline and discrete annotation protocols. Such cognitive load often results in low inter-rater agreement and unreliable data annotation [13], [14].

In this paper we introduce *AffectRank*: a real-time, discrete, rank-based annotation tool for video annotation and beyond. Earlier studies in the area of affective computing [2], [5], [6], [7], [3] have shown the advantages of rank-based emotion annotation for various purposes; none, however, investigates the impact of rank-based annotation on video annotation. Most importantly, to the best of our knowledge, no study compares the inter-rater agreement of rating-based versus rank-based continuous annotation or tests the efficacy of dissimilar annotation tools whatsoever. The reported study in this paper compares *AffectRank* against the popular and benchmarked FeelTrace tool showcasing the clear benefits of rank-based annotation in obtaining higher inter-rater agreement.

III. Annotation Procedure and Data Collection

This section describes the protocol followed for the experiments presented in this paper, the two tools developed for testing our hypothesis and the video datasets used for the annotation. We conclude this section with statistics from the annotation data obtained.

A. Protocol

We asked four annotators (1 female) to annotate five videos from two different datasets (see more about the specifics of the videos used in Section III-C). The annotators are all researchers within the areas of machine learning, artificial intelligence and games, and affective computing. All of them are well aware of the basic principles of arousal and valence, they all have participated in emotion annotation experiments in the past, and they all had further training in emotion annotation through a graduate course in affective computing.

We have created a web-based application for running our annotation experiments. Each annotator is logged in with her/his personal user name at the web application and, at the beginning of the annotation process, s/he is provided with detailed information about the purpose of the experiment and the core properties of the arousal-valence circumplex model of affect as defined by Russell [9]. Then the annotator is requested to follow a tutorial to get him/herself familiarized well with both annotation interfaces and the annotation process per se. The tutorials allow the users to test both annotation tools on a sample video that is different from the five videos used in this study.

Once the annotator feels comfortable using the annotation tool s/he proceeds to the main part of the experiment. The annotator is either presented with the FeelTrace or the *AffectRank* tool first and has to complete the corresponding tutorial. The order of tool presentation is randomized to minimize potential order effects introduced to our data. In both annotation schemes the annotator can pause the annotation process at any time and continue at a later stage. We implemented the pause feature for easing the fatigue that increases naturally during manual data annotation [1] in an attempt to minimize possible effects in our data collection. In addition, by selecting 5 short videos to show (annotated with both tools resulting to 10 videos) we aimed to keep the experimentation time at a reasonable window of around 40 minutes for each annotator. Pilot experiments showed that 30 to 40 minutes of annotation time are a good compromise between data quantity and quality with respect to user motivation and fatigue.

B. Annotation Tools and Interface

For assessing the capacity of *AffectRank* we compare it against a custom-made version of FeelTrace [4] which is arguably the most popular continuous affect annotation tool for videos. This section provides the details of the two tools used in the experiments of this paper and summarizes their differences.

The custom-made FeelTrace annotation tool (see Fig. 1) follows the basic principles of continuous emotion annotation on the arousal-valance plane. The annotator is presented with the circumplex model of affect depicted as a two-dimensional plane of arousal and valence. The arousal axis spans from inactive $(-)$ to active $(+)$ whereas the valence axis spans from unpleasant $(-)$ to pleasant $(+)$. The user activates the green dot in the origin of the axes and moves it freely within the circle in real-time to indicate the current values of arousal and valence. When moved, the dot leaves an animated trace of earlier positions as depicted in Fig. 1. Mouse positions (coordinates) are stored as a two dimensional vector with values lying within $[-1, 1]$. Data logging for the FeelTrace tool follows the specifications of [4]. The interface records every mouse movement and later resamples the signal at a constant sampling rate of 5 samples per second. Compared to the standard FeelTrace tool we have made a number of improvements as also suggested in [1]. In particular we have placed both the annotation tool and the video in the

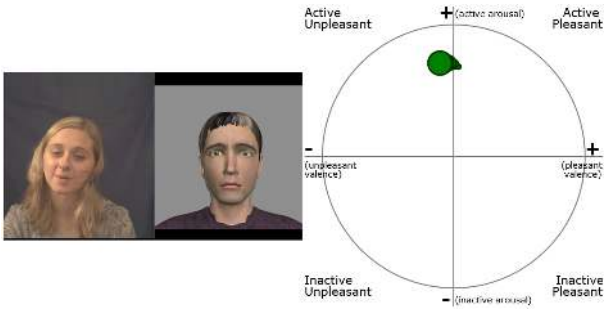


Fig. 1. The custom-made FeelTrace tool for real-time continuous annotation.

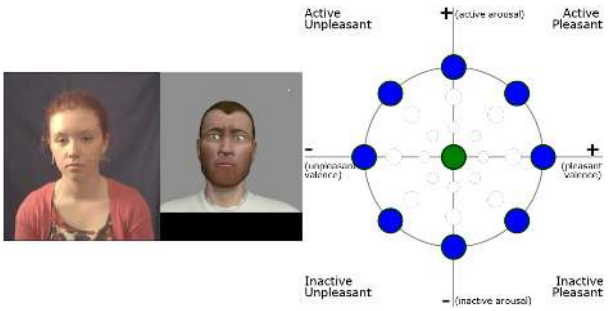


Fig. 2. *AffectRank*: the real-time, discrete, rank-based annotation tool introduced in this paper.

same window minimizing annotator distraction and we have improved the general usability of the tool as the user is not required to constantly click on the mouse for data to be logged.

The *AffectRank* annotation tool (see Fig. 2) uses the same arousal-valence representation and axes labels but, in contrast to FeelTrace, it requests annotators to indicate a *change* in arousal, valence, or both, only when they judge that such a change occurs (i.e. users annotate in real time but not continuously). Users are presented with 8 viable options (blue circles in Fig. 2) covering all the possible changes in the arousal valence plane: active, active pleasant, pleasant, inactive pleasant, inactive, inactive unpleasant, unpleasant, active unpleasant (see Fig. 2). The white circles appearing in Fig. 2 have been designed for animation purposes only. Evey time the user selects amongst the 8 options the corresponding white circles turn into green to better illustrate the selection.

The differences between the two annotation tools are described herein. *AffectRank* is a *discrete*-based emotion annotation tool both with respect to time and the arousal-valance space. Annotation happens only when the user clicks on possible arousal, valence, or arousal and valence change (discrete time) while the annotator can only pick from a predetermined number of states of change (8 in this case). On the contrary, the custom-made FeelTrace tool allows, by nature, for *continuous* annotation both with respect to time and the state space. Annotators can freely select any point in the arousal-valence plane while the mouse position is logged continuously. Clearly FeelTrace allows for more granularity during annotation. The final, yet critical, difference between the two is that *AffectRank* forces the annotator to **rank** affect in a *relative* fashion (i.e. to indicate a change in the arousal-valence plane) whereas annotators of the custom-made FeelTrace tool **rate** in a real-time *absolute* fashion.

C. Video Datasets

For testing our hypothesis across different video contexts we have used videos from two dissimilar datasets. Two out of five videos were selected from the freely available[1] SEMAINE video dataset [15] and three more videos were picked from the Eryi game-playing dataset. This section outlines the key properties of the datasets and the corresponding videos selected from them

SEMAINE [15] is a large audiovisual database containing interactions of people with agents in emotionally colored conversations. Recordings of high quality (5 high-resolution, high framerate cameras, and 4 microphones) from a total of 150 participants is included in the database. The database contains a total of 959 conversations with various agents lasting approximately 5 minutes each. Two videos were randomly selected from this dataset for the purposes of our experiences. The first video features a participant's interaction with the agent *Spike* (who is constitutionally angry; see Fig. 2) and the second features participant interaction with agent *Obadiah* (who is gloomy; see Fig. 1) [15].

The videos of the Eryi dataset were collected from research students of the Institute of Digital Games, University of Malta, for the purpose of modeling player experience (in particular frustration and engagement) in platformer games using a multimodal approach. The full dataset contains 13 game sessions of the 2D platformer game *Eryi's Action* (Xtal Sword, 2012) which is played by 13 participants. The Eryi dataset is not publicly available yet. The recording of the Eryi dataset takes place using one Kinect sensor placed just above the computer monitor recording the facial and head movements of the participant (see Fig. 3). Beyond the videos recorded, the dataset contains synchronized and detailed in-game information which is displayed during the annotation procedure at the top left of the video (see top left image of Fig. 3). For the experiments presented in this paper we selected videos from three different participants. The three participants were picked for their high expressiveness during gameplay with the working assumption that non-expert annotators would find the resulting videos easier to annotate.

D. Data Collected

The data collected across videos, participants and annotation tools is summarized in Table I. Compared to the continuous sampling of FeelTrace, *AffectRank* produces a smaller and variable amount of annotations (see two examples in Fig. 4). In this paper we argue that these fewer annotations are more reliable as they correspond to significant and clear changes of perceived affect. As already observed in [1] and seen in the continuous annotation examples of Fig. 4 raters tend to agree in relative terms (i.e. trend) but not in absolute terms (i.e. intensity of emotion).

[1] http://semaine-db.eu/

Fig. 3. A snapshot from the Eryi dataset.

TABLE I
NUMBER OF ANNOTATIONS ACROSS VIDEOS (V1-V5), ANNOTATORS (A1 TO A4) AND ANNOTATION TOOLS (FEELTRACE VS. *AffectRank*). V1 TO V2 AND V3 TO V5 ARE THE VIDEOS OBTAINED FROM THE SEMAINE AND THE ERYI DATASET, RESPECTIVELY.

	FeelTrace	*AffectRank*			
	A1 - A4	A1	A2	A3	A4
V1	955	16	12	20	16
V2	1000	19	26	22	15
V3	1575	45	38	26	15
V4	1550	36	60	29	12
V5	1700	35	64	32	18

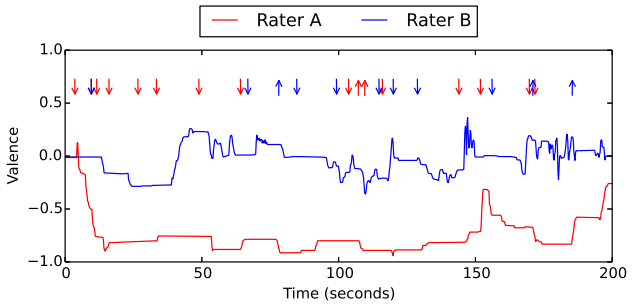


Fig. 4. Valence annotations of the same video by two raters using FeelTrace (continuous lines) and *AffectRank* (arrows).

IV. RESULTS AND ANALYSIS

This section presents the coefficient considered for comparing inter-rater agreement across the two different annotation tools (Section IV-A) and the key set of results obtained.

A. Test Statistic: Krippendorff's Alpha

To test the key hypothesis of the paper we need a measure of inter-rater reliability (agreement) that would be able to cater for both interval and ordinal values obtained from FeelTrace and *AffectRank*, respectively. While Cronbach's α [16] is the dominant coefficient for estimating the inter-rater agreement in the psychometrics and the affect annotation literature (e.g. in [1], [15]) it is not applicable to ordinal data and therefore cannot be used for a direct comparison between the two annotation tools.

Krippendorff's α [17], on the other hand, is a versatile statistic that measures the degree of agreement obtained among observers who label, categorize, rate, or rank a given set of objects in terms of the values of a given variable. The metric is rather generic as it can support any number of observers and several types of observations (such as nominal, ordinal, and interval), and it is able to handle missing data. The above properties make Krippendorff's α the ideal test statistic for the comparison between interval (FeelTrace) and ordinal (*AffectRank*) annotations available in our datasets. The obvious benefit of selecting such coefficient for our purposes is that the computed inter-rater reliabilities are comparable across any number of annotators, annotation data types and unequal sample sizes obtained via the different annotation schemes.

According to Krippendorff's alpha, the degree of reliability (α) between a number of raters is as follows

$$\alpha = 1 - (D_o/D_e) \tag{1}$$

where D_o is the observed disagreement between the raters and D_e is the expected disagreement. For space considerations we omit the detailed formulas for D_o and D_e and refer the interested reader to [17]. Clearly, *perfect* reliability and *absence* of reliability is, respectively, indicated by α values of 1 and 0. If $\alpha < 0$ disagreements amongst raters are systematic and lie beyond what can be expected by pure chance.

Note that an annotated dataset is expected to yield different Cronbach's and Krippendorff's α values. Cronbach values depend on the variance of annotated values in relation to the variance of the sum of all annotations. Krippendorff values, on the other hand, depend on the differences between the annotated values and the frequency of occurring values across annotators.

B. Inter-rater Agreement Comparison: General Methodology

To make the comparison between continuous and discrete annotation possible one needs to discretize time with predetermined time windows so that continuous values and discrete values are comparable within the same time windows. This is the traditional practice for the analysis of continuous annotation (e.g. see [15], [1]). We have partitioned the obtained data by considering two time windows in this paper: 3 and 5 seconds. More time windows were considered but those proved to provide either over-detailed information for affect annotation (time windows smaller than 3 seconds) or very few data points for comparison (time windows larger than 5 seconds). The two selected time windows give us a representative picture of how time discretization impacts inter-rater agreement across the two tools used.

Once data is partitioned within time windows the next step is to preprocess the continuous and discrete values to make the comparison fair. For *AffectRank* every time arousal and/or valence is increased (or decreased) within a time window we add (or subtract) 1 from the accumulated value within that window. We then compare the values and derive the relative change in arousal/valence between two subsequent

TABLE II
SAMPLE SIZES FOR THE CALCULATION OF KRIPPENDORFF'S α

	FeelTrace		*AffectRank*	
	3 sec	5 sec	3 sec	5 sec
Arousal	571	344	325	391
Valence	571	344	445	418

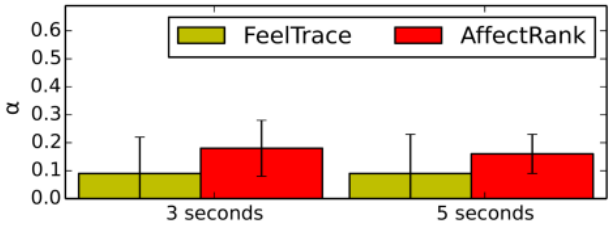
time windows. For FeelTrace, on the other hand, we explore two ways of treating the obtained values. In Section IV-C we first treat FeelTrace ratings as numerical (interval) values and average them for the comparison against the ordinal data obtained from the *AffectRank* tool — averaging rating values from real-time continuous annotation is a common practice within affective computing (e.g. in [15]). Then in Section IV-D we calculate maximum rating deviations across the two dimensions of arousal and valence which indicate noteworthy changes in those values. By following the second approach we convert FeelTrace annotation ratings into ranks which, as a practice, has evidenced advantages for affective modeling [6]. The converted FeelTrace ranks are compared against both the standard FeelTrace ratings and the ranks obtained from *AffectRank*.

C. Average Ratings (FeelTrace) vs. Ranks (AffectRank)

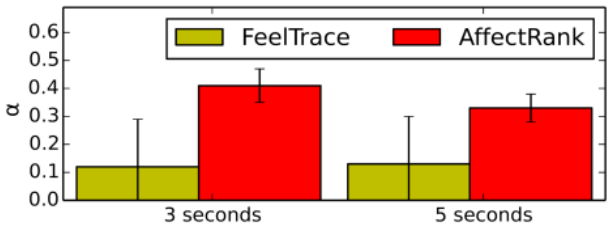
The continuous rating values of FeelTrace are averaged within the two time windows selected. Before delving into the comparative analysis against *AffectRank* we present the Cronbach's α values for FeelTrace as a baseline for inter-rater reliability obtained from the tool. The values are 0.69 for arousal and 0.9 for valence when the 3 second time window is applied (571 samples). The corresponding values for the 5 second time window are 0.62 and 0.83 (344 samples). Both results indicate that FeelTrace manages to yield high inter-rater agreement (as measured by Cronbach's α) for both affective dimensions in the videos tested. Further, it appears that valence is the affective dimension that was easier for annotators to agree upon.

For each annotation tool and affective dimension (arousal, valence) we calculate inter-rater agreement across all possible rater pairs, and in total, via the Krippendorff's α coefficient. The results obtained from the analysis are presented in Fig. 5 and Table II. As a general observation from this first round of experimentation one can derive that *AffectRank* yields more reliable data as the average α values are higher across both time windows explored. Moreover, it is evident that — independently of annotation tool used — α values are higher for valence. This seems to indicate that valence is easier to annotate within the selected videos.

As stated earlier, *AffectRank* not only offers a rank-based alternative to FeelTrace but also a discrete version of it with 8 options for the annotator to pick from. An obvious question is then how much of that observed increase in inter-rater agreement is due to the emotion-discrete (nominal) representation of *AffectRank* and how much of it is due to the rank-based (ordinal) nature of it. To address this question of tool validity



(a) Arousal.



(b) Valence.

Fig. 5. Krippendorff's α values for the two time windows and annotation tools. Standard deviations are calculated across the five videos.

we treat the data from *AffectRank* as nominal (8 classes in total) assuming that annotators did not annotate a change (rank) but rather a class and we recalculate the α coefficients. The α coefficients obtained for nominal *AffectRank* values are 0.15 (arousal) and 0.27 (valence) for the 3 second window and 0.18 (arousal) and 0.29 (valence) for the 5 second window. This shows that the nominal representation of *AffectRank* — i.e. annotators treating the eight discrete options as classes — yields lower inter-rater agreement for valence compared to the ordinal representation. The inter-rater agreements of the nominal *AffectRank* are still higher compared to the ones obtained from the continuous FeelTrace annotations. We can therefore conclude that the nominal *AffectRank* (i.e. a discrete version of FeelTrace) contributes to higher inter-rater agreement. However, it is primarily the rank-based annotation feature of *AffectRank* that elevates the α values to much higher levels (e.g. up to 0.41 for the valence dimension).

D. Ranked Ratings (FeelTrace) vs. Ranks (AffectRank)

We follow the same approach as in the previous set of experiments with the only difference that we now treat ratings obtained from FeelTrace naturally as ordinal data (as suggested in [6]). To do so we have picked a small distance margin (0.005 in this paper) above which a change in arousal and/or valence is considered a data point within each time window. Higher margins than 0.005 gave limited data points for any viable comparison. Results obtained for FeelTrace data following this approach are depicted in Fig. 6.

By observing the α values of Fig. 6 it becomes clear that the transformation of rating values to ranks is beneficial for achieving higher inter-rater agreement. Compared to the raw FeelTrace average values (see Fig. 5) the FeelTrace ordinal values yield higher inter-rater reliability for arousal with a

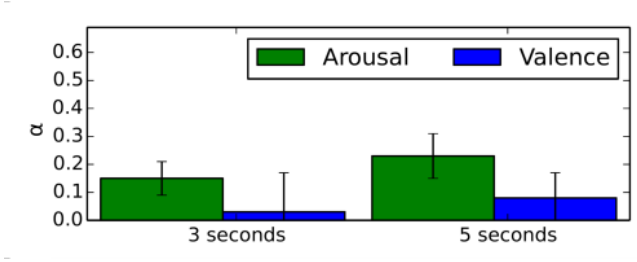


Fig. 6. Ranking FeelTrace ratings: Krippendorff's α values for the two time windows and affective dimensions. Standard deviations are calculated across the five videos.

insignificant drop in valence. This finding further validates the evidence provided in [6] which suggests that ratings should be naturally converted to ordinal values (ranks) for more reliable affect detection. Compared to *AffectRank* (see Fig. 5) the FeelTrace ordinal annotations yield much lower α values for valence whereas the difference in arousal is insignificant. While treating ratings as ordinal values increases inter-rater agreement it is far more reliable as a practice to ask annotators directly to rank amongst options (as in *AffectRank*) instead of asking them to rate with absolute values within the arousal valence plane (as in FeelTrace) [3].

V. Discussion and Future Directions

This initial study serves as the base for exploring the benefits of rank-based annotation in a proof-of-concept experiment. The focus of the paper is not on presenting yet another large annotated corpus and analyzing it rather than on introducing a new way of annotating and showcasing its clear benefits over standard rating-based annotation practices. Even though the experiments presented and the data collected proved to be sufficient for validating our key hypothesis more experiments with more annotators and more annotated videos will be required to further strengthen our validated hypothesis. It is important to note, however, that the data collection protocol followed in this paper is a good compromise between annotation time, and data quality as 40 min provides a reasonable time window for reliable data collection that keeps annotators motivated on the task. Indicatively, when asked, all annotators found the time spend on the task appropriate and expressed that they would not have wanted to annotate further videos.

This paper complements findings of several studies showcasing the complexity of affect annotation in the arousal valance plane. To ease the complexity of the task and study each affective dimension independently we intent to modify *AffectRank* for allowing the annotation of one affective dimension at a time (e.g. following the design principles of GTrace [10]). In that way, more affective dimensions, such as dominance, can be investigated in future annotation experiments. Furthermore, other popular annotation tools beyond FeelTrace — such as the self-assessment manikin and AffectButton — can offer a comprehensive set of comparisons against *AffectRank*; whether that is for video annotation of other types of content.

An obvious question of researchers with limited prior experience on ordinal data is how to use and further process the ranks obtained [3]. Non-parametric statistical methods such as the *Wilcoxon signed-rank test* [18] and *Kendall's Tau* [19] can be used to calculate the correlation between a hypothesized order and the observed ranks — see e.g., [6]. The non-parametric *Kruskal-Wallis* [20] and *Friedman's* [21] tests for three (or more) groups of ranks are also applicable. Furthermore, if one wishes to build computational models that predict those ranks a large set of algorithms such as linear discriminant analysis, Gaussian processes, artificial neural networks, support vector machines and deep networks are available. These methods are derived from the sub-area of machine learning named *preference learning* [22], [23], [24]. A number of such methods are currently included in the open-access, user-friendly and accessible Preference Learning Toolbox[2] (PLT) [25]

A possible next step is to attempt to machine learn the mapping between video properties and annotations for detecting affect. Given the findings of this paper and the evidence provided in [6] we expect that the generated affect models built on the *AffectRank* data to be more accurate — compared to models built on FeelTrace annotation data — and closer to the underlying ground truth. Note that, while the annotations produced by *AffectRank* are not continuous, we are still able to derive an underlying continuous affect model from rank annotations via preference learning [22], [23], [6].

VI. Conclusions

Motivated by recent findings in affective modeling the core hypothesis we attempted to validate in this paper was that relative (ordinal) affect annotation yields more reliable data compared to absolute annotation. To test this hypothesis we introduced the *AffectRank* tool which allows for real-time discrete-based annotation of content in a relative fashion (i.e. via ranks). We compared *AffectRank* against an improved version of the popular FeelTrace continuous annotation tool for the annotation of videos. The key findings of our study suggest that the ordinal annotations of *AffectRank* yield higher inter-rater agreement compared to the FeelTrace rating annotations. The agreement amongst *AffectRank* annotators is higher even when FeelTrace annotations are naturally converted to ranks.

We believe that this paper offers a solid foundation towards a paradigm shift within affective computing: from rating-based to rank-based emotion annotation. The core results presented confirm the speculations of earlier studies in affect annotation [1] and suggest that rating-based annotation can be detrimental to advances in affect sensing and modeling [3], [6].

Acknowledgment

The authors would like to thank all annotators that participated in the reported experiments. We would also like to thank Gary Hili and Ryan Abela for providing access to the Eryi dataset. The work is supported, in part, by the EU-funded FP7 ICT iLearnRW project (project no: 318803).

[2]http://sourceforge.net/projects/pl-toolbox/

REFERENCES

[1] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[2] 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.

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

[4] 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 *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[5] S. Tognetti, M. Garbarino, A. Bonarini, and M. Matteucci, "Modeling enjoyment preference from physiological responses in a car racing game," in *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*. IEEE, 2010, pp. 321–328.

[6] H. Martinez, G. Yannakakis, and J. Hallam, "Dont classify ratings of affect; rank them!" *Affective Computing, IEEE Transactions on*, vol. 5, no. 3, pp. 314–326, 2014.

[7] Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 762–774, 2011.

[8] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1, pp. 5–32, 2003.

[9] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[10] R. Cowie and M. Sawey, "Gtrace-general trace program from queens, belfast," 2011.

[11] 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, pp. 133–155, 2008.

[12] 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, pp. 283–290, 2007.

[13] 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 *5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, 2006, p. 22.

[14] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2376–2379.

[15] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 5–17, 2012.

[16] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

[17] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage, 2012.

[18] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, pp. 80–83, 1945.

[19] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, pp. 81–93, 1938.

[20] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[21] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[22] J. Fürnkranz and E. Hüllermeier, *Preference learning*. Springer, 2010.

[23] G. N. Yannakakis, "Preference learning for affective modeling," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–6.

[24] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *Computational Intelligence Magazine, IEEE*, vol. 8, no. 2, pp. 20–33, 2013.

[25] V. E. Farrugia, H. P. Martínez, and G. N. Yannakakis, "The preference learning toolbox," *arXiv preprint arXiv:1506.01709*, 2015.