The Ordinal Nature of Emotions: An Emerging Approach

Georgios N. Yannakakis, *Senior Member, IEEE,* Roddy Cowie, and Carlos Busso, *Senior Member, IEEE*

Abstract—Computational representation of 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 people to assign a value of intensity or a class value 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 values to subjective notions is better aligned with the underlying representations.This paper draws together the theoretical reasons to favor ordinal 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 ordinal 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. More broadly, the thesis that emotions are by nature ordinal is supported by both theoretical arguments and evidence, and opens new horizons for the way emotions are viewed, represented and analyzed computationally.

Index Terms—Emotion annotation; labeling; ranks; ratings; classes, ordinal data; preference learning

1 Introduction

One of the most basic challenges in affective computing is annotation—that is, taking direct records of emotionally significant events (audio, video, physiological, etc.), and attaching labels that describe the way people assess the emotions involved. It is *basic* because human assessments are the model that affective computing usually aims to match. Reproducing human assessments is what counts as success in most applications; and the means of achieving that is usually learning from a database of labeled records. The challenge arises because the channels through which people can externalize their impressions are frustratingly narrow. What goes on inside the head during an emotional experience is manifestly far more complex and diverse than any kind of response that annotators can be expected to make, and nobody should expect practicable ways of externalizing it to be perfect.

The theme of this paper is that experience with that challenge invites a *shift*. Exploring approaches with obvious attractions has thrown up difficulties that need to be acknowledged. As a result, different groups have moved, more or less independently, to explore techniques that are less obvious, but that avoid key difficulties. The aim of the paper is to articulate the case for that kind of shift, and to give an overview of the work that shows how the emerging approach can be carried forward. The techniques that we highlight are *ordinal*. They rely on annotators to rank two or more samples which may (or may not) be represented by a scale. The scale may take various forms—for instance intensity, or proximity to an archetypal state, or some theoretical attribute.

To some extent, emphasizing ordinality is about acknowledging realities of measurement. Ordinal information seems to be what people can deliver reliably. However, there are other levels. Computationally, there is the question of how to use ordinal information. Psychologically, there is the question of what the priority of order tells us about the way emotion is represented. That feeds back into computation, because if human assessments are the model that affective computing aims to match, then it matters to reproduce the underlying representations as closely as possible. We emphasize the level of measurement, but we also point to what is known on the other levels.

The case involves several steps. To provide context, we first summarize the problems in current practice and the work from other disciplines, which provide an incentive to develop relative conceptions of emotion (Section 2). Then in Section 3 we present a holistic perspective of affective computing through an ordinal lens. In particular, we first outline the two core ordinal approaches for collecting annotation data (Section 3.1), we then describe the different (good, bad and ugly) ways we can process those labels within affective computing (Section 3.2) and in Section 3.3 we present statistical methods and machine learning approaches for modeling affect from ordinal data. In Section 4, we consider measurement issues involving reliability and validity (Section 4.1). We also present a number of indicative case studies showcasing the benefits of relative annotation across several affective computing domains. The paper concludes with a discussion of the most common objections, limitations, and challenges as well as the general capacities of ordinal annotation (Section 5). Taken together, the arguments and evidence presented in this paper make a case for a paradigm shift in the way emotions are described computationally, annotated, and modeled.

- *G, N. Yannakakis is with the Institute of Digital Games, University of Malta. E-mail: georgios.yannakakis@um.edu.mt*
- *R. Cowie is with Queen's University, Belfast. Email: r.cowie@qub.ac.uk*
- *C. Busso is with The University of Texas at Dallas. Email: busso@utdallas.edu*

TABLE 1
Cronbach's α [1] coefficient values obtained in [2] for functionals (mean value, mean rise and mean fall of the trace) associated with each trace dimension. Note that $\alpha = 0.7$ is almost always considered acceptable; $\alpha = 0.6$ is the lowest value commonly considered acceptable.

	Intensity	Valence	Activation	Power	Expectation
Mean	0.74	0.92	0.73	0.68	0.71
Mean Rise	0.74	0.49	0.53	0.39	0.58
Mean Fall	0.68	0.45	0.55	0.55	0.51

2 Incentives to Consider Ordinal Conceptions

Early work on affective computing took up two contrasting conceptions of emotion: categorical (reflected in everyday language and basic emotion theories); and numerical (reflected in dimensional theories). Logic always indicated that there was ground between them; however, affective computing gave it limited attention. The argument of this section is that in the light of experience, there are substantial reasons to revisit the intermediate ground. They are of three broad types. First, within affective computing, other approaches have hit obstinate difficulties. Second, longstanding theoretical arguments, in philosophy and psychology, offer ways to understand the difficulties, and highlight a different kind of conception. Third, recent technical work in multiple disciplines shows that ideas like that can be translated into practice. Taken together, those provide strong encouragement to explore ordinal conceptions. In addition, they incorporate ideas that are a valuable resource for affective computing.

In this background section we initiate the debate on the grounds of affective computing (Section 2.1) and then discuss theoretical arguments and empirical evidence across the disciplines of philosophy (Section 2.2), psychology (Section 2.3), marketing (Section 2.4), behavioral economics (Section 2.5), neuroscience (Section 2.6) and ultimately AI and machine learning (Section 2.7).

2.1 The Affective Computing Context

In measurement theory, ordinal measurement lies between two familiar alternatives: nominal and interval. Affective computing initially explored the familiar options. Early research in affective computing gravitated towards *nominal* description—associating a sample with an emotion class ('angry', 'happy', etc.) [3], [4]. It became clear early on that there were problems with that approach, most obviously because emotion in realistic situations rarely conformed to a single category [5], [6]. That prompted interest in descriptive schemes based on different psychological theories. The obvious candidates involved *interval* measurement, but applied to attributes of emotion—'dimensions' such as valence and activation [7].

There is now a substantial body of information about interval approaches using annotation tools like FeelTrace. FeelTrace [8] is a freely available software that allows real-time emotional annotation based on Russell's two-dimensional (arousal-valence) circumplex model of affect [9] and is arguably one of the most popular continuous affect annotation tools. Through FeelTrace annotators can provide a continuous time series of interval values (e.g., that lie within $[-1, 1]$) for arousal, valence or any other emotional dimension such as dominance and intensity (see Fig. 1). The key issue is reliability. Early studies showed adequate reliability for FeelTrace ratings, using records chosen so that each portrayed a distinct and fairly consistent emotion, and comparing mean dimensional ratings across records. For example, Savvidou [10] reported coefficients of concordance of 0.963 for valence and 0.978 for activation. However, one of the obvious applications of interval description is to track change within a record. When the same techniques were applied to records where emotion changed substantially, the reliability of ratings within a record fell to problematic levels: the average α values for valence and activation were 0.53 and 0.69, respectively. Those findings have been followed up and amplified.

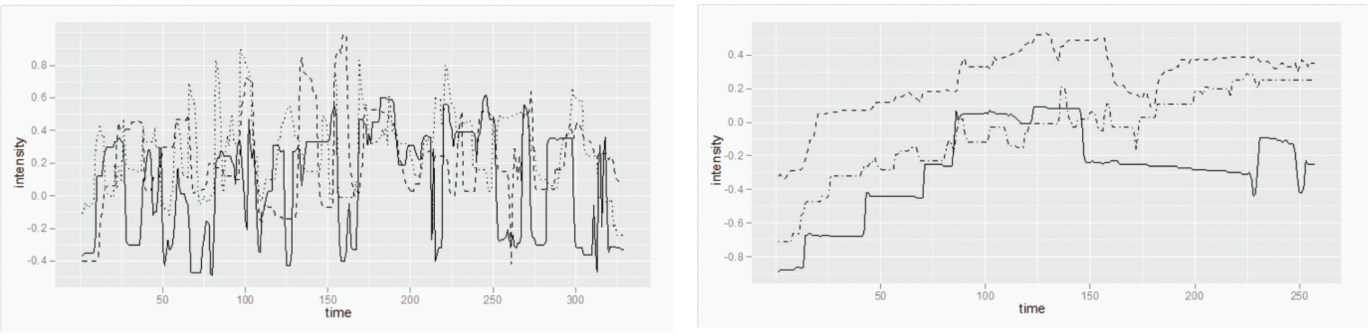
An obvious issue is separating theoretical constructs and measurement types. If categories are fundamental to emotion, as many theorists argue, the ideal might be to combine them with interval measurement (*how much anger is the person feeling?*). Different studies have compared interval measures based on categories and theoretical dimensions, using different measures of agreement. They agree that dimensional ratings give higher reliability [11], [12]. However, they also confirm that interval measurement—whether it is applied to dimensions or categories—ceases to be reliable when there is substantial change within a record. A study of the SEMAINE database (annotated with FeelTrace) made the point by reporting agreement on the size of rises and falls within samples [2]. Table 1 shows some key comparisons. Figure 1 gives a concrete illustration of what lies behind that. All of the traces are of the same record, in the same experiment. They are grouped to show that different individuals gave very different accounts of the way the emotion changed over the time period (about 300 seconds). Those cases illustrate a conclusion that was reached in different ways by different groups. There are substantial problems with both of the more familiar alternatives—nominal and interval. That conclusion points strongly to exploring the option that lies between them, which is ordinal measurement.

The difficulties that have been outlined were not completely predictable. People might have been able to deliver interval descriptions in real time on at least some dimensions. However, limited ability to do that is not a surprise.

2.2 Philosophical Background

A strong tradition in philosophy has presented emotional experience as fundamentally qualitative. Recent discussions develop the tradition in a way that it is useful to have in the background. What has been called the standard view compares them to propositions: a major alternative compares them to perceptions [13]. In either case, they are inherently structured. They are about something [14], and embedded in narratives [15]. That explains why context affects recognition so radically [16]: it lets us see what expressive behaviors are about, and at least parts of the narrative.

Notice that this complexity can easily bring together elements that are evaluated very differently: we see the evil



(a)

(b)

Fig. 1. Individual differences between annotators using the FeelTrace annotation tool to label the same clip (from the SEMAINE dataset [2]). Some (a) report rapid fluctuation, others (b) much more gradual change.

wolf meet sweet Red Riding Hood; we feel how bad the outcome might be, but know it actually ends well. Finding a single number to describe the experience that creates is not simply a noisy task: it is mystifying that we can do it at all [17]. Even ordering structures with multiple attributes is not easy, but the basis for it is clearer. Hansson [18] has proposed a unified formal structure of values and norms in his *ceteris paribus* (or *everything else being equal*) preference theory. Hansson claims that a very large proportion of daily decisions and choices are based on comparisons that pick out key features, and assume that the rest even out—hence '*ceteris paribus*'. That kind of principle provides a basis for ranking emotions.

2.3 Psychological Perspectives

Describing experience is a core problem in psychology, and so it offers great deal of relevant material—theory and empirical work, general and directly relevant to emotion.

Psychophysical research has explored methods of measuring experience since Fechner (in 1860). Early research concluded that pairwise comparison was the only sound approach. It was a century later when Stevens showed that 'magnitude estimation', based on numerical estimates, could be reliable, within limits and given suitable precautions [19]. Interval measurement methods in labeling are a form of magnitude estimation. In effect, the issue is whether this is a case where magnitude estimation is appropriate.

Psychophysical findings challenged naïve intuitions, and several levels of response emerged. On the level of measurement, several kinds of mixed scale were found useful. Prime examples are Likert and Semantic Differential scales [20]. There are many variants. For example self-assessment manikins (SAMs), much used in affective computing, belong in a type of semantic differential where points on the scale are given meaning in their own right (in the case of SAM, by a picture) [21]. Similarly, graded pair comparisons ask which of two cases is higher on a scale, and by how much (very much more, much more, more, a little more, or just barely more) [22]. The differences between mixed scales are generally subtle, and we treat them as ordinal. The labels obtained via Likert-based mixed scales are ordinal even though they are very often (mis-)treated as interval values [17].

A more radical, theoretical response was *adaptation level theory*, which took shape in the 1940s, and was consolidated by Helson in 1964 [23]. He viewed experience as fundamentally relational. It signals departure from a default level, the adaptation level, which is a weighted mean of previous stimuli. In addition to recent experiences, which act as an immediate referent, a longer-term process, linked to conditioning, establishes more enduring comparators, consisting of memories ordered in terms of intensity. Hence, numerical descriptions are a proxy for describing where, in a sequence built and anchored by comparisons, a particular experience lies.

Those theoretical ideas apply directly to affect, as Helson himself noted. He reported evidence that affective estimates are relative to an adaptation level. That has an obvious implication: it is very doubtful to treat equal ratings at different points in a rating session as if they meant the same thing. On a more abstract level, he considered it obvious that there are different kinds of pleasantness or unpleasantness. However, the different experiences can still be ordered, and hence they can be treated as dimensional for some purposes. 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 [24].

Related ideas feature in many later theories. They include *relative judgment models* [25], [26] suggesting that experience with stimuli gradually creates our internal context, and discussions of *anchors* [27], against which we rank any new 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 [28].

There is at least one account of that kind which deals in depth with issues closely related to affect labeling. Stewart et al. [28], extended theories of relative judgment to economic decision making. Their work 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. That theoretical tradition provides a valuable perspective on the issues that are considered in this paper. The argument is emphatically not that the evidence forces us to accept that kind of picture. The point is that the problems of data collection within affective computing are not unique, and not necessarily technicalities. There is good reason to ask whether they point to questions at quite a deep level about the way human affect works.

Given the theory, it is perhaps surprising that psychology does not offer much direct empirical evidence on levels of measurement. That may be because so much work has focused on nominal responses (i.e., emotion categories). However, at least one series of studies compares ordinal and interval responses to affective stimuli, and, in line with the theory, it shows a clear advantage for pairwise comparison.

The comparisons involve emotion intensity, reported by using either a numerical scale, or graded pair comparison. The scales reconstructed from graded pair comparisons are more satisfying than the numerical estimates in multiple ways. They show higher reliability (alpha coefficients $0.94 - 0.98$ as against $0.82 - 0.87$) [29]. They also agree better with theoretical predictions [22]. The reconstructed scales also satisfy tests of metric structure [29], suggesting that the problem with interval measurement lies not in the underlying experience, but in the task of externalizing it. Note, though, that that applies to intensity: Table 1 indicates that it stands out from other dimensions (particularly in agreement on rises and falls). It cannot be taken for granted that the same holds for other scales.

The concepts underlying this paper have been explored in several other disciplines beyond psychology which we cover briefly in the remainder of this section. The results and evidence from these disciplines are also relevant for the study of emotions.

2.4 Marketing and Social Psychology

The fundamental role of affect for human behavior in general and decision making, in particular, has been at the core of behavioral and social psychology. Experiments by Zajonc et al. [30] attempted to shed light on the relationship between affect and preference and revealed that a mere exposure to certain options is sufficient for a subject to develop a positive preference for those options. It seems that the more often those options are presented to us the more we tend to prefer them and react positively.

The study of values in *social psychology* has traditionally been linked to the study of surveys that would be able to capture appropriately the perceptions of such social values [31]. Theoretically, the stance has been that rankings are a superior method as social values (being subjective constructs) are inherently comparative and competitive [32]. As that suggests, *relative choice* among options appears to be the best underlying mechanism for capturing concepts that are subjective [33]. One of the largest concentrated efforts within social psychology [31] reveals that the latent variable structure of the two measures (ratings vs rankings) are different.

In *marketing* research values are traditionally measured with the use of rank-based questionnaires [32]. As 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 his/her hierarchy of values in a self-reporting manner [32]. Numerous studies have attempted to find the golden standard for assessing subjectively defined notions in marketing research by comparing the most popular measurement approaches: ratings and rankings. The empirical evidence in that area is strong. For instance, a large scale study involving over $3,500$ students across 19 counties [34] 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. Similarly to marketing and social psychology, the paired comparison method is dominant in studies involving image perception which attempt to assess visual discomfort, viewing experience and stereoscopic image quality of various multimedia types [35], [36], [37].

2.5 Behavioral Economics

Foundational to behavior is decision making and foundational to decision making is judgment. How do we value different options and decide to take one of the many? Several theories of psychology have relied on the central concept of *distance* among options: when one compares two options she naturally assesses how close they are to each other. It turns out however that the comparison is not always symmetric. For instance the distance between A and B is not the necessarily the same as the distance between B and A. This phenomenon is what Tversky called *features of similarity* [38] according to which people compare things across a number of noticeable features, not just one. The number of similar noticeable features determines the degree of similarity.

Conventional decision-making theory—such as expected utility theory and foraging theory [39], [40]—represents the values of possible actions in some normative form and assumes that options are evaluated in an absolute manner, without considering other alternatives. Evidence from behavioral psychology, however, suggests that the valuation of choice depends largely on the composition of the options available. Naturally when people are faced with a large set of options they report increasing difficulty in assigning values to all. Examined extensively, both animals and humans seem to rely on context-dependent preferences that reflect the particular alternative options they have available [41], [42].

Another foundational aspect of behavioral economics is the principle of humans acting as *irrational agents* when they take decisions. The judgment heuristics theory introduced by Kahneman and Tversky [43]—which is built on the bounded rationality theory of Simon [44]—captures the various factors that make us act irrationally. Such factors, named *heuristics* by Kahneman and Tversky, were carefully studied and reported in a series of papers that had a major impact across several disciplines. The first heuristic in support of our thesis is *representativeness* [45]: according to that heuristic when people make judgments they compare the value of the option to be judged to an internal model they maintain in their brains. Another heuristic relevant to the ordinal approach is *anchoring* [43] which is the bias our brains maintain systematically against any decision we make or problem we attempt to solve. It is the context (social, emotional, spatial, temporal, and so on) which defines a reference point against which we evaluate our options. *Framing* [46], similarly to anchoring, is the way we make decisions based on the way options are presented to us: if for instance a reward over a bet dilemma is inverted and instead is presented as a loss over a bet dilemma we will most likely accept the reward in the former case and the bet in the latter case. Overall, judgment heuristics suggest that when humans assess options they do not assign a value to their options but they rather rely on a *change*. Kahneman put it simply as follows: "*...it is safe to assume that changes are more accessible than absolute values*" [47].

2.6 Neuroscience

In *neuroscience*, Damasio [48] 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.

The extensive study of the orbitofrontal cortex (OBC) is particularly relevant to the message of this paper since it has been correlated to rewards and their processing. Brain activity in OBC seems to be a good predictor of the motivational value of a reward [27], [49]. There is also evidence in monkeys [50] suggesting that their brain—in particular OBC—encodes values in a relative fashion. Further evidence show that neurons in the monkey lateral intraparietal cortex (LIP) encode a relative form of saccadic value which is explicitly dependent on the values of the other available alternatives [51]. Similar results have been reported for the human medial OBC [52] and LIP [53].

Further neuroscientific evidence that supports our ordinal stance can be found in studies investigating the *divisive normalization* phenomenon. Divisive normalization is the input-related activity of a neuron which is divided by the summed activity of a large pool of neighboring neurons [53]. The phenomenon has been widely observed in sensory systems—explaining responses such as contrast gain control in the retina—but recent neurophysiological evidence shows that divisive normalization extends to higher-order cortical areas of the brain which are involved in decision making [53]. In particular the firing rates of a neuron are increased by increases in the value of the represented option whereas they are suppressed by increases in the value of the alternative options available [51]. Within decision-making divisive normalization creates context dependence: our brain's encoding of an option's value is explicitly dependent on the value of other available alternatives [53].

2.7 Artificial Intelligence and Machine Learning

The notion of *preference* is nowadays central in *artificial intelligence* and machine learning [54]. The theoretical grounding of learning from preferences [55] is based on humans' limited ability to express their preferences *directly* in terms of a specific value function [56]. 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 valuing 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 available options.

Centrally to the message of this paper, rating, class and ranking data of a survey or an annotation process can be viewed as different forms of expressing a preference about a subjective notion. A preference can be seen alternatively as the building block of a global ordinal relationship that exists among the various instances of the notion we attempt to capture. The set of preferences available can, in turn, encapsulate the underlying phenomenon; be it an emotion, an opinion or a decision. In Section 3.3 the paper provides an extensive discussion about the use of *preference learning* as a way to model and predict preferences, or ordinal data at large.

2.8 Summarizing the Background

Section 2 noted the problems associated with nominal and interval approaches to emotion annotation. Logically, that makes it natural to consider ordinal approaches; but there would be grounds to be wary if that meant stepping into unknown territory, theoretically and practically. In fact, reviewing related disciplines shows a very different picture. This section indicates that they offer a wealth of relevant material, which offers practically useful models at multiple levels, and reaffirms the intellectual case for thinking of emotion in ordinal terms.

3 Affective Computing: An Ordinal Perspective

In this section we view affective computing from an ordinal perspective and we detail the various phases involved in this process. In particular, we present the ways by which annotation data can be collected (Section 3.1), the resulting data that can be processed (Section 3.2), and ultimately the methods that we can use to analyze and model data (Section 3.3). To better illustrate the material presented in this section the reader is referred to the three phases (columns) depicted in Fig. 2; each phase corresponds to a subsection.

3.1 Annotation: First-order vs. Second-order

Prior to delving into the details of data analysis and machine learning from an ordinal perspective we first need to view the process of obtaining reliable and valid ground truth

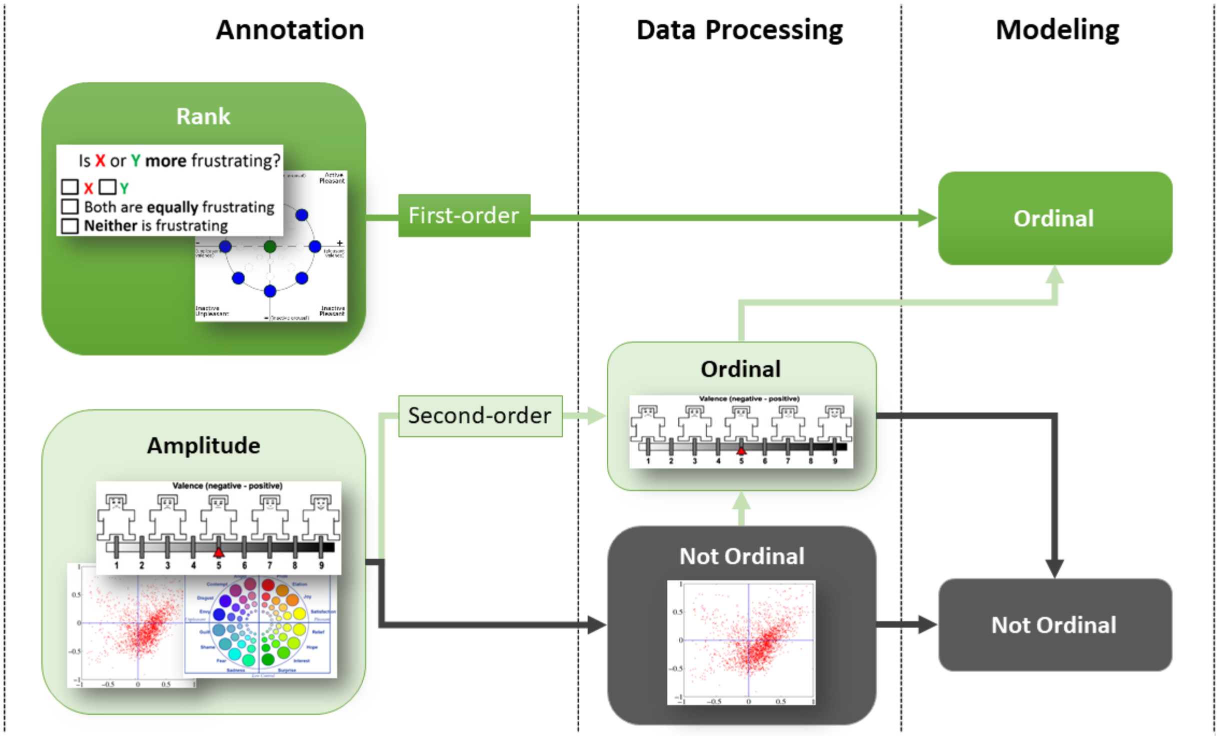


Fig. 2. An ordinal-centric taxonomy of approaches through the phases of annotation, data processing and data modeling. With dark green (gray in print) and light green (light gray in print) color, respectively, we illustrate the *first-order* and *second-order* approaches for the ordinal analysis of annotations. With gray color (dark gray in print) we indicate data processing or modeling approaches that do not follow the ordinal perspective. Sections 3.1, 3.2, and 3.3, respectively, cover the phases of annotation, data processing, and modeling.

labels. We can distinguish two approaches that ultimately lead to an ordinal analysis of data: the direct, or *first-order* approach, and the indirect, or *second-order*, approach. In this section we discuss the core differences between a first- and a second-order approach to emotion annotation.

Given our ordinal-centric perspective we identify two main ways through which one can collect annotation data (see Fig. 2).

- **Rank**: Annotators are asked to give a preference among two or more options or a *rank*. In this case the labels are relative and hence the annotation data is ordinal. An example of a time-discrete rank-based annotation scheme is the forced-choice questionnaire that requests responders to rank two or more options; a popular variant is the 3 alternative forced-choice questionnaire asking the annotator to either report her preference over two options, or instead report that there is no noticeable difference between the options. Another example of a time-continuous rank-based annotation tool is *AffectRank* [57]—a rank annotation variant of FeelTrace—that is covered in detail in Section 4. Given that the data collected are ordinal by nature, the analysis may rely on non-parametric statistics and preference learning as suggested in Section 3.3. We refer to this direct approach of annotation data collection and analysis as *first-order* and we represent it with dark green (or gray in print) color in Fig. 2.

  As mentioned extensively in Section 2 the first-order (ranking) approach is dominant in domains such as marketing, decision making, preference handling, computational social choice, recommender systems, machine learning, and combinatorial optimization. For instance, more than $3,000$ datasets are included in the $\{$PrefLib$\}$ library of preferences for public use[1]; examples include the Sushi dataset which contains data about the preferences of people over variations of Sushi[2] and the LETOR [58] package of datasets for research on learning to rank within information retrieval.[3] In affective computing the approach has given us a number of accessible datasets such as the *Platformer Experience Dataset*[4] which contains visual cues, behavioral data and pairwise preferences of the experience of Super Mario Bros players [59] and the *Mazeball*[5] dataset which contains physiological responses (skin conductance and heart rate variability), behavioral data and pairwise experience preferences of puzzle gamers [60]. There is also a dataset containing rank-based affect annotations of sound effects [61] (*Sonancia Crowdsourcing DataSet*[6]).
- **Amplitude**: Annotators are asked to give an *amplitude* or a magnitude label to the stimulus they are

[1] Available at: http://www.preflib.org/
[2] Available at: http://www.kamishima.net/sushi/
[3] Available at: https://www.microsoft.com/en-us/research/project/letor-learning-rank-information-retrieval/
[4] Available at: http://ped.institutedigitalgames.com/
[5] Available at: http://hectorpmartinez.com/
[6] Available at: http://www.autogamedesign.eu/sonancia

presented with. The amplitude can represent a continuous (interval) value—such as an arousal trace—a class (nominal)—such as a categorical label for a facial expression—or even an ordinal value such as a Likert item response.

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 (first-order) would seem to be the *ideal* approach. That is not always possible, however. By looking at the alternative way of obtaining annotations (i.e., *amplitude*) we may distinguish four possibilities for their analysis (see Fig. 2):

- Annotations are inherently **ordinal**, such as responses from SAM and Likert questionnaires or the Geneva Wheel Model, and:
  - following the **second-order** approach (see Fig. 2) they are processed as ordinal data (see Section 3.3), or
  - they are not processed as ordinal data
- Annotations are **not ordinal**, such as arousal traces or categorical labels of facial expressions, and:
  - following the **second-order** approach (see Fig. 2) they are processed as ordinal data given that an underlying order exists within the annotations, or
  - they cannot be processed as ordinal data.

Within the second-order approach—as indicated by the light green color (or light gray color in print) in Fig. 2—it is recommended that whenever possible we treat amplitude data of any type as ordinal and process them this way. In the following sections we discuss the various aforementioned data types and corresponding analysis in further detail.

3.2 Data Processing

According to our thesis if ordinal data is available (first-order) or if the data has a meaningful underlying order (second-order) then naturally the analysis should follow the ordinal path. But, how should we process other data types, and how should we machine learn from data types that are not necessarily ordinal? Following the taxonomy of Stevens [62] we can distinguish three data types that we can obtain from an emotion annotation task: *interval*, *nominal*, and *ordinal*. The first two types are *not ordinal* in principle but, under particular assumptions, could be converted to ordinal data and processed as such (see middle column of Fig. 2).

The remainder of this section explores what the thesis of the paper implies for the various data processing 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* practices associated with each. These practices are depicted in Fig. 3, respectively, as white, dark gray and light gray table cells.

3.2.1 Annotations Are Interval (Not Ordinal)

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 [17]. The most popular rating-based question is a Likert item [63] 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 [64] and the Self-Assessment Manikin [65]. When annotations come in an interval form we can treat them as such or alternatively treat them as nominal or ordinal data.

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 (see top left cell of Fig. 3). 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 is problematic for subjective constructs such as emotions: it misrepresents 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 (see the two arousal classes at the middle left cell of Fig. 3). This is another dominant practice in affect modeling (e.g., see studies on the SEMAINE dataset [66]) but recent evidence suggests that such practice introduces a multitude of biases in data and thus takes us further away from the underlying ground truth [60]. Furthermore, creating dichotomized labels from interval data creates unavoidable problems where similar samples around the boundary are artificially placed in different classes [67].

Any attempt to derive an ordinal scale from interval data that characterize subjective notions appears to be a good practice to follow [68], [69], [70]. In the example of the bottom left cell of Fig. 3 the arousal values are not considered; instead the points across the arousal dimension are ordered based on their trace value. Several studies have transformed values of affect to ordered ranks and then derived affect models via preference learning. As we will see in Section 4 such a transformation improves cross-validation capacities [60], [71], [72].

3.2.2 Annotations Are Nominal (Not Ordinal)

The second annotation data type one may obtain comes in *nominal* (or class) form. Nominal data are mutually exclusive labels which are not ordered and can be stored as words or text (e.g., Male, Female) or given a numerical label (e.g., Male is 0, Female is 1); it is important to note that numbered labels do not (and should not) imply that there is an underlying order. Nominal data, however, 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 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.

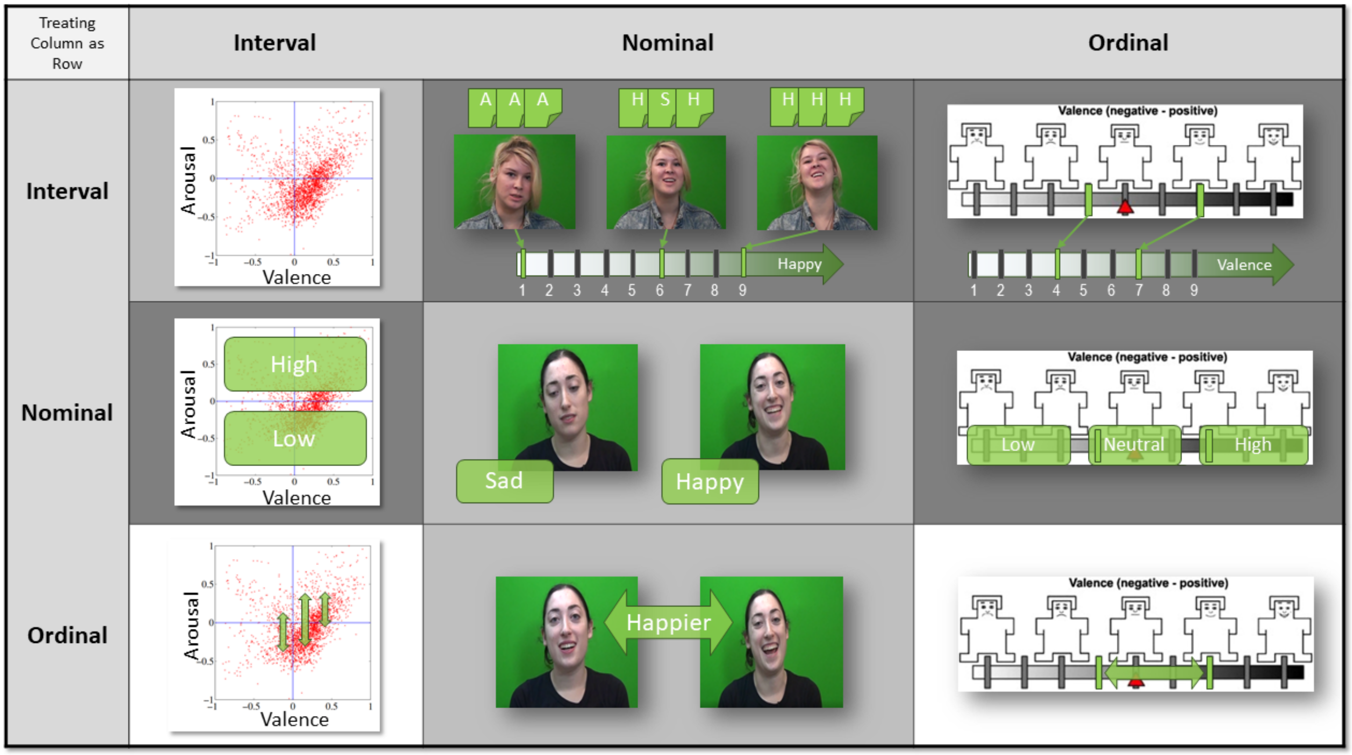


Fig. 3. Processing data in affective computing: Treating column data types as row data types. White, dark gray, and light gray table cells, respectively, illustrate the *good*, *bad* and the *ugly* practices according to the thesis of this paper. By good we refer to approaches that are theoretically 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.

Deriving interval scores out of nominal values seems flawed unless there is an underlying order across the classes or nominal data can be aggregated somehow (see top center cell of Fig. 3). An attempt towards that direction is presented in [73]; 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 idea is illustrated at the top center cell of Figure 3, where three videos are annotated by three evaluators. The first video was consistently annotated as anger (i.e., A, A, A); the second video received two votes for happiness and one vote for sadness (i.e., H, S, H); the third video was consistently annotated as happiness (i.e., H, H, H). If *happiness* is the target emotion, these three videos can be mapped into an interval score by considering their individual evaluations. The third video was consistently annotated as happiness, therefore, it is mapped to the positive extreme of the interval. The second video received two votes for happiness, so it is mapped lower than the third video, but higher than the first video, which did not receive any vote for happiness. 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 (see middle center cell of Fig. 3). 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 [9]. 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 [74].

An order cannot be easily derived from classes which are unordered—e.g., happiness and sadness (see bottom center cell of Fig. 3). Indicatively, Lotfian and Busso [73] 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. [75] 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, first-order, approach instead: to ask annotators to rank samples directly (e.g., is sample A happier than sample B?).

3.2.3 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 [70], sounds [61], video screenshots [57], [76], or videos [77]. 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 SAM is ordinal by nature [17]. Such data is generally treated as interval values, however—for instance, by averaging the obtained annotation values; see the example at the top right cell of Fig. 3 in which two SAM valence labels are assigned numbers (4 and 7). 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 [17], [68], [78].

Another popular practice is to treat ordinal data as nominal and view the problem as classification. In the example of the middle right cell of Fig. 3 the SAM valence scale is converted to three classes (low, neutral and high) and hence the first label belongs to the neutral class whereas the other label belongs to the high class. Recent studies, however, compared the use of ordinal labels as they are against the use of ordinal labels as classes and showcased the benefits of the former in yielding more general affect models [60].

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 [60], [72] and reliability [57]. In the example of the bottom right cell of Fig. 3 we only know that the right label has a higher value of valence than the left label; however, the value difference between them is unknown. The studies presented in Section 4 provide additional evidence for the superior nature of relative affect annotation and its analysis for affect modeling.

3.3 Modeling

Independently of which approach one follows to obtain ordinal labels (first- or second-order) data is ultimately stored in a rank or pairwise format and ready to be analyzed statistically or derive affect models from. A popular objection against the use of ordinal labels is the lack of statistical tools and methods to process them—note that section 5 addresses this and other common objections directly. As a response to this objection this section outlines the palette of data analysis tools and statistical methods available for the processing of ordinal data.

3.3.1 Statistical Methods for Ordinal Data

Standard data summarization approaches based on averages or standard deviations are strictly not applicable on ordinal data. Instead, approaches for analyzing ordinal data should rely on *non-parametric* statistics such as *Spearman's* rank correlation (e.g., as in [77]). It is important to note that Norman [79] showed empirically in one dataset that Pearson's correlation is robust enough when compared against Spearman's rank correlation. Such evidence could support the use of standard parametric correlation tests that treat ordinal data as interval values. Nevertheless such practice ignores the nature of the data and instead views ordinal labels of emotion as magnitudes of emotion, hence adding bias to the underlying ground truth [60].

Considering statistical factor analysis for ordinal data one may use the *Wilcoxon signed-rank test* [80] which bypasses inter-personal subjective differences by comparing only within-participant ranks. A common alternative is *Kendall's* τ [81] that can be used to calculate the correlation between the hypothesized order (e.g., A is happier than B) and the obtained labeled ranks—see e.g., [60]. Further the *Mann–Whitney* [82], the Kruskal–Wallis [83] and Friedman's [84] tests for three (or more) groups of ranks are also directly applicable to ordinal data. Finally the Bradley-Terry model may be considered which uses a linear function mapping paired preferences to probabilities within an interval scale [35].

When it comes to the estimation of inter-rater agreement—a typical analysis when several raters are involved—Cronbach's α [1] is the dominant coefficient in the affect annotation literature. Cronbach's α, however, is not applicable to ordinal data and therefore cannot be used to estimate the agreement across annotators' labels that are either first-order ordinal or they are processed as ordinal in a second-order manner. Krippendorff's α [85], on the other hand, is a rather generic statistic that can measure the degree of agreement among annotators through several annotation types (including nominal, ordinal, and interval) and it is also able to handle missing data. It is thus recommended as a versatile statistic of inter-rater agreement for annotators who can either label, categorize, rate, or even rank stimuli in terms of emotion. The measure has been used in a number of studies which compare the inter-rater reliability of ordinal labels against other label types [57], [86].

In addition to a simple statistical analysis one may wish to machine learn the ordinal data. The next section focuses on preference learning algorithms [54], [55], the natural approach to process ordinal data and derive models from this data.

3.3.2 Preference Learning

Preference learning (PL) is a subfield of supervised learning dedicated to the processing of ordinal labels. The preference learning paradigm as an approach for affective modeling was first introduced by Yannakakis in 2009 [69]. Since then numerous studies in affective computing have used preference learning for affect detection and retrieval through images [87], [88], [89], [90], videos [76], [91], music [70], [92], sounds [61], speech [60], [75], [93], games [60], [94], [95] and text [96].

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 preference learning problem by using an appropriate formulation. Linear statistical models, such as ordinal regression, linear discriminant analysis and large margins, and non-linear approaches, such as Gaussian processes, shallow and deep artificial neural networks, and support vector machines (SVMs), are directly applicable.

A popular derivation for preference learning 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 preference learning method for SVMs [97].

An alternative formulation for preference learning 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 [98], or can be learned from data using deep learning structures as performed via convolutional neural networks [89], [94], [99], via RankNet [100], or via neuroevolution [60], [69]. Studies have demonstrated that all aforementioned methods provide compelling results [69], [73], [89], [90], [93], [99], [101]. For the interested reader, a number of preference learning methods including RankSVM, neuroevolutionary preference learning and preference learning via backpropagation are contained in the preference learning toolbox (PLT) [102]. PLT is an open-access toolkit[7] built and constantly updated for the purpose of easing the processing of ordinal labels. Similar arguments can be made for labeling relative scores, where approximations can be made to reduce the number of pairwise comparisons to annotate n different samples.

3.3.3 Global Ranking via Pairwise Preferences

It is worth noting that for n different samples, a direct evaluation of all possible pairwise comparisons involves $\frac{n(n-1)}{2}$ assessments. As n increases, the number of comparisons becomes unfeasible ($O(n^2)$). Inferring the global ranking out of these comparisons can be formulated as a sorting problem with noisy pairwise evaluations. There are various algorithms to approximate the global ranking using a subset of the possible comparisons. Some of these methods are directly applicable in affective computing, through which the required pairwise comparisons between the samples are dictated by the algorithm; see for example the work of Jamieson and Nowak [103]. Other methods rely on pairwise comparisons between randomly selected samples; see for example the work of Wauthier et al. [104]. The complexity of sorting algorithms is $O(n \log_2 n)$. Under realistic assumptions, it is possible to approximate the ranking using less comparisons. For example, Jamieson and Nowak [103] suggested that if the samples can be embedded in a lower d-dimensional Euclidean space that respects the ranking between samples, the full ranking can be obtained with $O(d \log_2 n)$ actively selected comparisons. In practice, studies have successfully estimated global ranking even by sampling from possible pairwise comparisons [105].

3.3.4 Preference Learning Applications for Affective Computing

Any application in emotion recognition can be formulated as a ranking problem in which preference learning 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 [106]. 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 [73]. Applications of emotional retrieval include solutions for health care domains [107], [108]. In longitudinal studies relying on remote assistant technologies, rank-based emotion retrieval can provide an ideal framework for a 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 [109]. Emotion-aware recommendation systems are also an important application area for preference learning using ordinal labels (e.g., selecting music or sounds conveying emotions that match the current affective preference of the user [61], [70]).

The breadth of applications expand to video-based [57], [76], [77], image-based [35], [37], [37], speech-based [106], music-based [70], [110] or physiology-based [111] emotion recognition for health, educational or entertaining [78] purposes. The next section covers a few successful applications directly showcasing the benefits of ordinal annotation and processing for affect modeling.

4 Affective Computing Case Studies

Ordinal conceptions have rarely been identified as a major issue in affective computing, but there has been a gradual growth of published studies where ordinal techniques were used. This section offers the first reasonably comprehensive overview of those studies. We emphasise affect annotation studies that *compare* ordinal annotations *against* other annotation forms (e.g., class-based or interval-based protocols) within the domains of video, face, body, music and sounds, speech, and game experience annotation. We begin in Section 4.1 by making explicit the measures that we use to compare approaches, i.e., *reliability* and *validity*. Those measures are then used to consider relevant studies, grouping them by domain. Section 4.8 then integrates across all the domains, and summarises the evidence of comparative benefits for ordinal labels.

4.1 Performance Measures

In the previous sections we argued for the advantages of ranks as an emotion annotation tool. Thus, naturally, our thesis depends on measures of performance that show such advantages. In many disciplines such as engineering

[7] Available at: http://plt.institutedigitalgames.com/

and in several problems with objectively defined reference points (ground truths), accuracy can be measured as the difference between the data observations and the reference value. However, such a measure is not available for subjective notions, as the underlying ground truth is unknown. Turning to affective computing, a necessary starting point is identifying ways to measure the performance of annotation. Two kinds of measure are available, involving *reliability* and *validity*. Those notions in psychometrics are directly linked to the notions of precision and accuracy, respectively. The first estimates the degree of repeatability of the observation whereas the latter estimates the proximity of an observation to the underlying true value (ground truth). Our thesis takes both into account.

4.1.1 Reliability

Reliability is a standard concept in quantitative research methods, and it is a major concern whenever human respondents are used to provide data for analysis. It has already been referred to in Section 2. Two different types of reliability are relevant to this paper: inter-rater and test-retest reliability.

Inter-rater reliability is the degree of agreement among a number of annotators. Estimates of inter-rater reliability yield a score of consensus across the answers (e.g., ratings) of all annotators that participated in a study involving a particular construct. In contrast, when the same annotation task is given to the same annotator at different times, there is a need to test whether the subject is consistent with regards to the construct across the time instances (test-retest reliability). Test-retest reliability is also relevant for the thesis of this paper. Both reliability measures yield superior results for relative (rank-based) annotation, compared to other absolute annotation methods, as showcased by the case studies detailed in the remainder of this section.

4.1.2 Validity

The validity of annotation labels is the degree to which the annotation measures the phenomenon we claim it does (in contrast, reliability measures the degree to which our observations agree with each other). Several types of validity are available but, in this paper, we will focus solely on empirically measured validity, defined as criterion validity. 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 detailed in this section, and others in the literature showcase the superior generalizability of ordinal approaches to modeling affect.

Other measures of validity should be mentioned. Physiological measures have been used in affective computing [112] and they have an obvious appeal, because it is easy to think of the changes that they measure as ground truth. They certainly have a contribution to make. However, on the standard philosophical view mentioned earlier, the physiological changes involved in an emotion are part of a structured whole, and it would be logically confused to treat a measure of the part as ground truth for the whole. An interesting alternative is also illustrated in work mentioned earlier: it is agreement with theory. Where that has been used, it favors ordinal techniques [22]; however, few theories are robust enough to provide that kind of test. Both avenues are interesting for the future.

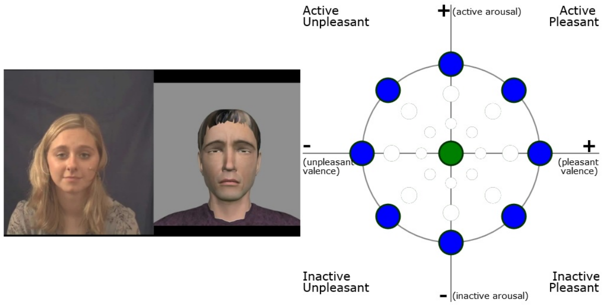


Fig. 4. *AffectRank*: the rank-based annotation tool introduced in [57]. *AffectRank* is inspired by FeelTrace but it allows the real-time annotation of arousal and/or valence in a relative fashion. *AffectRank* is freely-available at https://github.com/TAPeri/AffectRank.

4.2 Videos

While Metallinou and Narayanan [68] and Soleymani et al. [77] 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* [57] is a freely-available, rank-based version of FeelTrace which asks the annotator to indicate a *change* in arousal and/or valence while watching a video. As seen in Fig. 4 the annotator has 8 options to pick from (blue circles) during annotation based on positive or negative changes of arousal and/or valence. In that regard, the labels obtained are discrete events (changes of arousal/valence) in time. The evaluation study of [57] compared the inter-rater reliability between FeelTrace and *AffectRank* for the video annotation of two datasets: the SEMAINE [113] 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. 4). *AffectRank* yields superior reliability even when FeelTrace ratings are treated as ordinal data. The key findings of [57] 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.

While *AffectRank* focuses on a first-order approach for analyzing ordinal labels, *RankTrace* [114] is an affect annotation tool that provides a continuous trace for second-order analysis. Similarly to Gtrace [115], *RankTrace* allows for the annotation of one affect dimension at a time (see Fig. 5). *RankTrace*, however, *does not constrain* the user within annotation bounds (e.g., within [-1, 1]) as typically practiced in affect annotation via traces, and it uses a wheel-like hardware as a more natural means of user interfacing with continuous annotation [116]. The labels obtained via *RankTrace* are continuous unbounded time series (i.e., affect traces). In [114] a number of players used *RankTrace* to annotate their tension levels by watching their video-captured playthroughs of a horror game [61], [117]. During the game their skin conductance was also measured as an estimate of the ground truth of tension. Based on the annotation traces

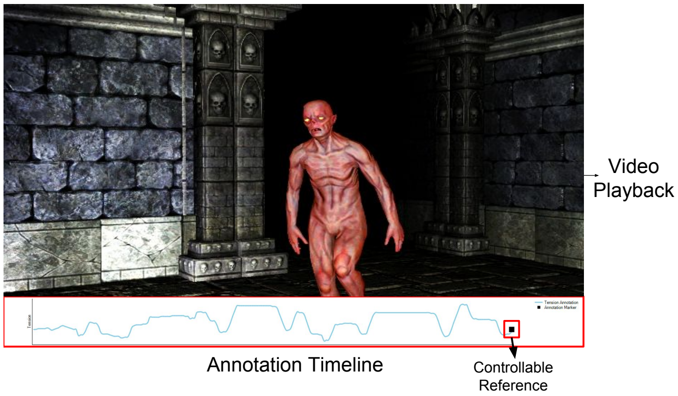


Fig. 5. The *RankTrace* tool introduced in [114] is inspired by GTrace and allows participants to annotate their emotional experience using a wheel-like controller in real-time, while watching a video. *Ranktrace* is freely-available at http://www.autogamedesign.eu/software.

and the ground truth data Lopes et al. [114] compared two types of approaches for the second-order analysis of data: approaches that treat the trace as *absolute* values (mean and integral within a time window) against *relative* approaches which rely on changes in the trace (amplitude and average gradient). Results from a rank correlation analysis show that annotation features which assess relative annotation changes within the window are better and more robust linear predictors of the ground truth. Such findings suggest that treating a continuous annotation signal in a relative (ordinal) fashion, e.g., via the its gradient, yields features of higher predictive capacity and, hence, validity. The core findings of Melhart et al. [118] in the AMIGOS dataset [119] are also in alignment with the key message of the above studies. The ordinal transformation of SAM labels in that study yields more accurate predictors of arousal (as elicited through short videos) than the predictors built on classes of SAM labels.

4.3 Face

Another domain of the ordinal approach via the use of preference learning in affective computing is in face analysis. Baltrušaitis et al. [105] proposed to estimate the intensity of facial action units (AUs) using rank-based methods. The underlying assumption is that the ordinal comparison of the intensities of AUs is more reliable than the estimation of absolute scores. To establish pairwise preferences the method considers pairwise comparisons between two images of the same subject that in turn determine if the intensities are either equal or different. The local ranking is then formulated as a multi-class classification problem implemented with SVMs (equal, lower or higher). A strength of the approach is that the local ranking does not consider the difference between the values of the AU intensities. Instead, it only considers their relative order, thereby increasing the robustness of the approach. The local pairwise comparisons are aggregated to create a global score describing the intensity of the AUs using a Bayesian model. The key results of that study demonstrate that the framework based on local rankings performs higher than other state-of-the-art methods. The performance gain was particularly high for AUs that are less frequent in the corpora. The experimental evaluation also showed that the ranking method generalized better across databases. Similar conclusions on the across-dataset generalization capacity of ordinal over standard interval or class labels were reached in [118], [120].

Another interesting research direction is to detect changes in facial expression. For example, Khademi and Morency [121] proposed to detect changes in AU within neighboring frames. The problem was formulated as whether the expression increased, decreased or remained the same. The approach achieved better performance over absolute methods, showing robustness against individual differences.

In a similar vein, Walecki et al. [90] designed an ordinal regression model with the aim to learn and infer jointly the intensities of multiple AUs. The ordinal approach to AU intensity estimation outperforms independent modeling of AU intensities as well as the state-of-the-art classification approaches of AU intensity estimation. Similar findings showcasing the supremacy of ordinal analysis were obtained via variational Gaussian process auto-encoders [122]. The models in that study were constructed in a supervised manner by imposing the ordinal AU intensity labels to the manifold.

4.4 Body

At the intersection of game-based affective interaction and body-based affective interaction we meet a number of studies showcasing the benefits of ordinal annotation for emotion detection. In particular, Kim et al. [123] used an ordinal method to annotate entertainment during collaborative child play; the derived entertainment models relied on the non-verbal features of turn-taking and body movement during play. Initially annotators were asked to provide levels of engagement in a nominal fashion (low, medium, high) but doing so resulted in poor inter-rater agreement. However, when the annotation questionnaire considered relative levels of engagement the inter-rater agreement was improved significantly. Further a ranking-based approach for training SVMs on the ordinal labels outperforms significantly conventional SVM classification which is trained on nominal labels. The same group in a follow up study [124] introduces a ranking algorithm that fuses characteristics and temporal dynamics of ranks by combining the omission probability of each rank and the transition probability between ranks in time. The algorithm yields even higher validation accuracies on the task of body-based engagement detection.

Using a corpus of abstract body postures, Pasch et al. [125] showcase that there is a high correlation between pairwise preference and rating labels for annotating different affective states. As found in Yannakakis and Hallam [78], however, there are considerable mismatches that are not investigated further.

The study of Rienks and Heylen [126] conducted ordinal annotations of dominance for small group interactions. The analysis showed high consistency between annotators. However, the ordinal annotations were not compared to other alternative absolute methods.

4.5 Speech

Recent work in speech-based affect recognition has demonstrated the benefits of using preference learning with ordinal labels [71], [72], [73], [93], [127]. Using time-continuous evaluations for arousal and valence provided by FeelTrace, the above studies defined preferences between pairs of speech samples and compared preference learning (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. 6a). The evaluation also revealed two important observations. First, preference learning 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 [93], [94].

Within the domain of speech-based emotion detection it is worth noting that there have been attempts for developing annotation tools that consider the relative change of the emotion's intensity over time [128]. Some comparative results of the use of such tools demonstrate that ordinal representations are more reliable than nominal representations for emotion labeling from spontaneous speech. Siegert et al. [129] investigate contextual factors that may contribute to the validity of the label annotations of speech and showcase that the knowledge about past is needed to assess the affective state. In agreement with Yannakakis and Martinez [57], they further show that the inter-rater reliability of ordinal labels is higher than the agreement achieved with nominal labels [86].

In another study, it was proposed to define ordinal labels by considering trends in the time-continuous labels [72]. 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 [12] 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. 6b). The arrows denote 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. 6b). The core findings suggest that extracting ordinal labels with QA provides better classifiers, increasing the accuracy of the emotion rankers.

Other related studies have formulated speech emotion recognition problems as detection of changes in the emotional content [130], [131] and detection of deviations from neutral patterns [106], [132].

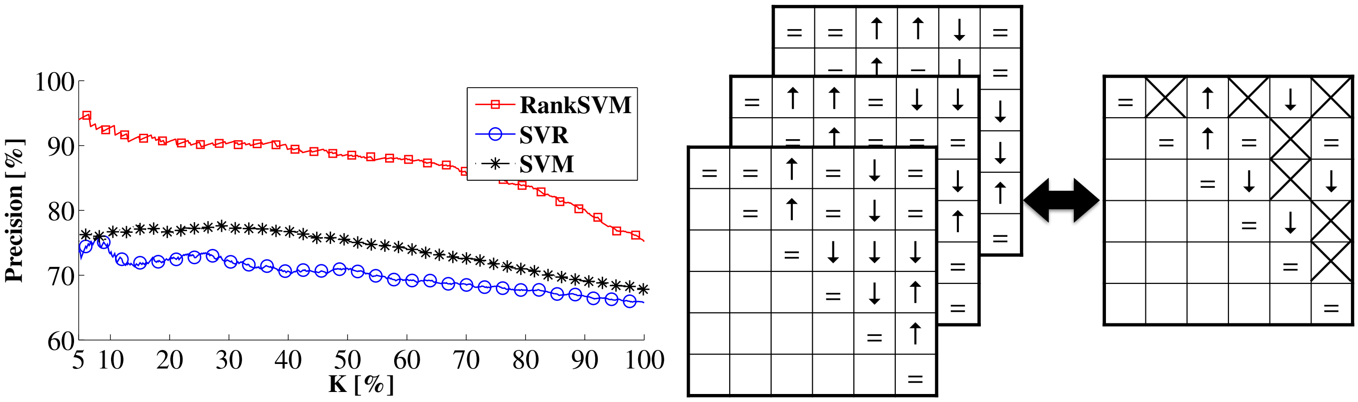
4.6 Music and Sounds

As much as with the other types of stimuli, the emotional intensity of music has been found to be best annotated via ordinal approaches. The work of Yang et al. [70], [110] is seminal in this domain of affective computing. In their studies they used a ListNet approach [133]—that employs a non-linear, radial basis function ANN—to learn to rank the emotional intensity of music samples. In contrast to pairwise preferences ListNet uses lists directly as learning instances and attempts to minimize the deviation between the ground truth ranking and the estimated one. The music corpora in Yang et al. [70], [110] consisted of a large set of famous English and Chinese pop songs and feature extraction relied on the melody, timbre and rhythm of the songs. Affect was annotated using both a rating and a ranking approach along the dimensions of arousal and valence. The key findings of these studies reveal that ranking is easier to use for annotating songs compared to a Likert scale and that its test-retest reliability is higher than that of rating scales. Most importantly, the preference learning algorithms employed perform consistently well for all datasets examined, they are more robust with regards to parameter tuning and, finally, they are more efficient to conventional methods (e.g., support vector regression) that are trained directly on the rating labels. The benefits of the rank-based approach of Yang et al. for music annotation have also been adopted by recent studies that examine the modeling of emotion as elicited by sound effects [61] and music videos [118] of the DEAP dataset [134].

4.7 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 [95], [135], prey-predator [17], [57], [60], horror [61], and physical interactive games [78] among many other game genres. Most notably for the purposes of this section, Yannakakis and Hallam compared rating versus ranking annotations of first person experience in both a prey predator and a physical interactive game [78]. Their subjects were asked to use 5-point (used in the prey-predator game) or 10-point (used in the 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 (i.e., ranks yield higher degrees of test-retest reliability).

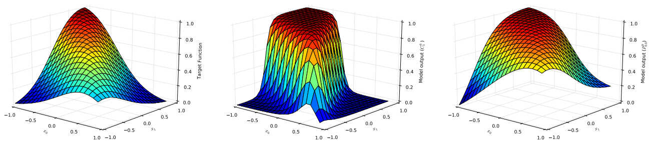
In another study, Martinez et al. [60] worked on the hypothesis that the best way of analyzing ratings of affect is to treat them naturally as ordinal data. To test their hypothesis they compared models of affect that are the result of converting ratings to classes (classification) versus ordinal models



(a) Rankers versus other methods

(b) Finding trends in traces

Fig. 6. Case studies on speech: (a) improved *precision at K* (P@K) of RankSVM over regression (SVR) and binary classification (SVM) for arousal [71], (b) QA framework to identify trends from emotion annotation traces [72].



(a) Ground truth

(b) Classification

(c) PL

Fig. 7. A hypothesized (artificial) ground truth function (z-axis) which is dependent on two attributes, x_1 and x_2 (Fig. 7a), the best classification model (Fig. 7b) and the best preference learned model (Fig. 7c) [60].

that are trained directly via preference learning (second-order analysis). They used three datasets for their analysis: an artificial dataset, a dataset from the MazeBall game containing physiological signals and gameplay data [99] and the SAL [136] 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 7 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. [111] compare different types of stress annotation with the aim of finding the best possible approximation to the underlying ground truth. In particular they compare, in a first-order manner, 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.

Similarly to Lopes et al. [114], Camilleri et al. [120] used the *RankTrace* tool to annotate arousal across three very different games (e.g., horror, puzzle, and shooter) and then attempted to build *general* affect models across all three games. The model was trained via preference learning on in-game data and physiological data of the players to predict the level of arousal in unseen games. To test for generality the performance of the model was assessed through a form of *leave-one-game-out* cross-validation. By treating the arousal traces in a relative fashion and ranking the average gradient value of the trace across subsequent time windows Camilleri et al. managed to achieve a cross-game validation accuracy which was significantly higher than the baseline. When, instead, they treated the arousal trace in an absolute fashion and ranked the average values of subsequent windows, the arousal model could not surpass the baseline performance. With this empirical evidence at hand it is safe to conclude that the second-order processing of annotation traces as ordinal data not only yields more valid but also more general models of affect. As covered earlier in Section 4.3 similar results were obtained across datasets for general arousal prediction [118] and for general AU intensity estimation [105] using a local ranking approach.

4.8 A Summary

In this section we summarize the evidence solicited from the affective computing and psychometrics literature demonstrating the comparative advantages of the ordinal approach with regards to both reliability and validity. The summary of all findings is presented in a table format in which we place the various case studies covered in this paper with regards to the domain (or modality), the ordinal approach followed for collecting and processing the data (first-order vs. second-order), and the performance measure (reliability vs. validity) upon which they show that the ordinal approach is beneficial (see Table 2). It is directly observable from the table that the ordinal approach is rather robust across domains, performance measures and methodologies followed. While the first-order comparative studies are currently fewer than the studies following a second-order approach, the message of this paper remains solid: regardless of the approach adopted (first or second order), treating and processing data as ordinal is beneficial for the reliability and/or validity of the affect models. The table also suggests there is still room for further investigation and opportunities for future research in ordinal labeling and processing. The blank cells of the table call for further studies and analysis.

TABLE 2
Summary of comparative advantages of the ordinal approach (first-order and second-order) with regards to the performances measures of reliability and validity and across the affective computing domains presented in this paper.

Domain	Measure	First-Order	Second-Order
Video	Reliability	[57]	[57]
	Validity		[114], [118]
Face	Reliability		
	Validity		[90], [105], [121], [122]
Body	Reliability	[126]	[123]
	Validity		[123], [124]
Speech	Reliability		[86]
	Validity		[60], [71], [72], [73], [93]
Music/Sounds	Reliability	[70], [110]	
	Validity	[70], [110]	[118]
Games	Reliability	[78], [111]	
	Validity		[60], [120]

5 Outstanding Questions

Even quite recently, it seemed natural to structure an overview of labeling round two contrasting options—the traditional categorical descriptions, and tracing (assumed to be interval) [7]. From a review of the relevant literatures, it seems clear that ordinal approaches are at the very least a strong alternative, and there is a strong case for saying that they should be the default option. There is research from a range of disciplines to confirm that that position is intellectually well-founded, and from within affective computing to confirm that it is technically realistic. However, a few broad kinds of question clearly remain. There are questions about the limitations of the ordinal approach (Section 5.1); about the future challenges (Section 5.2); and finally about the intellectual status of the approach—whether it is simply a matter of solving technical problems, or bears on questions of general intellectual interest (Section 5.3).

5.1 Objections

The **first broad question** is whether obvious objections can be answered. A few key cases are worth considering.

More Must Be Better: It is natural to assume that the more information labelers' responses contain, the better they will specify the relevant impression. However, the exact opposite is what emerges from all the material presented in this paper—studies in affective computing, evidence from other disciplines, and psychological theorizing about the nature of emotions: it appears that less is more. Trying to achieve absolute annotation is adding noise rather than valid information. A specific concern is that relative descriptions cannot capture the intensity of an emotion, only its relation to a comparator. However, the paper has presented evidence that given appropriately chosen comparators, the intensity is not lost. Functions that model affective ranks (e.g., a preference learned neural network [60], [94]) can directly output intensity values. That links to the next objection.

Anchors: Ranking procedures require at least one reference point. The need to provide a reference point adds practical work; and theoretically, results can only be relative to it. We argue that far from being a problem, the introduction of explicit reference points is a major strength. The theoretical background discussed earlier indicates that whether or not they are explicit, baselines play an integral part in any type of reporting scheme. It is a major attraction of comparative procedures that this reference to baselines is not left to unconscious and uncontrolled adjustments. Instead, the baseline is a real option that is used as a reference during the annotation. Once again, what appears to be a limitation actually encapsulates a core strength.

Ipsative Nature: Comparative measures in some contexts are described as *ipsative*—that is, the scale is peculiar to the individual. So, for instance, the worst one person has experienced may be far from the worst that another has; and what is worse for one may not be worse for another. Issues like that are said to make comparison across individuals impossible. Again, what matters is that ordinal methods help to bring real issues into the open, rather than give a misleadingly reassuring impression. Once in the open, they can be addressed. The first of the issues raised is dealt with in the way that was just mentioned, by providing anchors. The second is addressed by measuring reliability. Testing reliability indicates that the comparisons are less ipsative than the hidden transformations that labelers use to map their internal experiences onto an apparently objective scale.

Exponential Growth: An obvious limitation of methods based on pairwise comparisons is that the number of possible comparisons increases as the square of the number of stimuli. This, in turn, limits the usability of the method for large sets of stimuli. The issue is a real one, but the literature has proposed several protocols and algorithmic approaches that reduce the number of comparisons required for a result to be valid. Indicatively in Li et al. [35] the full paired comparison method was compared against pair comparison selection algorithms in the task of assessing visual discomfort. The adaptive square design method introduced in Li et al. [35] reduces the number of comparisons substantially while providing accurate and robust results against observation errors and interdependence of comparisons. See Section 3.3.3 for further discussion on this issue.

Cognitive Load and Completion Time: Ranking questions can be cognitively demanding. The amount of concentration required by the respondent is directly linked to the number of options to order. The time it takes for an annotator to answer a ranking question is also directly proportionate to the number of options that need to be ordered or ranked [137]. Again, though, there are efficient methods. Studies have shown that if respondents are presented with a short question, given a predefined set of possible answers, and are asked to rank their top n (n is usually 3 or 5) answers, the validity of ranking responses reaches the highest possible value [138].

Statistical Analysis: Using ordinal data restricts the statistical methods that can be used. Standard descriptive statistics such as mean values and standard deviations are not applicable. Parametric tests are not applicable either. Part of the answer is that multiple data visualization methods and data processing techniques are available for handling preferences and ranks, from classical correlation analysis, to statistical tests for significance, and further to modern machine learning approaches. Several of these methods have been covered in the paper (Section 3.3). The other part is that yet again, this is about exposing an issue that ought to be faced rather than covered up. If familiar analytic techniques do not fit emotion, then it is misguided to insist on descriptive techniques that make it look as if they do.

5.2 Challenges

Certainly objections can be put in different ways, but these seem to cover most of the issues that recur. Outright objections should be separated from the **second broad question**, which is whether there are outstanding challenges that require work. A few can be pointed out reasonably easily.

Segmentation: Comparative techniques highlight the problem of finding appropriate units to compare. Most obviously, simply ranking extracts will not reveal whether some show high internal variation; and if there is significant variation within an extract, there will be parts that its rank does not describe at all well. Using short extracts can minimize that problem, but it deprives labelers of context that may be crucial. The hope that tracing techniques might avoid those issues was one of their attractions [7]; but in the event, emotional variation is precisely where they are weakest (see Section 2). There are various approaches to finding appropriate units. In the speech modality, there is long-standing interest in prosodically defined units [139]. In the visual modality, more recent work has used expressive patterns to identify informative regions [140]. It underlines the complexity of the issue: for instance, different parts of the face show different timecourses [141]. The immediate bearing of those issues is on temporal resolution. It is an area where there is a great deal to be learned, but again, that is a general problem: ordinal techniques highlight the fact that it is genuinely challenging.

It is natural to believe that the higher reliability of ordinal approaches is due to their lower resolution; at least compared to higher resolution interval values. When it comes to temporal segmentation of continuous annotation, for instance, time intervals should be adapted if ranks are to be compared against ratings [142]. While the resolution of discrete ranks versus continuous absolute values can be a valid concern there exist accurate and appropriate methods for comparing the two via standard windowing methods. An empirical approach is to attempt different time resolutions and compare the reliability obtained with ranks (first-order), with ordinal labels that are obtained through intervals (second-order) and with interval values (e.g., traces) directly. Statistical methods such as Krippendorff's α [85] can be used to compare the reliability across different data types as, for instance, performed by Yannakakis and Martinez [57]. That comparative study showcased the inter-rater reliability gains of both first-order and second-order processing of continuous video annotations over standard continuous annotation (ratings). Our message in this paper is that whether an annotation tool allows for first-order ordinal labels (e.g., *AffectRank*) or not (e.g., *FeelTrace*), following the ordinal path of data processing can be beneficial for the reliability of the labels obtained. It is also important to note that our thesis stands for any type of annotation (discrete or continuous) and for any type of domain as showcased through the several case studies covered in Section 4.

Efficient Algorithms: At first sight, efficiency appears to be a severe problem for comparative approaches. It is already clear that the worst fears are avoidable (see Section 3.3). However, psychophysics, dealing with related problems, makes considerable efficiency gains by using methods that take advantage of "the experimenter's prior knowledge and the observer's responses on past trials" to determine what is presented next [143]. It remains to be seen whether similar gains can be made here. Another layer of questions about efficiency comes into play when we consider the issue of segmentation (if the number of segments can be reduced, it should be). Overall, identifying optimal methods remains an interesting challenge.

Validation: It is a feature of ordinal data that a great variety of specific techniques, first- and second-order, can be used to obtain it (see Section 3.2). There is an obvious case for work that establishes how they compare. In particular, there is a need to study comparative validity (see Section 4.1): that is, to identify the techniques most likely to yield results that allow systems trained on them to function as they should.

5.3 Intellectual Views

The **third broad question** remains. Most of the focus in this paper has been on technical means to satisfy a practical end—providing training material. The question is whether there is any deeper intellectual reason to be interested in the area.

Learning the Language of Emotion: The general answer is that research in the area is linked to a question that is deep, and long-standing; and that affective computing may have a particular role in answering. The question is where ordinality enters the picture of emotion.

One possibility is that ordinality is relatively superficial. It is a product of the way people go about externalizing underlying experiences which, in and of themselves, are marked by straightforward numerical parameters (such as intensity, valence, goal conduciveness, etc). If that is the case, then effects involving ordinality are mostly of technical interest: they tell us about reporting emotion rather than emotion itself. Some of the evidence that was mentioned invites that interpretation [29].

The other possibility is that ordinality goes to the *core* of emotion. Salient aspects of an experience (such as valence, intensity, etc.) are defined by the place that an underlying representation of the situation occupies relative to a range of comparators that are active at the time. As Section 2 indicated, accounts of that kind are not eccentric: related ideas have been proposed in many areas. If that is the case, then it is important for understanding human beings; and affective computing has a particular role in developing the idea. The reason is that it is hard to imagine how the idea could be consolidated without sophisticated modeling. Affective computing has the means to carry out that kind of modeling, and the motive: it is basic to reconstructing emotional experiences from what people say about them.

Deciding between those possibilities is not a technicality. It is not far-fetched to say that it is about unraveling the underlying language of emotion.

6 Conclusions

This paper has presented and supported the thesis that emotions are by nature *ordinal*. 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, behavioral economics, marketing research, artificial intelligence and, not least, affective computing. Our fundamental aim is to make it clear that this is not a fringe issue. Attempts to work with interval or categorical 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.

Acknowledgments

This paper was supported in part by the FP7 Marie Curie CIG project AutoGameDesign (project no: 630665) and the H2020 project Envisage (project no: 731900). This work was also funded by National Science Foundation under grant IIS-1453781 (CAREER Award).

References

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

[2] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: An overview," *International Journal of Synthetic Emotions*, vol. 3, no. 1, pp. 1–17, January-June 2012.

[3] M. Pantic and J. Rothkrantz, "Automatic analysis of facial expressions: State of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, December 2000.

[4] F. Dellaert and T. P. A. Waibel, "Recognizing emotion in speech," in *International Conference on Spoken Language (ICSLP 1996)*, vol. 3, Philadelphia, PA, USA, October 1996, pp. 1970–1973.

[5] K. R. Scherer and G. Ceschi, "Lost Luggage: A Field Study of Emotion? Antecedent Appraisal," *Motivation and Emotion*, vol. 21, no. 3, pp. 211–235, September 1997.

[6] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Communication*, vol. 40, no. 1-2, pp. 117–143, April 2003.

[7] 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.

[8] 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.

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

[10] S. Savvidou, "Validation of the feeltrace tool for recording impressions of expressed emotion," Ph.D. dissertation, Queen's University Belfast, 2011.

[11] L. Devillers, R. Cowie, J. C. 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 *International conference on Language Resources and Evaluation (LREC 2006)*, Genoa,Italy, May 2006, pp. 1105–1110.

[12] 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.

[13] J. Deigh, "Concepts of emotions in modern philosophy and psychology," in *The Oxford Handbook of Philosophy of Emotion*, P. Goldie, Ed. Oxford University Press, May 2012, pp. 17–40.

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

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

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

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

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

[19] S. Stevens, "To honor Fechner and repeal his law," *Science*, vol. 133, pp. 80–86, January 1961.

[20] N. J. MacKinnon, "Measuring self-sentiments," in *Self-Esteem and Beyond*. Springer, 2015, pp. 71–95.

[21] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[22] M. Junge and R. Reisenzein, "Indirect scaling methods for testing quantitative emotion theories," *Cognition and Emotion*, vol. 27, no. 7, pp. 1247–1275, 2013.

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

[24] 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.

[25] D. Laming, "The relativity of 'absolute'judgements," *British Journal of Mathematical and Statistical Psychology*, vol. 37, no. 2, pp. 152–183, 1984.

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

[27] 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.

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

[29] M. Junge and R. Reisenzein, "Metric scales for emotion measurement," *Psychological Test and Assessment Modeling*, vol. 58, no. 3, pp. 497–530, 2016.

[30] R. B. Zajonc, "Attitudinal effects of mere exposure," *Journal of personality and social psychology*, vol. 9, 1968.

[31] D. F. Alwin and J. A. Krosnick, "The measurement of values in surveys: A comparison of ratings and rankings," *Public Opinion Quarterly*, vol. 49, no. 4, pp. 535–552, 1985.

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

[33] M. L. Kohn, "Reassessment 1977," *Class and Conformity: A Study in Values, 2nd ed., pp. xxv-lx, Univ. of Chicago Press, Chicago. CHANGES IN QUALITIES*, vol. 235, 1977.

[34] 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.

[35] J. Li, M. Barkowsky, and P. Le Callet, "Boosting paired comparison methodology in measuring visual discomfort of 3dtv: performances of three different designs," in *Stereoscopic Displays and Applications XXIV*, vol. 8648. International Society for Optics and Photonics, 2013, p. 86481V.

[36] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.

[37] J.-S. Lee, L. Goldmann, and T. Ebrahimi, "Paired comparison-based subjective quality assessment of stereoscopic images," *Multimedia tools and applications*, vol. 67, no. 1, pp. 31–48, 2013.

[38] A. Tversky, "Features of similarity," *Psychological review*, vol. 84, no. 4, p. 327, 1977.

[39] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior (commemorative edition)*. Princeton university press, 2007.

[40] D. W. Stephens and J. R. Krebs, *Foraging theory*. Princeton University Press, 1986.

[41] A. Tversky and I. Simonson, "Context-dependent preferences," *Management science*, vol. 39, no. 10, pp. 1179–1189, 1993.

[42] S. Shafir, T. A. Waite, and B. H. Smith, "Context-dependent violations of rational choice in honeybees (apis mellifera) and gray jays (perisoreus canadensis)," *Behavioral Ecology and Sociobiology*, vol. 51, no. 2, pp. 180–187, 2002.

[43] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[44] H. A. Simon, "Theories of bounded rationality," *Decision and organization*, vol. 1, no. 1, pp. 161–176, 1972.

[45] D. Kahneman and A. Tversky, "Subjective probability: A judgment of representativeness," *Cognitive psychology*, vol. 3, no. 3, pp. 430–454, 1972.

[46] A. Tversky and D. Kahneman, "The framing of decisions and the psychology of choice," *Science*, vol. 211, no. 4481, pp. 453–458, 1981.

[47] D. Kahneman, "A perspective on judgment and choice: mapping bounded rationality." *American psychologist*, vol. 58, no. 9, p. 697, 2003.

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

[49] C. Padoa-Schioppa and J. A. Assad, "Neurons in the orbitofrontal cortex encode economic value," *Nature*, vol. 441, no. 7090, p. 223, 2006.

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

[51] K. Louie, L. E. Grattan, and P. W. Glimcher, "Reward value-based gain control: divisive normalization in parietal cortex," *Journal of Neuroscience*, vol. 31, no. 29, pp. 10 627–10 639, 2011.

[52] R. Elliott, Z. Agnew, and J. Deakin, "Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans," *European J. of Neuroscience*, pp. 2213–2218, 2008.

[53] K. Louie, M. W. Khaw, and P. W. Glimcher, "Normalization is a general neural mechanism for context-dependent decision making," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 6139–6144, 2013.

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

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

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

[57] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *ACII*, 2015, pp. 574–580.

[58] T. Qin, T.-Y. Liu, J. Xu, and H. Li, "Letor: A benchmark collection for research on learning to rank for information retrieval," *Information Retrieval*, vol. 13, no. 4, pp. 346–374, 2010.

[59] K. Karpouzis, G. N. Yannakakis, N. Shaker, and S. Asteriadis, "The platformer experience dataset," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 712–718.

[60] 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.

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

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

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

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

[65] 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.

[66] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias, "Nonverbal feedback on user interest based on gaze direction and head pose," in *Proceedings of SMAP*, 2007, pp. 171–178.

[67] S. Mariooryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier," *IEEE Trans. on Affective Computing*, pp. 119–130, 2017.

[68] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Automatic Face and Gesture Recognition*. IEEE, 2013, pp. 1–8.

[69] G. N. Yannakakis, "Preference learning for affective modeling," in *Proceedings of ACII*. IEEE, 2009, pp. 1–6.

[70] 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.

[71] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *IEEE ICASSP*, 2016, pp. 5205–5209.

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

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

[74] 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*. Oxford University Press, 2013, pp. 110–127.

[75] 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, 2015.

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

[77] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proceedings of the 2nd ACM workshop on Multimedia semantics*. ACM, 2008, pp. 32–39.

[78] G. N. Yannakakis and J. Hallam, "Rating vs. preference: a comparative study of self-reporting," in *ACII*, 2011, pp. 437–446.

[79] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in health sciences education*, vol. 15, no. 5, pp. 625–632, 2010.

[80] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[81] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[82] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[83] 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.

[84] 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.

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

[86] I. Siegert, R. Böck, and A. Wendemuth, "Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 17–28, 2014.

[87] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *IEEE ICIP*, 2008, pp. 117–120.

[88] 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.

[89] D. Tran, R. Walecki, O. Rudovic, S. Eleftheriadis, B. Schuller, and M. Pantic, "DeepCoder: Semi-parametric variational autoencoders for automatic facial action coding," in *IEEE International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, October 2017, pp. 3190–3199.

[90] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "A copula ordinal regression framework for joint estimation of facial action unit intensity," *IEEE Transactions on Affective Computing*, 2017.

[91] J. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *IEEE ICME*, 2009, pp. 1436–1439.

[92] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions," in *ISMIR*, 2008, pp. 325–330.

[93] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *IEEE ICASSP*, 2017, pp. 4995–4999.

[94] 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*, 2014, pp. 34–41.

[95] S. Tognetti, M. Garbarino, A. Bonarini, and M. Matteucci, "Modeling enjoyment preference from physiological responses in a car racing game," in *IEEE CIG*. IEEE, 2010, pp. 321–328.

[96] 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.

[97] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *ICANN*, 1999, pp. 97–102.

[98] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *ICML*, 2005, pp. 137–144.

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

[100] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *ICML*, 2005, pp. 89–96.

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

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

[103] K. Jamieson and R. Nowak, "Active ranking using pairwise comparisons," in *Advances in Neural Information Processing Systems (NIPS 2011)*, vol. 24, Granada, Spain, December 2011, pp. 2240–2248.

[104] F. Wauthier, M. Jordan, and N. Jojic, "Efficient ranking from pairwise comparisons," in *International Conference on Machine Learning (ICML 2013*, vol. 28, Atlanta, GA, USA, June 2013, pp. 109–117.

[105] T. Baltrušaitis, L. Li, and L. P. Morency, "Local-global ranking for facial expression intensity estimation," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 111–118.

[106] 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.

[107] 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.

[108] 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," *J. of Clinical Oncology*, pp. 5748–5752, 2007.

[109] 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.

[110] Y.-H. Yang and H. H. Chen, "Music emotion ranking," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 1657–1660.

[111] C. Holmgård, G. N. Yannakakis, H. P. Martinez, and K.-I. Karstoft, "To rank or to classify? Annotating stress for reliable PTSD profiling," in *ACII*, 2015, pp. 719–725.

[112] D. Küster and A. Kappas, "Measuring emotions online: Expression and physiology," in *Cyberemotions*. Springer, 2017, pp. 71–93.

[113] 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.

[114] P. Lopes, G. N. Yannakakis, and A. Liapis, "Ranktrace: Relative and unbounded affect annotation," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2017.

[115] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with emotionml," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 709–710.

[116] A. Clerico, C. Chamberland, M. Parent, P.-E. Michon, S. Tremblay, T. H. Falk, J.-C. Gagnon, and P. Jackson, "Biometrics and classifier fusion to predict the fun-factor in video gaming," in *Computational Intelligence and Games (CIG), 2016 IEEE Conference on*. IEEE, 2016, pp. 1–8.

[117] P. Lopes, A. Liapis, and G. N. Yannakakis, "Targeting horror via level and soundscape generation," in *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*, 2015.

[118] D. Melhart, K. Sfikas, G. Giannakakis, G. N. Yannakakis, and A. Liapis, "A study on affect model validity: Nominal vs ordinal labels," in *Proceedings of the IJCAI workshop on AI and Affective Computing*, 2018.

[119] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *arXiv preprint arXiv:1702.02510*, 2017.

[120] E. Camilleri, G. N. Yannakakis, and A. Liapis, "Towards general models of player affect," in *Affective Computing and Intelligent Interaction (ACII), 2017 International Conference on*, 2017.

[121] M. Khademi and L.-P. Morency, "Relative facial action unit detection," in *IEEE Winter Conference on Applications of Computer Vision (WACV 2014)*, Steamboat Springs, CO, USA, March 2014, pp. 1090–1095.

[122] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic, "Variational gaussian process auto-encoder for ordinal prediction of facial action units," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 154–170.

[123] J. Kim, K. P. Truong, V. Charisi, C. Zaga, V. Evers, and M. Chetouani, "Multimodal detection of engagement in groups of children using rank learning," in *International Workshop on Human Behavior Understanding*. Springer, 2016, pp. 35–48.

[124] J. Kim, K. P. Truong, and V. Evers, "Automatic temporal ranking of children's engagement levels using multi-modal cues," *Computer Speech & Language*, vol. 50, pp. 16–39, 2018.

[125] M. Pasch, A. Kleinsmith, and M. Landoni, "Human rating of emotional expressions-scales vs. preferences." in *PhyCS*, 2014, pp. 240–245.

[126] R. Rienks and D. Heylen, "Automatic dominance detection in meetings using easily detectable features," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*. IEEE, 2005.

[127] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Interspeech 2018*, Hyderabad, India, September 2018.

[128] R. Böck, I. Siegert, M. Haase, J. Lange, and A. Wendemuth, "ikannotate–a tool for labelling, transcription, and annotation of emotionally coloured speech," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 25–34.

[129] I. Siegert, R. Böck, and A. Wendemuth, "The influence of context knowledge for multi-modal affective annotation," in *International Conference on Human-Computer Interaction*. Springer, 2013, pp. 381–390.

[130] Z. Huang, J. Epps, and E. Ambikairajah, "An investigation of emotion change detection from speech," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1329–1333.

[131] Z. Huang and J. Epps, "Detecting the instant of emotion change from speech using a martingale framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5195–5199.

[132] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.

[133] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 129–136.

[134] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[135] M. Garbarino, S. Tognetti, M. Matteucci, and A. Bonarini, "Learning general preference models from physiological responses in video games: How complex is it?" *Affective Computing and Intelligent Interaction*, pp. 517–526, 2011.

[136] K. Karpouzis, G. Caridakis, R. Cowie, and E. Douglas-Cowie, "Induction, recording and recognition of natural emotions from facial expressions and speech prosody," *JMUI*, pp. 195–206, 2013.

[137] J. M. Munson and S. H. McIntyre, "Developing practical procedures for the measurement of personal values in cross-cultural marketing," *Journal of Marketing Research*, pp. 48–52, 1979.

[138] K. Peng, R. E. Nisbett, and N. Y. Wong, "Validity problems comparing values across cultures and possible solutions." *Psychological methods*, vol. 2, no. 4, p. 329, 1997.

[139] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M.Westerdijk, and S. Stroeve, "Approaching automatic recognition of emotion from voice: A rough benchmark," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 207–212.

[140] Y. Kim, J. Chen, M.-C. Chang, X. Wang, E. M. Provost, and S. Lyu, "Modeling transition patterns between events for tempo-

ral human action segmentation and classification," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE international conference and Workshops on*, vol. 1. IEEE, 2015, pp. 1–8.

[141] Y. Kim and E. M. Provost, "Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction, pages=92–99, year=2016, organization=ACM*.

[142] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott, "A psychometric evaluation of the facial action coding system for assessing spontaneous expression," *Journal of Nonverbal Behavior*, vol. 25, no. 3, pp. 167–185, 2001.

[143] B. Farell and D. G. Pelli, "Psychophysical methods, or how to measure a threshold and why," *Vision research: A practical guide to laboratory methods*, vol. 5, pp. 129–136, 1999.

Georgios N. Yannakakis (S'04-M'05-SM'14) is a Professor and Director of the Institute of Digital Games, University of Malta. He received the Ph.D. degree in Informatics from the University of Edinburgh in 2006. Prior to joining the Institute of Digital Games, UoM, in 2012 he was an Associate Professor at the Center for Computer Games Research at the IT University of Copenhagen. He does research at the crossroads of artificial intelligence, computational creativity, affective computing, advanced game technology, and human-computer interaction. He pursues research concepts such as user experience modeling and procedural content generation for the design of personalized interactive systems for entertainment, education, training and health. He has published more than 220 papers in the aforementioned fields and his work has been cited broadly. His research has been supported by numerous national and European grants (including a Marie Skłodowska-Curie Fellowship) and has appeared in *Science Magazine* and *New Scientist* among other venues. He is currently an Associate Editor of the IEEE TRANSACTIONS ON GAMES and used to be Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES journals. He has been the General Chair of key conferences in the area of game artificial intelligence (IEEE CIG 2010) and games research (FDG 2013). Among the several rewards he has received for journal and conference publications he is the recipient of the *IEEE Transactions on Affective Computing Most Influential Paper Award* and the *ACII 2017 Best Paper Award*. He is a senior member of the IEEE.

Roddy Cowie Roddy Cowie is Emeritus professor of Psychology at Queens University, Belfast. He has used computational methods to study a range of complex perceptual phenomena — perceiving pictures, the experience of deafness, what speech conveys about the speaker, and, in a series of EC projects, the perception of emotion, where he has developed methods of measuring perceived emotion and inducing emotionally colored interactions. Key outputs include pioneering papers on emotion recognition in human-computer interaction (2001) and the emotional states that are expressed by speech (2003), as well as special editions of Speech Communication (2003) and Neural Networks (2005), and the HUMAINE Handbook on Emotion-Oriented Systems (2011).

Carlos Busso (S'02-M'09-SM'13) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests include digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human interaction, in-vehicle active safety system, and machine learning methods for multimodal processing. He was the general chair of ACII 2017. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.