2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)

The Ordinal Nature of Emotions

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

Roddy Cowie
Queen's University, Belfast
r.cowie@qub.ac.uk

Carlos Busso
The University of Texas at Dallas
busso@utdallas.edu

***Abstract*—Representing computationally everyday emotional states is a challenging task and, arguably, one of the most fundamental for affective computing. Standard practice in emotion annotation is to ask humans to assign an *absolute* value of intensity to each emotional behavior they observe. Psychological theories and evidence from multiple disciplines including neuroscience, economics and artificial intelligence, however, suggest that the task of assigning reference-based (*relative*) values to subjective notions is better aligned with the underlying representations than assigning absolute values. Evidence also shows that we use reference points, or else *anchors*, against which we evaluate values such as the emotional state of a stimulus; suggesting again that ordinal labels are a more suitable way to represent emotions. This paper draws together the theoretical reasons to favor relative over absolute labels for representing and annotating emotion, reviewing the literature across several disciplines. We go on to discuss good and bad practices of treating ordinal and other forms of annotation data, and make the case for preference learning methods as the appropriate approach for treating ordinal labels. We finally discuss the advantages of relative annotation with respect to both reliability and validity through a number of case studies in affective computing, and address common objections to the use of ordinal data. Overall, the thesis that emotions are by nature relative is supported by both theoretical arguments and evidence, and opens new horizons for the way emotions are viewed, represented and analyzed computationally.**

***Keywords*—emotion annotation; ranks; ratings**

I. Introduction

Describing everyday emotional states is a complex and nontrivial task [1], and therefore so is labeling data for affective computing [2]. Several particular challenges affect labeling. People's impressions of emotions are a basic source, but the channels through which people can externalize them are frustratingly narrow. Standard practice in affective computing involves *absolute* annotation. Categorical labels (e.g., happiness and anger) or attribute descriptors (e.g., arousal and valence) try to locate the state of the subject at a point in a tree or a continuous space. Annotators are doing that when they observe an emotional behavior, and report the label or score that best reflects their perception. However, emotions are inherently structured. They are about something [3], which is why context is crucial to understanding [4]. They are also situated within narratives [5]. Trying to assign numbers to emotions is not simply a noisy task: there are a multitude of theoretical and practical reasons to doubt that subjective notions function as numbers in the first place [6].

In spite of theoretical doubts, it made sense to explore the use of absolute annotation for representing and annotating emotion. However, the outcome underlines the doubts. This paper makes the case for engaging with a fundamentally different theoretical stance, where emotions are regarded as intrinsically *relative*. If so, their annotation and analysis should follow the ordinal path. To support this thesis we first review the theoretical arguments from psychology, and the empirical evidence coming from other disciplines, which favor relative conceptions of emotion (Section II). We then consider measurement issues, involving reliability and validity, which show the advantages of a relative annotation approach (Section III). In Section IV we describe the different annotation data types and the various (good, bad and ugly) ways these can be processed within affective computing. Then in Section V we present the preference learning paradigm for deriving affect models from ordinal data as demonstrated via a plethora of applications. Successful case studies showcasing the benefits of relative annotation for videos, speech and games are also presented (Section VI). The paper concludes with a discussion on the most common objections to the use of ordinal annotation (Section VII). Taken together, the arguments and evidence make a case for a paradigm shift in the way emotions are described computationally, annotated, and modeled.

II. Why Relative over Absolute?

Multiple disciplines, including philosophy, psychology, economics, and artificial intelligence have studied the problem of measuring subjective variables, and taking decisions based on the results. In this section we survey the way various disciplines have developed ideas about relative perspectives and their importance for our decision making, as a way of dealing with subjective notions such as affect.

A. Psychological Perspectives

A long standing thesis in psychology holds that while humans are efficient at discriminating among options [7], that is not matched by their ability to assign accurate absolute values for the intensity of what we perceive—and therefore, contrary to what we might imagine, does not depend on it. We are particularly bad at giving absolute values when the estimates involve subjective variables, such as the tension, frequency and loudness of sounds, the brightness of an image, or the arousal level of a video [7].

The theory related to that issue comes from various strands, and has a long history. As a way of structuring it, it makes sense to begin with themes that have early roots. *Adaptation level theory* was an approach to sensory experience that took

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

shape in the 1940s, and was drawn together by Helson in 1964 [8]. The central idea is that experience signals departure from a default level, the adaptation level, which is a weighted mean of previous stimuli. It is striking that an early paper by Russell, who became the best-known advocate of a dimensional account, stressed the relevance of adaptation level theory to emotion [9]. Similar points are made in diverse theories, including *relative judgment models* [10] suggesting that experience with stimuli gradually creates our internal context, and discussions of *anchors* [11], against which we rank any forthcoming stimulus or perceived experience. The accounts agree that our choice about an option is driven by our internal ordinal representation of that particular option within a sample of options; not by any absolute value of that option [12]. A particularly well-developed account, due Stewart et al. [12], extended theories of relative judgment to economic decision making. It models the subjective value involved in a decision as the outcome of a series of pairwise ordinal comparisons within a sample of attribute values drawn from memory; which, in turn, determine the final rank of the decision within the sample of options. The theory explicitly rejects appeals to underlying psycho-economic scales. Instead, it offers a Helsonian picture. Binary, ordinal comparisons to material held in memory provide the basis of subjective value. The value reflects its rank in a sample biased towards recent experience, but also longer-term information about the distribution of relevant attributes.

There is a straightforward application of the above theories to affect annotation. As with sensory dimensions, affective estimates are relative to an adaptation level. The implication is that it is very doubtful to treat equal ratings at different points in a rating session as if they meant the same thing. More fundamentally, Helson considers how experiences are related. He considers it obvious that there are different kinds of pleasantness or unpleasantness. The dimensional character comes from the fact that instances that are of different kinds can still be ordered. Ordering involves comparison with reference experiences. It is part of the adaptation level paradigm that a body of recent experiences act as immediate referents, but Helson also points a longer-term process, linked to conditioning, that establishes more enduring comparators. On that picture, numerical descriptions are a proxy for describing where, in a sequence built by comparisons, a particular experience lies.

The picture of emotions' relative nature fits most of the observations that are outlined in this paper. The argument is emphatically not that the evidence forces us to accept that kind of picture. It is that it provides a valuable perspective on the problems with reliability and disputes about practice in the collection and analysis of data within affective computing. It is not unreasonable to suspect that they reflect the need to rethink models of the way affect works at quite a deep level.

B. Other Perspectives

The concepts underlying this paper have been explored in several other disciplines beyond psychology. The results and evidence are relevant for the study of emotions but are not covered in detail due to space considerations. In *marketing research* values are traditionally measured with the use of rank-based questionnaires. According to Rokeach [13] societal or ethical values are acquired, internalized and organized in a hierarchical manner. The ranking approach naturally helps the respondent to discover, reveal and crystallize [13] his/her hierarchy of values in a self-reporting manner. The empirical evidence in that area is strong. For instance, a large scale study involving over $3,500$ students across 19 counties [14] compared ratings and rankings for addressing the recurring problems of response style differences and language biases in cross-national research. The findings support the ranking approach: they show that it is more effective at reducing response biases in cross-cultural settings.

In *neuroscience*, Damasio [15] reports extensive experiments on the role of emotion in decision making. They imply that each time we are presented with a stimulus, we construct and store an anchor (or a *somatic marker*) which is eventually a mapping between the presented stimulus and our affective state. We then use these somatic markers as drivers for making choices between options. Given its unique role, affect can naturally guide our attention towards preferred options and, in turn, simplify the decision process for us. There is also evidence in monkeys [16] suggesting that their brain—in particular, the orbitofrontal cortex (OBC)—encodes values in a relative fashion. Similar results have been reported for the human medial OBC [17].

Arguments in *philosophy* also emphasize the fundamental role of comparison in subjectively defined values. The *ceteris paribus* (or *everything else being equal*) preference theory by Hansson [18] attempts to offer a unified formal structure of values and norms. Hansson claims that most of our daily decisions and choices are based on preferences *ceteris paribus* of two relata, A and B—as the building block of our decision making process. It rests on the problem of comparing structures with multiple attributes—which, as noted at the beginning of the paper, emotional experiences typically do, because they involve a context and a narrative.

The notion of *preference* is nowadays central in *artificial intelligence* and machine learning [19]. The theoretical grounding of learning from preferences [20] is based on humans' limited ability to express their preferences *directly* in terms of a specific value function [21]. That limitation holds even if the underlying scale of the notion we wish to asses is ordinal (e.g., in the case of ratings). This inability is mainly due to the subjective nature of a preference and the notion we express a preference about, and the substantial cognitive load required to give a specific value for each one of the available options we have to select from; and (recalling Hansson) each one of the options is characterized by a number of attributes (or the context) that we consider. Thus instead of rating our options directly it is far *easier* and more *natural* to express preferences about a number of limited options; and this is what we end up doing normally. As the *relative* comparison between pairs of options is less demanding (cognitively) than the *absolute* assessment of a set of single options, pairwise preferences are easier to specify than exact value functions about the available options.

III. Performance Measures

This paper argues for the advantages of ranks as an emotion annotation tool. Clearly, that depends on identifying a measure of performance that shows these advantages. In this section we outline the two core criteria that against which annotation strategies can be measured: *reliability* and *validity*.

Inter-rater reliability is the degree of agreement among a number of annotators. This performance measure yields superior results for relative (rank-based) annotation, compared to other absolute annotation methods, as showcased by the case studies detailed in this paper. Testing whether the annotator is consistent over time (test-retest reliability) is also relevant for the thesis of this paper (see Section VI).

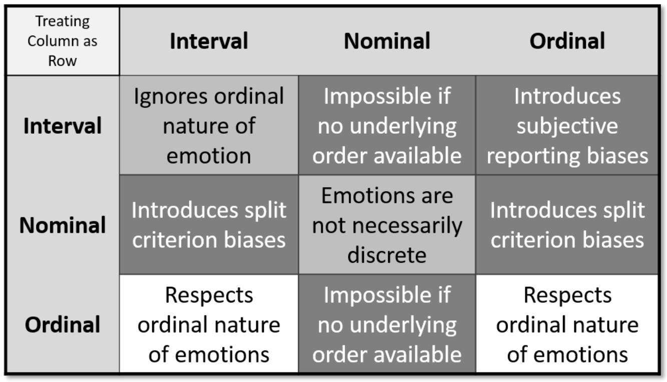
The validity of annotation is the degree to which the annotation construct measures the phenomenon we claim we measure. While interlinked to an extent, validity and reliability are different notions; the latter measures the degree to which our observations about a phenomenon are consistent. Validity in this paper is measured by the process of cross-validation in statistics and machine learning. Cross-validation examines the degree to which the result of a statistical analysis on data can generalize to unseen (independent) data. Several case studies in Section VI, and others in the literature showcase the superior generalizability of ordinal approaches to modeling affect.

IV. Data Analysis for Affect Modeling

There are both theoretical and empirical arguments in favor of ordinal approaches to affect annotation and affect modeling. If so, obtaining ordinal labels in the first place would seem to be the ideal approach. That is not always possible, though. So, how should we process other data types, and how should we machine learn from data types other than ordinal? Following the taxonomy of Stevens [22] we can distinguish three data types that we can obtain from an emotion annotation task: *interval*, *nominal*, and *ordinal*. The next sections is what the thesis of the paper implies for the various data analysis practices followed in affective computing. That leads to an outline that covers the three different data types used, and considers the *good*, the *bad* and the so-called *ugly* data analysis practices associated with each. These practices are depicted in Fig. 1, respectively, as white, dark gray and light gray table cells.

A. Annotations are Interval

Interval data represent an affect state or dimension with a scalar value or a vector of values. Intervals are often confused with ratings and the terms are used interchangeably; however, ratings are not interval but rather ordinal values [6]. The most popular rating-based question is a Likert item [23] in which users are asked to specify their level of agreement with a given statement. Popular rating-based questionnaires for affect annotation include the Geneva Wheel model [24] and the Self-Assessment Manikin (SAM) [25]. When annotations come in an interval form we can treat them as such or alternatively treat them as nominal or ordinal data.



Treating Column as Row	Interval	Nominal	Ordinal
Interval	Ignores ordinal nature of emotion	Impossible if no underlying order available	Introduces subjective reporting biases
Nominal	Introduces split criterion biases	Emotions are not necessarily discrete	Introduces split criterion biases
Ordinal	Respects ordinal nature of emotions	Impossible if no underlying order available	Respects ordinal nature of emotions

Fig. 1. Practices in affective modeling: Treating column data types as row data types. White, dark gray, and light gray table cells, respectively, illustrate the *good*, *bad* and the *ugly* practices according to the thesis of this paper. By good we refer to approaches that are theoretically sound and compatible with the key message of the paper. By bad we refer to approaches that are technically flawed or even impossible and are also incompatible with the ordinal approach advocated in this paper. Finally, by ugly (perhaps too aggressively) we refer to approaches that are possible but nevertheless incompatible to the key message of the paper.

If interval data is treated as such then a form of regression is naturally implied. For instance, one can think of attempting to approximate the absolute interval traces of arousal or valence using FeelTrace. This is a dominant practice in affective computing and it is also theoretically solid from a machine learning perspective. However, as advocated in this paper, the approach of approximating absolute values of subjective constructs such as emotions in models that misrepresent the ground truth of emotion.

Treating interval values as nominal data, instead, implies that one needs to first classify continuous annotations (e.g., from FeelTrace) and then create models via classification. This is another dominant practice in affect modeling (e.g., see studies on the SEMAINE dataset [26]) but recent evidence suggests that such practice introduces a multitude of biases in our data and thus takes us further away from the underlying ground truth [27]. Furthermore, creating dichotomized labels from interval data creates unavoidable problems where similar samples around the boundary are artificially placed in different classes [28].

Any attempt to derive an ordinal scale from interval data that characterize subjective notions appears to be a good practice to follow [29]–[31]. Several studies have transformed values of affect to ordered ranks and then derived affect models via preference learning: the transformation improves cross-validation capacities [27], [32], [33].

B. Annotations are Nominal

The second annotation data type one may obtain comes in *nominal* (or class) form. Nominal data are mutually exclusive labels which are not ordered, such as sex or unordered affective states. Note, however, that nominal data sometimes take the form of a *preference* involving two or more options (for instance, they may indicate preference for the timbre of one sound in a list, or the warmth of one image in a set). There, an order of preference is implied—or is inherent—and underlies the observations. Binary nominal data that have a meaningful underlying order can also be viewed as borderline nominal. Examples include yes or no answers to questions such as *do you think this is a sad facial expression?* or *is the user in a high- or a low-arousal state?* In all such instances we argue that data can be safely treated as ordinal.

Deriving interval scores out of nominal values seems flawed or impossible unless there is an underlying order across the classes; e.g., an attempt is presented in [34]. The approach, however, leveraged on individual evaluations instead of consensus labels. The key idea was to create a probabilistic score per emotional category by considering the inter-evaluator agreement. The framework also considered relationships between emotional categories. For example, a sample receiving the label *excitement* increases its *happiness* score since these emotions are related. This is only possible if individual evaluations are available; otherwise, converting nominal values into interval scores is not feasible or appropriate.

Nominal data is ideal for multi class machine learning problems when emotional content is described in terms of categorical emotions. The common approach in affective computing is to ask multiple evaluators to select an emotional category after watching or listening to a stimulus. The individual evaluations are then often aggregated creating consensus labels. Forced-choice responses where an evaluator has to select an emotion out of a list create inaccurate descriptors, however. Depending on the options, the same stimulus can be annotated with different emotions [35]. Furthermore, nominal labels do not capture any within-class differences (i.e., different shades of happiness). As a result, the nominal labels tend to be noisy yielding poor inter-rater agreement, especially when the list of emotions is large [36].

An order cannot be easily derived from classes which are unordered—e.g., happiness and sadness. Indicatively, Lotfian and Busso [34] used a probabilistic score to define preferences between samples. The study established preferences when the difference between the probabilistic score of two samples was greater than a margin. On a similar basis, Cao et al. [37] also derived preferences from categorical emotions; in their study, every sentence labeled as happy was preferred over sentences labeled with another emotion. One drawback of this approach, however, is that it is not possible to establish preferences between samples from the same class. We argue for a more direct approach: to ask annotators to rank samples directly (e.g., is sample A happier than sample B?).

C. Annotations are Ordinal

Ordinal data can be obtained via rank-based annotation protocols. The annotator is asked to rank a preference among options such as two or more images, musical pieces [31], sounds [38], or video screenshots [39], [40]. On its simplest form, the annotator compares two options and specifies which one is preferred under a given statement (*pairwise preference*). With more than two options, the annotator is asked to provide a ranking of some or all the options. Examples of rank-based questions include: *was that level more engaging than this level? which facial expression looks happier? is the user more aroused now?*

Data obtained through the common rating-based annotation tools in affective computing such as FeelTrace or SAM is ordinal by nature [6]. Such data is generally treated as interval values, however—for instance, by averaging the obtained annotation values. While this is the dominant practice in psychometrics at large, there is extensive evidence for its invalidity and the numerous subjective reporting biases such analysis introduces to data [6], [29], [41].

Another popular practice is to treat ordinal data as nominal and view the problem as a classification task. Recent studies comparing the use of ordinal affect labels as ordinal against the use of ordinal labels as classes showcase the superiority of the first in yielding more general models of affect [27].

Finally treating ordinal data as ranks and viewing the problem of affect modeling as a preference learning task both respects the nature of the data and yields affect models of supreme validity [27], [33] and reliability [40]. The studies presented in Section VI provide additional evidence for the superior nature or relative affect annotation and its analysis for affect modeling.

V. So, I Have Ranks; How do I Derive my Models?

Given the different data types that we can obtain from our annotations, the next step is naturally to process it statistically or derive affect models which rely on this data. A popular objection to the use of ordinal labels is the lack of statistical tools and methods to process them. Section VII addresses common objections directly, but this section focuses on preference learning (PL) [19], [20], the natural approach to process ordinal (affect annotation) data and derive (affect) models from this data. PL is a subfield of supervised learning dedicated to the processing of ordinal labels. The PL paradigm as an approach for affective modeling was first introduced by Yannakakis back in 2009 [30]. Since then numerous studies in affective computing have used PL for affect detection and retrieval through images [42], [43], videos [39], [44], music [31], [45], sounds [38], speech [27], [37], [46], games [27], [47], [48] and text [49].

There are several algorithms and methods available for the task of preference learning. Most of them reduce the problem to pairwise comparisons where the task is to determine whether one sample, A, is preferred over another sample, B, (i.e., $A \succ B$). The results of the pairwise comparisons are used to rank the samples. It is important to note that *any* supervised learning algorithm can be converted to a PL problem by using an appropriate formulation. Linear statistical models, such as linear discriminant analysis and large margins, and non-linear approaches, such as Gaussian processes, shallow and deep artificial neural networks, and support vector machines, are applicable for learning to predict ranks.

A popular derivation for PL consists of using binary classifiers. Let ϕ be the feature vector of sample x. If x_i is preferred over x_j (i.e., $x_i \succ x_j$), the objective is to find a hyperplane w such that $w(\phi_i - \phi_j) > 0$, which is equivalent to a binary classification problem where the features are the subtraction of their respective feature vectors. This problem can be solved by any binary classifier; e.g., RankSVM is the equivalent PL method for support vector machines (SVMs) [50].

An alternative formulation for PL is training a function f that maintains a higher preference for the preferred option; for example, if $x_i \succ x_j$ then $f(\phi_i) > f(\phi_j)$. There are several approaches to create this function: for example, it can take a parametric distribution as done with Gaussian processes [51], or can be learned from data using deep learning structures as performed via convolutional neural networks [47], [52] or via RankNet [53], or via neuroevolution [27], [30]. Studies have demonstrated that all aforementioned methods provide compelling results [30], [34], [46], [52], [54]. For the interested reader, a number of PL methods including RankSVM, neuroevolutionary preference learning and PL via backpropagation are contained in the preference learning toolbox (PLT) [55]. PLT is an open-access toolkit built and constantly updated for the purpose of easing the processing of ordinal labels.

A. *Preference Learning for Affective Computing: Applications*

Any application in emotion recognition can be formulated as a ranking problem in which PL algorithms are trained to predict ordinal labels. Examples of applications include forensic analysis where the goal is to prioritize the videos or audio to be analyzed by selecting a subset of recordings with target emotional content (e.g., threatening behaviors). Another example is in identifying emotionally salient regions, relying on relative emotional changes [56]. Computational tools that are able to rank emotions are also suitable for emotion retrieval, where the goal is to identify examples associated with a given emotional content [34]. Applications of emotional retrieval include solutions for health care domains [57], [58]. In longitudinal studies relying on remote assistant technologies, rank-based emotion retrieval can provide an ideal framework for healthcare practitioner to identify and review relevant events from patients with emotional disorders. Emotion retrieval from speech can facilitate better solutions for call centers. It can also facilitate the collection of natural emotional speech databases [59]. Emotion-aware recommendation systems are also an important application area for PL using ordinal labels (e.g., selecting music or sounds conveying emotions that match the current affective preference of the user [31], [38]).

The breadth of applications expand to video-based, [39], [40], speech-based [56] or physiology-based [60] emotion recognition for health, educational or entertaining [41] purposes. The next section covers a few successful applications directly showcasing the benefits of ordinal annotation and processing for affect modeling.

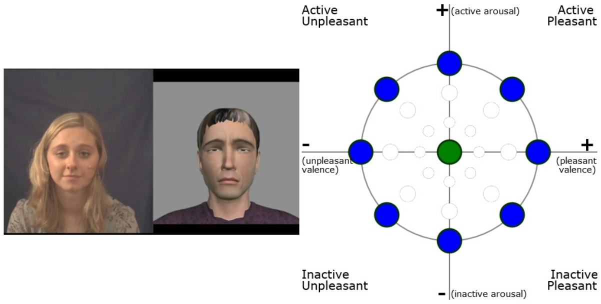


Fig. 2. *AffectRank*: the rank-based annotation tool introduced in [40]. *AffectRank* is inspired by FeelTrace but it allows the real-time annotation of arousal and/or valence in a relative fashion.

VI. Case Studies Relevant for Affective Computing

In this section we outline a number of studies across various domains of affect computing that support the main argumentation and evidence of this paper. None of these studies is new; however, they are put together for the first time, thereby, collectively making our thesis stronger. We focus on important affect annotation studies that compare ordinal annotations against class-based and/or rating-based protocols within the domains of video, speech, and game experience annotation. Please note that the list is not exhaustive, because space is limited.

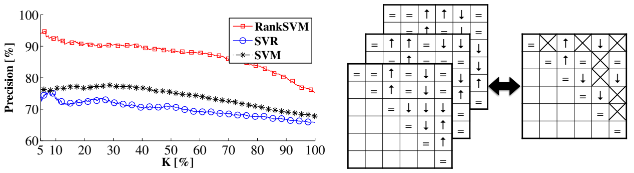
A. *Videos*

While Metallinou and Narayanan [29] have long indicated the need of tools that would allow for a relative annotation of videos it is only very recently that such tools were introduced. The annotation tool named *AffectRank* [40] is a freely-available[1], rank-based version of FeelTrace which asks the annotator to indicate a *change* in arousal and/or valence while watching a video. The evaluation study of [40] compared the inter-rater reliability between FeelTrace and *AffectRank* for the video annotation of two datasets: the SEMAINE [61] and the Eryi game dataset. The obtained results validate the hypothesis that *AffectRank* provides annotations that are significantly more reliable than the annotations obtained from FeelTrace (see Fig. 2). *AffectRank* yields superior reliability even when FeelTrace ratings are treated as ordinal data. The key findings of [40] further support the thesis of this paper by demonstrating that the dominant practice in continuous video affect annotation via rating-based labeling has negative effects.

B. *Speech*

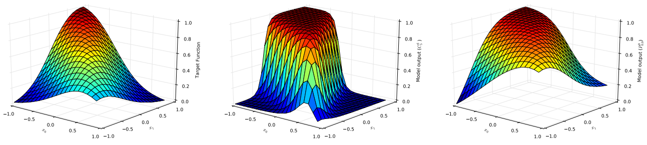
Recent work in speech-based affect recognition has demonstrated the benefits of using PL with ordinal labels [32]–[34], [46]. Using time-continuous evaluations for arousal and valence provided by FeelTrace, the above studies defined

[1] https://github.com/TAPeri/AffectRank



(a) Rankers versus other methods (b) Finding Trends in Traces

Fig. 3. Case studies on speech: (a) improved *precision at K* (P@K) of RankSVM over regression (SVR) and binary classification (SVM) for arousal [32], (b) QA framework to identify trends from emotion annotation traces [33].



(a) Ground Truth (b) Classification (c) Preference learning

Fig. 4. A hypothesized (artificial) ground truth function (z-axis) which is dependent on two attributes, x_1 and x_2 (Fig. 4a), the best classification model (Fig. 4b) and the best preference learned model (Fig. 4c) [27].

preferences between pairs of speech samples and compared PL (via RankSVM) against binary classification and regression for modeling arousal and valence. The task consisted of determining whether the value of the attribute of one sample was above or below the median value across the corpus (i.e., median split). This formulation applies directly to binary classifiers, where the positive and negative classes are defined according to the median split. For implementing regression, the predicted scores were sorted by selecting the samples at the top and bottom of the list. For preference learning, samples were ranked according to the emotion attributes, selecting samples in the extremes. The evaluation demonstrated that preference learning provided over 10% increase in cross-validation performance compared to the other two methods (see Fig. 3a). The evaluation also revealed two important observations. First, PL makes better use of the training set. Even when the margin that defines a preference is large, most of the data is still included in the ordinal dataset. Second, the results seem to saturate for RankSVM as the number of pairwise comparisons increases over $5,000$ in the training set. We expect that deep architectures will be able to handle a bigger dataset, achieving better results [27], [46], [47].

In another study, it was proposed to define ordinal labels by considering trends in the time-continuous labels [33]. Each dialog is annotated by multiple evaluators creating a trace per rater. A common observation is that these traces are noisy with low inter-evaluator agreement. Instead of averaging the traces across evaluators, the qualitative agreement (QA) framework [62] was used to identify segments where most of the evaluators agreed on trends (e.g., increase or decrease in the values of the traces). This framework leverages consistent information provided in the, otherwise, noisy traces. The emotion annotation traces are segmented into bins, and their average are compared creating an individual matrix per evaluator (right side of Fig. 3b). The arrows denotes increasing or decreasing trends between bins. All the individual matrices are then combined creating a consensus matrix with the consistent trends (left side of Fig. 3b). The core findings suggest that extracting ordinal labels with QA provides better classifiers, increasing the accuracy of the emotion rankers.

C. Games

The literature on the benefits of ordinal annotation in video games is rich. Several studies have explored both first-person and third-person ordinal annotation of playing experience and player affect. Indicatively, ordinal annotation protocols have been explored in racing games [48], prey-predator [6], [27], [40], horror [38], and physical interactive games [41] among many other game genres. Most notably for the purposes of this paper, Yannakakis and Hallam compared rating versus ranking annotations of first person experience in both a prey predator and a physical interactive game [41]. Their subjects were asked to use 5-point (prey-predator) or 10-point (physical interactive game) Likert items versus a ranking protocol to answer questions about the experience of the games they just played. The affective states they explored spanned from *fun* to *frustration*, to *excitement* and *boredom*. Their key findings reveal that rater consistency (reliability) is higher when ranking protocols are used across both games. Further their evidence suggests that the order of answering affects ratings more than ranks. In other words, ranks yield higher degrees of test-retest reliability.

In another study, Martinez et al. [27] worked on the hypothesis that the best way of analyzing ratings of affect is to treat them naturally as ordinal data. To test their hypothesis they compared models of affect that are the result of converting ratings to classes (classification) versus ordinal models that are trained directly via preference learning. They used three datasets for their analysis: an artificial dataset, a dataset from the MazeBall game containing physiological signals and gameplay data [52] and the SAL [63] corpus which contained 739 1-second-long speech segments. The main findings of their study validate their hypothesis and further support the thesis of this paper. Models trained via preference learning outperform the classification models of affect in terms of cross-validation. Figure 4 showcases how much closer a preference learned model can reach a hypothesized (artificial) ground truth, compared to a classification model.

Importantly for the thesis of this paper, Holmgaard et al. [60] compare different types of stress annotation with the aim of finding the best possible approximation to the underlying ground truth. In particular they compare annotations indicating the most stressful event in a game (class-based annotation) versus a rank-based approach by which subjects compare stress across game events. Their findings reveal that the ordinal annotations are more accurate predictors of the phasic driver of skin conductance which is assumed to be a reliable indicator of underlying stress.

VII. Can Less Be More?

In this section we discuss a number of traditional objections to the use or ranks or relative constructs in affective computing and human computer interaction at large. There are certainly more objections we could discuss; however, we put an emphasis on the most important ones given the limited space.

More is Better: It is natural to believe that more information about a subjectively defined notion such as an emotion is a good property. It is also safe to believe that if one merely compares emotions in a relative fashion and does not specify an absolute value about them the resulting analysis suffers from lack of resolution and intensity. However, all studies presented in this paper, the evidence collected from other disciples and, finally, the psychological framework about the relative nature of emotions collectively suggest the exact opposite: that *less is more*. It appears that the additional information offered by absolute annotation is only biasing the search for valid and reliable models of affect. Further, at first sight it might seem that the *intensity* of an emotion is completely lost if it is expressed in a relative fashion as it is only compared to an anchor or a reference point. Empirical evidence from this paper showcase that the intensity is not lost; it is instead lying under the provided ranks. The function that models affective ranks (e.g., a preference learned neural network [27], [47]) can directly output intensity values of the modeled affect. In other words, intensity is not only present when ordinal labels are used but is also free of reporting biases caused through absolute annotation.

Anchors: To rank means inherently to compare. To be able to compare one needs a point of reference. Ranking-based annotations require at least one reference point which is usually found as an option in the question. While the requirement of a reference point might seem a core limitation of rankings we argue that it is their biggest strength. As we discussed thoroughly in this paper it is theoretically grounded that humans maintain a baseline when using any type of reporting scheme. Be it a class or a rating question an annotator will make a comparison based on other items within the scale or earlier responses in similar questions and contexts. The power of ranks is that this baseline extraction process does not have to happen unconsciously or intuitively; it is *forced*. Further the baseline is not some mere approximation of preference indicated by a scale or distorted by memory; it is a real option one uses as a reference during the annotation. Once again, this property of ranks appears to be a limitation but it instead encapsulates one of their core strengths.

Statistical Analysis: Given their ordinal nature it is not always possible to apply conventional statistical methods to ordinal data. Standard descriptive statistics such as mean values and standard deviations are not applicable. Parametric tests are not applicable either. There are still, however, multiple data visualization methods and data processing techniques that span from classical correlation analysis to statistical tests for significance and further to modern machine learning approaches that are available for handling preferences and ranks. This paper already covered several of the methods (Section V) but the reader is also referred to [6], [40] for more details.

VIII. Conclusions

This paper has presented and supported the thesis that emotions are by nature *relative*. We do not claim that it is a novel thesis. On the contrary, we have taken pains to show that it reflects established ideas in many literatures—psychology, philosophy, neuroscience, marketing research, artificial intelligence and, not least, affective computing. The research in affective computing allows us to identify good and bad practices for the analysis of interval, nominal and ordinal annotations; and it provides several practical ways of processing ordinal affective labels via preference learning. It also provides studies, across various domains of affective computing, which showcase the advantages of treating emotions as relative notions. Our fundamental aim is to make it clear that this is not a fringe issue. Attempts to work with absolute annotation, including our own, have shown that problems we knew in principle do not turn out to be unimportant in practice. The cumulation of evidence says that it makes sense to look in a concerted way at the alternative that various teams, including our own, have been exploring.

References

[1] R. Cowie, "Describing the forms of emotional colouring that pervade everyday life," in *The Oxford Handbook of Philosophy of Emotion*, P. Goldie, Ed. Oxford, UK: Oxford University Press, January 2010, pp. 63–94.

[2] R. Cowie, C. Cox, J. C. Martin, A. Batliner, D. Heylen, and K. Karpouzis, "Issues in data labelling," in *Emotion-Oriented Systems*, P. Petta, C. Pelachaud, and R. Cowie, Eds. Springer Berlin Heidelberg, July 2011, pp. 213–241.

[3] P. Goldie, "Emotions, feelings and intentionality," *Phenomenology and the Cognitive Sciences*, vol. 1, no. 3, pp. 235–254, September 2002.

[4] L. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, October 2011.

[5] P. Goldie, *The mess inside: narrative, emotion, and the mind*. Oxford, UK: Oxford University Press, September 2012.

[6] G. N. Yannakakis and H. P. Martínez, "Ratings are overrated!" *Frontiers in ICT*, vol. 2, p. 13, 2015.

[7] 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.

[8] H. Helson, "Adaptation-level theory," 1964.

[9] J. A. Russell and U. F. Lanius, "Adaptation level and the affective appraisal of environments," *Journal of Environmental Psychology*, vol. 4, no. 2, pp. 119–135, 1984.

[10] N. Stewart, G. D. Brown, and N. Chater, "Absolute identification by relative judgment." *Psychological review*, vol. 112, no. 4, p. 881, 2005.

[11] 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, pp. 173–178, 2008.

[12] N. Stewart, N. Chater, and G. D. Brown, "Decision by sampling," *Cognitive psychology*, vol. 53, no. 1, pp. 1–26, 2006.

[13] M. Rokeach, *The nature of human values*. Free press, 1973.

[14] T. Johnson, P. Kulesa, Y. I. Cho, and S. Shavitt, "The relation between culture and response styles: Evidence from 19 countries," *Journal of Cross-cultural psychology*, vol. 36, no. 2, pp. 264–277, 2005.

[15] A. R. Damasio, "Descartes' error: Emotion, rationality and the human brain," 1994.

[16] L. Tremblay and W. Schultz, "Relative reward preference in primate orbitofrontal cortex," *Nature*, vol. 398, no. 6729, pp. 704–708, 1999.

[17] R. Elliott, Z. Agnew, and J. Deakin, "Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans," *European Journal of Neuroscience*, vol. 27, no. 9, pp. 2213–2218, 2008.

[18] S. O. Hansson, "What is ceteris paribus preference?" *Journal of Philosophical Logic*, vol. 25, no. 3, pp. 307–332, 1996.

[19] S. Kaci, *Working with preferences: Less is more*. Springer Science & Business Media, 2011.

[20] J. Fürnkranz and E. Hüllermeier, *Preference learning: An introduction*. Springer, 2010.

[21] C. Domshlak, E. Hüllermeier, S. Kaci, and H. Prade, "Preferences in AI: An overview," *Artificial Intelligence*, vol. 175, no. 7-8, pp. 1037–1052, 2011.

[22] S. S. Stevens, "On the Theory of Scales of Measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[23] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[24] K. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[25] J. Morris, "Observations: Sam: The self-assessment manikin—an efficient cross-cultural measurement of emotional response," *Journal of advertising research*, vol. 35, no. 6, pp. 63–68, 1995.

[26] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, "Non-verbal feedback on user interest based on gaze direction and head pose," in *Proc. of Int. Workshop on Semantic Media Adaptation and Personalization (SMAP)*. IEEE Computer Society, 2007, pp. 171–178.

[27] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *Affective Computing, IEEE Transactions on*, vol. 5, no. 3, pp. 314–326, 2014.

[28] S. Mariooryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 119–130, January-March 2017.

[29] 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.

[30] 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.

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

[32] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5205–5209.

[33] S. Parthasarathy, R. Cowie, and C. Busso, "Using agreement on direction of change to build rank-based emotion classifiers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2108–2121, November 2016.

[34] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 490–494.

[35] J. A. Russell, "Forced-choice response format in the study of facial expression," *Motivation and Emotion*, vol. 17, no. 1, pp. 41–51, 1993.

[36] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[37] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, January 2015.

[38] P. Lopes, A. Liapis, and G. N. Yannakakis, "Modelling affect for horror soundscapes," *IEEE Transactions on Affective Computing*, 2017.

[39] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "From crowdsourced rankings to affective ratings," in *Multimedia and Expo Workshops, 2014 IEEE International Conference on*. IEEE, 2014, pp. 1–6.

[40] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 574–580.

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

[42] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *IEEE International Conference on Image Processing (ICIP 2008)*, San Diego, CA, USA, October 2008, pp. 117–120.

[43] S. Schmidt and W. G. Stock, "Collective indexing of emotions in images. A study in emotional information retrieval," *J. of the American Soc. for Information Science and Technology*, vol. 60, no. 5, pp. 863–876, 2009.

[44] J. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *IEEE International Conference on Multimedia and Expo*, Amsterdam, The Netherlands, 2009, pp. 1436–1439.

[45] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions," in *International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, PA, USA, September 2008, pp. 325–330.

[46] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4995–4999.

[47] H. P. Martínez and G. N. Yannakakis, "Deep multimodal fusion: Combining discrete events and continuous signals," in *Proceedings of the 16th Int. Conf. on Multimodal Interaction*. ACM, 2014, pp. 34–41.

[48] 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, pp. 321–328.

[49] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[50] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *International Conference on Artificial Neural Networks (ICANN 1999)*, Edinburgh, UK, September 1999, pp. 97–102.

[51] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *International conference on machine learning (ICML 2005)*, Bonn, Germany, August 2005, pp. 137–144.

[52] 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.

[53] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Int. Conf. on Machine learning*, Bonn, Germany, 2005, pp. 89–96.

[54] G. Yannakakis, M. Maragoudakis, and J. Hallam, "Preference learning for cognitive modeling: a case study on entertainment preferences," *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 39, no. 6, pp. 1165–1175, 2009.

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

[56] S. Parthasarathy and C. Busso, "Defining emotionally salient regions using qualitative agreement method," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3598–3602.

[57] M. Kranzfelder, A. Schneider, S. Gillen, and H. Feussner, "New technologies for information retrieval to achieve situational awareness and higher patient safety in the surgical operating room: the MRI institutional approach and review of the literature," *Surgical Endoscopy*, vol. 25, no. 3, pp. 696–705, March 2011.

[58] K. Pollak, R. Arnold, A. Jeffreys, S. Alexander, M. Olsen, A. Abernethy, C. Sugg Skinner, K. Rodriguez, and J. Tulsky, "Oncologist communication about emotion during visits with patients with advanced cancer," *Journal of Clinical Oncology*, vol. 25, no. 36, pp. 5748–5752, 2007.

[59] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[60] C. Holmgård, G. N. Yannakakis, H. P. Martinez, and K.-I. Karstoft, "To rank or to classify? annotating stress for reliable ptsd profiling," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 719–725.

[61] 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.

[62] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," September 2010, SEMAINE Report D6b.

[63] K. Karpouzis, G. Caridakis, R. Cowie, and E. Douglas-Cowie, "Induction, recording and recognition of natural emotions from facial expressions and speech prosody," *Journal on Multimodal User Interfaces*, vol. 7, no. 3, pp. 195–206, 2013.