Annotation and Processing of Continuous Emotional Attributes: Challenges and Opportunities

Angeliki Metallinou
Signal Analysis and Interpetation Lab (SAIL)
University of Southern California
California, USA
metallin@usc.edu

Shrikanth Narayanan
Signal Analysis and Interpetation Lab (SAIL)
University of Southern California
California, USA
shri@sipi.usc.edu

***Abstract*— Human emotional and cognitive states evolve with variable intensity and clarity through the course of social interactions and experiences, and they are continuously influenced by a variety of input multimodal information from the environment and the interaction participants. This has motivated the development of a new area within affective computing that treats emotions as continuous variables and examines their representation, annotation and modeling. In this work, we use as a starting point the continuous emotional annotation that we performed on a large, multimodal database, and discuss annotation challenges, design decisions, annotation results and lessons learned from this effort, in the context of existing literature. Additionally, we discuss a variety of open questions for future research in terms of labeling, combining and processing continuous assessments of emotional and cognitive states.**

***Index Terms*— emotional representations, continuous emotional annotation, continuous emotion estimation, dyadic multimodal database**

I. INTRODUCTION

In everyday social situations, humans are able to, in real-time, perceive, combine, process and respond to a multitude of information including lexical content of a conversation, nonverbal information such as facial and body gestures, subtle vocal cues, and context, i.e., events happening in the environment. Multimodal cues unfold, sometimes asynchronously, through time and continuously influence the participants' underlying affective and cognitive states, which are volatile, evolving and often ambiguous. Those key points, namely the multimodal nature of emotional expressions that makes it difficult to define precise starting and ending points of an expression, and the realization of the complex nature of emotion that is not well-captured by static, categorical labels motivate the use of continuous representations. Those are descriptions of emotional states, e.g., activation or valence, or cognitive states, e.g., engagement, that take continuous values and continuously evolve through time. The human annotation and automatic processing of such continuous attributes is an exciting, emerging topic within the affective computing domain.

The goal of this paper is to discuss the challenges and opportunities that lie in this area, focusing primarily on data labeling issues. We use as a case study the continuous emotional annotation of a multimodal emotional database of dyadic interactions, in order to discuss the various challenges that we faced, and the design decisions that we adopted in order to address them. Those challenges include modifying the annotation software to suit the needs of our experiment, training the participating annotators to have a good understanding of the data, the emotional attributes of interest and the annotation software, and defining what inter-annotator agreement would mean for the case of continuous ratings.

Our analysis of the resulting annotations shows that continuous rating of long clips containing nonprototypical emotional expressions is a challenging task, and humans tend to be better at at rating attributes in relative, rather than in absolute, terms. Another interesting question is the relation between detailed continuous ratings and a global (summative) rating, given by the same annotator. Our analysis indicates that global ratings are not simple averages of a continuous assessment, as different portions of a data recording are weighted differently, an effect that depends both on the rater and the annotated attribute.

Using this annotation work as a starting point, we further discuss a series of open questions that have been partially addressed in the literature and represent interesting future directions. For example, since continuous annotations are performed real-time, there are subject-specific delays between an emotional event and its annotation, which brings forward the issue of measuring, modeling and normalizing for such delays. Aggregation of multiple annotators' continuous subjective judgements, in a way that considers individual annotator variability, is another interesting direction. Finally, continuous representations could shed light into the way humans perceive and aggregate information over time, highlighting regions of interest or 'emotional saliency' that are weighted more when assessing an emotional experience.

II. RELATED WORK

Psychology researchers have proposed describing human emotional states in terms of certain continuous-valued attributes (dimensions); activation, valence and dominance being the most common [1], [2]. Activation describes how intense is the emotional experience, valence describes the level of pleasure related to an emotion, and takes positive and negative values for pleasant and unpleasant emotions respectively, while dominance describes the level of control of a person during an emotional experience. This approach can be seen as a more generic way to classify emotions, and it has been widely adopted in the affective computing

community. However, the dimensional attribute values are typically quantized into discrete levels [3], [4], [5]. Examples of work that avoid discretizing the emotional dimensions include [6], [7] where regression approaches, such as Support Vector Regression (SVR), were used to estimate continuous dimensional attributes from speech cues of presegmented utterances.

A small but increasing amount of works treats both time and emotion variables as continuous. Continuous emotional annotation across time has been facilitated by the introduction of tools such as Feeltrace, a freely available software that allows real-time emotional annotation of video [8]. Other continuous annotation software include EmuJoy [9], designed primarily for annotation of music emotional content, and Continuous Measurement System (CMS) [10], which was used in psychology studies for annotating continuous attributes such as smile intensity, level of positive emotion, from videos of infants. Recently, an updated, more flexible version of Feeltrace, called Gtrace was introduced [11].

The availability of these software tools has enabled continuous emotional annotation of a number of speech and multimodal emotional databases, including the Belfast Naturalistic Database [12], SEMAINE [13] and the CreativeIT database [14], that is discussed here. The availability of continuous annotations may allow a more generic and flexible treatment of emotions. However, certain works, including our own past work [15], have reported difficulty in obtaining consistent annotations of continuous attributes because of the subjectivity and the challenging nature of the task [16], [17]. This challenge is one of the main points of discussion for this paper, and it is further elaborated in Sec.V-A.

Modeling and estimating continuous ratings require methodologies that move away from multiclass classification schemes and instead focus on continuous estimation of such ratings. In [18] the authors describe a multimodal system to continuously track valence and activation of a speaker, using SVR and Long-Short Term memory (LSTM) regression, with LSTM being the best performing approach (LSTM is a variation of Recurrent Neural Networks). Similarly, single-modality systems were proposed in [19], [20] using SVR and LSTM neural networks for regression to continuously estimate valence and activation values from emotional speech. In our previous work, we have utilized a generative GMM-based method for continuous emotion estimation based on multimodal information, that is shown to outperform LSTM regression [21]. An unsupervised method for mapping the emotional content of movies in the valence-activation space was proposed in [22], [23] using low-level audio and video cues.

III. DATABASE DESCRIPTION

We use the USC CreativeIT database which is a multimodal database of theatrical improvisation, collected as a collaborative effort between engineering and theater [14]. It contains a variety of dyadic theatrical performances, that are either improvisations loosely based on scenes from theatrical plays or theatrical exercises where actors repeat sentences in a manner that conveys specific intent (e.g., accepting or rejecting behavior towards other). The actors were not instructed to produce specific emotions; instead, a variety of improvised emotional expressions and interaction dynamics occur as part of the performance. This design makes the emotional manifestations of the database especially challenging to annotate and analyze, since they are more subtle and diverse. The design was performed by a theater expert (director/teacher), and the participating actors were senior theater students. The performances were recorded under the guidance of the theater expert in order to ensure high quality performances. The final result is very close to an actual theatrical performance. Further data collection details can be found in [14].

The collected data contains full body Motion Capture information from both participants, speech obtained from close-talking microphones and videos from two HD cameras located at opposite sides of the recording space. There are 16 participating actors (9 female) who perform a total of 50 improvisations, ranging from 2-10 minutes. In this work, we focus on the data annotation process in terms of emotional content using the videos of the performances.

IV. DATA ANNOTATION

A. Why Continuous Annotations?

The CreativeIT database contains a variety of multimodal expressions and interaction dynamics that continuously unfold during the improvisation. Therefore, it is difficult to define precise starting and ending times of expressions since those are produced multimodally, or to segment interactions into units of homogenous emotional content. In unimodal databases, or databases that are spoken-dialog centric such as IEMOCAP [24] and VAM [25], it seems natural to segment a conversation into utterances as basic units for examining emotional content. In contrast, the CreativeIT database contains many nonverbal emotional expressions that happen asynchronously to speech or when the participant is silent. Such observations motivate the use of continuous attributes as a natural way to describe the emotional flow of an interaction.

The perceived emotional state for each participant was annotated in terms of the widely used dimensional attributes of activation, valence and dominance. This representation is well-suited to describe the complex and ambiguous manifestations of the CreativeIT database, which do not always have clear categorical descriptions. For our annotations, we used the Feeltrace software [8] and collected annotations of perceived activation, valence and dominance for each actor in each performance, taking continuous values in [-1,1].

B. Challenges and Design Decisions

Annotation of emotional content is an inherently subjective task that depends, among others, on the individual's perception, experiences and cultural background. The use of continuous descriptors seems to increase the level of complexity of the emotional annotation task, as it requires a higher amount

of attention and cognitive processing compared to non real-time, discrete annotation tasks. Apart from being a strenuous and time-consuming process, continuous annotation poses challenges in terms of obtaining inter-annotator agreement, as has been reported by several researchers. In [16] authors report that in 70% of continuous valence annotations of TV clips, the inter-annotator correlations are above the 0.5 threshold, a percentage that reduces to 55% and 34% for activation and dominance (power), respectively. In [17] authors report mean annotator correlations of 0.3 and 0.4 for valence and activation respectively for continuous self-annotations of felt emotion while watching movies. Our pilot annotation of a CreativeIT data subset resulted in median annotator correlations of around 0.5 for the three dimensional attributes [15].

The continuous rating of the SEMAINE database shows a more optimistic picture, where Cronbach's α values of continuous valence ratings are reported to be higher than 0.75 for 86% of the ratings (indicating acceptable consistency) [13]. Authors also report high α values (0.8-0.9) for consistency of functions over continuous dimensional attributes, e.g., mean. However such functions should be interpreted with caution as they do not necessarily correspond to a well-defined user perception. For example, the mean of a continuous rating over a clip does not generally correspond to the rater's perceived global rating of the clip, as discussed in Sec. V-C.

For the annotation of the CreativeIT data, we identified a number of potential factors that could be sources of annotation noise, and could increase the level of inter-annotator variability over what is naturally expected because of the challenging and subjective nature of the task. Below, we describe such noise factors and the resulting practical design decisions that we adopted in order to address them.

Annotator Motivation and Experience We recruited psychology students, most of whom had previous experience in emotional annotation, and were committed to weekly working requirements.

Definition of Emotional Attributes Although people typically learn to assess emotional content through social experiences, the definition of dimensional emotional attributes may be less intuitive for some annotators. The definitions of activation, valence and dominance attributes were explained through examples. We clarified that ratings are subjective, however annotators should be able to rationally explain their decisions based on verbal or nonverbal characteristics of the interaction [1].

The annotation instrument We observed a learning curve until annotators became comfortable with the use of Feeltrace (see also Sec.VI-A). Annotators were trained on how to use Feeltrace, they performed their first annotations multiple times to familiarize themselves with the software, and were later encouraged to perform each annotation as many times as needed until they were satisfied with the result.

[1]One could argue that such rationalization may not necessarily be desirable, since continuous emotional ratings may be affected by implicit emotional perceptions in additon to rational reflection. The relation between emotional perception, the rating as indicated by cursor/mouse movement and the corresponding verbal explanation is yet unclear and could be a potential direction of study.

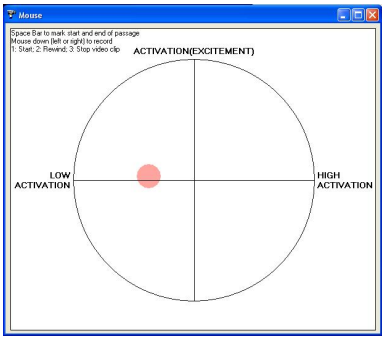


Fig. 1. Screenshot of the modified Feeltrace interface.

Understanding the type and range of emotional content in the dataset In order to facilitate the annotation process, we wanted annotators to be familiar with the type and range of emotional manifestations that appear in the database. Therefore, as part of their training, they had to watch in advance about a fifth of the recorded performances, randomly selected across different performances.

Person-specific annotation delays Since continuous annotations are performed in real time, we expect person-specific delays due to perceptual processing, between the time that an event happens and when its emotional content is annotated. In order to reduce such delays, we modified the Feeltrace interface so that annotators can focus their attention on one attribute each time, rather than two attributes that was the original design of Feeltrace. The modified Feeltrace interface for activation annotation is presented in Fig.1. The annotation is perfomed by moving the mouse, shown as a full circle, along the horizontal line, while watching the performance video in a separate window. It is interesting to note that a one-dimensional version of the Feeltrace interface later became available (software Gtrace [11]), indicating the need for such a one-dimensional annotation tool. Finally, to further reduce delays due to perceptual processing, we also instructed annotators to watch each video multiple times and have a clear idea of the emotional content before starting the real-time annotation.

C. Discrete Data annotation

We also collected discrete annotations of global emotional content of each performance. Emotional content was rated in terms of perceived activation, valence and dominance for each actor on a 9-point scale. Rating 1 denotes the lowest possible activation level, the most negative valence level, and the most submissive dominance level. Annotators were asked to give an overall rating that summarizes the particular attribute over the total recording. They were instructed to perform the overall rating right after completing the corresponding continuous annotation, such that they would have a recent impression of the annotated performance and their continuous assessment of its emotional content. The reason for collecting global annotations is two-fold; firstly we

wanted to enrich our annotation with more standard discrete labels for potential future use. Secondly, we want to study relations between global discrete and detailed continuous ratings provided by the same person, in order to shed light into the way humans summarize an emotional experience.

V. ANNOTATION RESULTS

The database contains 50 recordings, each rated for both actors in the dyad, therefore we have 100 actor-recordings. Seven annotators participated in total, rating overlapping portions of the database, so that each actor-recording would be rated by three or four people (88 out of the 100 actor-recordings were rated by 3 people). The resulting continuous annotations were pre-processed by lowpass filtering to remove high frequency noise artifacts. This section describes analysis of the annotation results that we obtained.

A. Annotator Agreement

Evaluator agreement is a straightforward concept when dealing with discrete labels; for example we can say that two annotators agree if the choose the same label. For continuous annotations this concept becomes less straightforward. Researchers processing continuous ratings generally assume agreement when two continuous ratings are correlated, e.g., using Pearson correlation as a metric, or when they are consistent in terms of the quantitative Cronbach's α coefficient (which is widely used in the cognitive sciences), or when they have small mean square difference with each other, in terms of their absolute values [18], [13].

To choose an agreement metric, it is important to understand how raters behave when rating continuous attributes. Fig. 2 shows an example of activation annotations by three annotators for the same actor-recording, and their average. Although annotators agree on the trends of the activation curve (mean correlation of 0.67), and recognize pronounced activation events, they do not agree on the actual activation values. Similar observations hold true for several of our obtained annotations. This suggests that people agree more when describing emotions in relative terms, e.g., whether there has been an increase or decrease, rather than in absolute terms. Rating emotions in absolute terms seems more challenging because of each person's internal scale when assessing an emotional experience (similar arguments are made in [26]). This motivates us to focus on the annotation trends, and to use correlation metrics, such as Pearson's correlation, and Cronbach's α to measure evaluator agreement.

To compute the emotional 'ground truth' for each recording (especially for facilitating subsequent computational modeling), we need to aggregate the multiple annotators' decisions. However, some ratings might appear inconsistent with the ratings of the majority of annotators. This issue is common in emotional labeling with categorical labels, where the emotional ground truth is often computed based on majority voting and minority labels are ignored, e.g. [24]. Here, we extend this notion on continuous ratings, using correlations as a basis for agreement. Specifically, we set a cut-off threshold for defining acceptable annotator agreement and for each actor-recording, we take the union of all annotator pairs with linear correlations greater than the threshold. Only this annotator subset is used to compute the ground truth for the corresponding actor-recording. If no annotators are selected then we assume that there is no agreement for that recording. Our threshold is empirically set to 0.45, which is similar to the correlation threshold used in [16] for defining agreement. This results in ground truth agreement in 80, 84 and 73 actor-recordings for the activation, valence and dominance class respectively, out of 100 in total. Interestingly, a comparable percentage of ground truth agreement (about 75%) was reported for annotation into categorical labels using this majority voting scheme, for the IEMOCAP database [24], an emotional database of improvised acting.

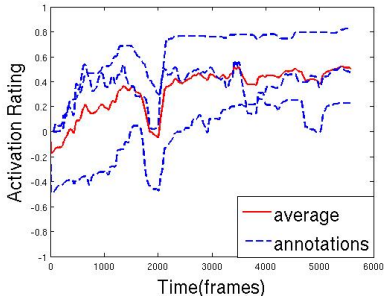


Fig. 2. Example of activation rating by three annotators.

This process selects consistent annotations and allows for aggregation methods such as averaging. Developing methodologies for combining multiple annotators' subjective judgements in a more informed way than averaging, potentially considering disagreeing annotators, is an important research problem, e.g., see [27], [28]. However, the continuous nature of our ratings makes such existing methodologies not directly applicable. An approach for weighted averaging of continuous annotations is proposed in [18], where annotator weights are computed based on their correlation with the rest of the annotators. This is a correlation-based, soft-selection scheme, which instead of ignoring uncorrelated annotators, it assigns them low weights. A more principled approach to uncover a continuous ground truth from multiple noisy continuous emotional annotations, in a way that also handles potential delays between the annotations, is proposed in [29], by combining Probabilistic Canonical Correlation Analysis (PCCA) and Dynamic Time Warping (DTW) approaches. Finally, generalized additive mixed models (GAMMs) have also been applied for combining continous emotional annotations [30].

To get an impression of annotator agreement over the database, we first compute the mean of the correlations between the selected annotators per actor-recording, and then compute the mean over all actor-recordings. Similarly, we also compute the Cronbach's α coefficient of the selected annotators per actor-recording, and then compute the overall mean. These measures are presented in Table I (third line).

TABLE I

MEASURES OF AGREEMENT OF THE SELECTED CONTINUOUS RATINGS FOR ACTIVATION, VALENCE AND DOMINANCE, AT DIFFERENT LEVELS OF ANNOTATION DETAIL

mean Pearson's correlation			mean Cronbach's α		
activation	valence	dominance	activation	valence	dominance
100 values per sec.					
0.60	0.63	0.59	0.75	0.77	0.74
1 value per sec.					
0.63	0.64	0.61	0.76	0.77	0.75

TABLE II

CRONBACH'S α AND ICC OF GLOBAL DISCRETE RATINGS FOR ACTIVATION, VALENCE AND DOMINANCE

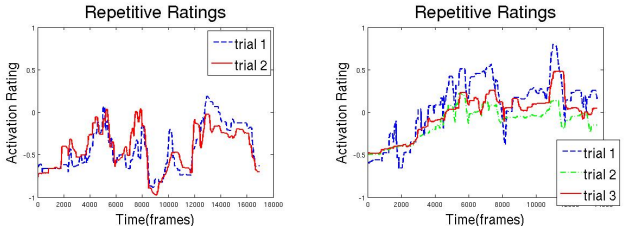
Cronbach's α			Intra-class correlation (Case 1 [31])		
activation	valence	dominance	activation	valence	dominance
0.72	0.78	0.67	0.69	0.78	0.65

For all tasks we have $\alpha > 0.7$ which indicates acceptable levels of annotator consistency, and good levels of correlation, given the challenging nature of the task. These measures were computed from detailed annotations with 100 values per sec. However, emotional states are slowly varying, therefore this degree of accuracy may not be necessary and it might be capturing annotation noise, unrelated to emotion. We also examine the effect of lower information detail by first averaging the selected annotations over windows of 3 sec., with 1 sec. overlap (1 value per sec.). These measures are presented in Table I (fifth line). We notice only a slight increase in annotator consistency, which indicates that our annotation pre-processing through lowpass filtering has removed most of the high frequency noise.

The consistency of the discrete global annotations was also examined by computing the α coefficient of global activation, valence and dominance ratings from all different annotators (no annotator selection is performed here). Those coefficients are presented in Table II. Overall, we notice that annotator consistency is at acceptable levels (around and over 0.7), except from dominance which is slightly lower. We also report consistency in terms of intra-class correlation (ICC), a related metric that considers the fact that different recordings are rated by different annotator subsets. Specifically, we use ICC Case 1 from [31], which assumes that each target is rated by a different set of annotators, randomly selected from a larger annotator pool.

B. Intra-Annotator Consistency: Repetitive ratings

As mentioned in Sec. IV-B, we encouraged annotators to perform repetitive ratings for their first recordings in order to familiarize themselves with Feeltrace. We also asked them to repeat ratings as many times as needed until they were satisfied with the result. The obtained repetitive ratings per annotator provide us with some information regarding how annotators may modify or refine their assessment of emotional content. Two examples of repetitive ratings are shown in Figs. 3(a)-(b), where in 3(a) we can observe slight changes in the timing activation peaks between the first and second trial, while in 3(b) the annotator seems to be refining his assessment regarding activation amplitude while progressing from trial 1 to 3. Overall, we measured the consistency for each set of repetitive trials using Cronbach's α. The mean computed α per annotator ranges between 0.8 and 0.9, except from an annotator with $\alpha = 0.7$. This gives an impression of individual annotator consistency. In the analysis presented in Sec. V-A we only used the final rating which represents the final decision of each annotator. Alternatively, one could also use the average of repetitive ratings when they exist, or could consider each annotator's α as a weight of annotator reliability when combining multiple annotators' decisions.



(a) (b)

Fig. 3. Two examples of annotators performing repetitive ratings of a recording.

C. Comparing Continuous and Discrete Annotations

The availability of both global discrete and detailed continuous ratings from the same annotator and for the same actor-recording, allows us to examine how annotators summarize continuous information to produce an overall judgement. We applied several functions to summarize each continuous rating into a number and examined how close the resulting functional is to the global rating given by the annotator. The functions include mean, median, maximum, minimum, first and third quantile (q1 and q3) of the recording. The discrete ratings were first shifted and rescaled to match the range of the Feeltrace annotations.

Table III shows the mean squared error (MSE) between the discrete ratings and different functionals over all actor-recordings. The last line is the MSE when we choose the closest to the discrete rating between q1 and q3. We notice that the discrete rating is generally closer to either q1 or q3 compared to the other metrics, although it varies per rating to which of the two functionals it will be closer. Hence, the global rating is more influenced by either the highest or the lowest values of a rating during a recording. Specifically, for 66% of the activation ratings the discrete rating is closer to q3, for 59% of the valence ratings the discrete rating is closer to q1, while for dominance there is an almost equal split. This suggests that global judgements of activation tend to be more influenced by the higher activated events of a recording, while global judgments of valence tend to be more influenced by the more negatively valenced events.

TABLE III

MEAN SQUARED ERROR BETWEEN THE DISCRETE RATINGS AND DIFFERENT FUNCTIONS OF CONTINUOUS RATINGS OVER ALL ACTOR-RECORDINGS.

function	activation	valence	dominance
mean	0.13	0.10	0.06
median	0.14	0.10	0.07
max	0.31	0.37	0.24
min	0.71	0.26	0.40
q1	0.22	0.12	0.12
q3	0.14	0.13	0.08
either q1 or q3	**0.07**	**0.06**	**0.03**

It also seems that different raters weight differently the same recording when making an overall decision; for example looking at the clips that were rated by 3 people (which is the large majority) in only about 40% all annotators were consistent as to the quantile that they weighted more, either that was q1 or q3 (this percentage is similar for the 3 attributes). These preliminary findings illustrate the complexity of the human cognitive processing when summarizing emotional content; this processing is influenced by the emotional aspect to be evaluated, the events that are being observed, as well as person-specific characteristics.

VI. OPEN QUESTIONS AND RESEARCH DIRECTIONS

A. Improving tools for continuous annotations

Performing continuous annotations is a challenging task, therefore the availability of suitable annotation tools is important to facilitate this process. Annotation software should ideally be customizable in terms of the number and type of attributes to be annotated, portable and easy to install for users with little technical experience, and easy to learn and use. Feeltrace, being one of the first freely available software for continuous annotation, has been a useful resource to the community. However, this early version was lacking in terms of portability and customization functionality, e.g., it only supported 2-dimensional annotation. Those issues were addressed by the development of its successor Gtrace. Also, we noticed that users were sometimes distracted by the two separate windows in Feeltrace (one for annotation and one for video viewing); in Gtrace where those windows are integrated into a common interface. Further improvements could focus on usability; for example Feeltrace and Gtrace are mouse operated and require continuously pressing the mouse to perform annotation. This can be tiring especially for annotation of long videos. Joystick-based options would be more natural, and even more enjoyable to the user. Regarding other continuous annotation tools, one could also look at the joystick-operated and freely available CMS [10].

B. Absolute vs Relative Ratings

Our annotation results suggest that humans are better at rating emotions in relative rather than absolute terms. Indeed, humans seem to have individual internal rating scales that are culture and personality dependent, among others. Therefore it is easier for multiple people to agree that, for example there has been an activation increase, rather than on the absolute values of activation. This issue was also discussed in [26], where authors propose a rating-by-comparison method for annotation of emotional content in terms of discrete dimensional labels. It would be interesting to study how such rating-by-comparison ideas could be reformulated in the context of continuous annotation, in an attempt to increase inter-annotator consistency.

C. Which attributes can be continuously rated?

Here, we focus on emotional content, however a variety of other attributes could be annotated in a continuous way depending on the application and the analysis focus, i.e., engagement, enjoyment, frustration, hostility, etc. Given that some attributes seem easier to rate than others, e.g., valence generally achieves higher consistency than activation and dominance, it would be interesting to examine to what extent different attributes can be rated consistently in a continuous way. This is also discussed in [16], where authors compare rating consistencies for different emotions and cognitive states. It seems however that this effect could be data dependent. For example, data collected for emotional studies are usually rich in emotional manifestations which may facilitate consistency in emotional annotation, while data collected to examine user experience, i.e., using an interface could be more appropriate for engagement or frustration annotation.

D. Multiple Subjective Ratings

The problem of combining multiple annotators continuous subjective ratings in a way that takes into account annotator-specific and recording-specific characteristics has been discussed in Sec V-A. A related problem is examining the differences in emotional assessment, and hence differences in the obtained annotations, between different annotator groups, e.g., with variable level of expertise. For example, our theatrical improvisations might be rated differently by theater experts as opposed to naive viewers (audience). A related issue pertains to the differences in self-assessment of emotional experience versus assessment of others, which is addressed in [32], or cultural differences in emotional assessment [30].

Such studies that compute statistical properties over user populations may require a large number of annotators per recording. This brings forward the relevant discussion in the literature regarding selecting few expert annotators or many naive annotators. In this work, we have followed the former approach, as a small number of expert annotators was relatively easy to recruit in a university environment, and to coordinate. A similar approach was followed for the SEMAINE database annotation [13], while for the Belfast Naturalistic database ([12]) the large number approach was adopted instead, with some clips being rated over 160 times. Obtaining a large number of annotations per recording can be greatly facilitated by the use of modern crowdsourcing tools, like the Amazon Mechanical Turk (MTurk) which has proven useful for various user studies [33], [34], or

translation tasks [35]. One caveat when using MTurk for subjective ratings is that the researcher should devise a method for assessing the attention of the rater and the quality of the annotations, in order to prevent potential careless annotators from contaminating the results, as discussed in [35], [33]. For the case of continuous ratings, there is yet no available online tool for continuous ratings, that would enable performing such annotations online and linking them to the MTurk service. Such a tool would allow researchers to harness the capabilities of crowdsourcing for large scale data annotation projects.

E. Annotator-specific delays

As discussed in Sec. IV-B, because the continuous annotations are done in real-time and due to cognitive processing, we expect a person-specific delay between the time that an event happens and when its emotional content is annotated. Here, we have tried to reduce such delays by asking the annotators to perform an emotional assessment of each recording before starting the real-time annotation. Other researchers have addressed this issue by performing post-processing of the obtained annotations, e.g., by slightly shifting them in time such that the correlation between multiple annotators' continuous ratings is empirically maximized [18]. A more sophisticated method for warping and combining multiple annotations with variable delays is proposed in [29], by combining Probabilistic Canonical Correlation Analysis (PCCA) and Dynamic Time Warping (DTW). More detailed studies of annotator delays would shed light into the timing of human cognitive processing with respect to events in the environment, and how it is affected by factors such as fatigue or interest. Such studies would require a careful experimental design, where for example each rater's response could be measured with respect to certain predefined emotional events in the videos.

F. Continuous ratings and saliency detection

Continuous ratings could reveal regions of an interaction that are characterized by abrupt changes or extreme values of emotional content. Those could be regions where interesting events happen, in the sense that they catch the viewer's attention or they are prototypical or salient examples of a certain emotional or cognitive state. Such regions may also be weighted more heavily when a person summarizes an emotional experience. Availability of continuous annotations would help us understand what constitutes the salient content of an interaction, and would pave the way towards emotional event detection and summarization in social interactions.

VII. FUTURE DIRECTIONS

In this paper, we have discussed challenges and opportunities regarding annotation and processing of continuous emotional attributes, which is an emerging topic in the affective computing domain. Our discussion has focused primarily on data annotation issues in the context of existing literature, and using as a case study the continuous emotional annotation process that we performed on the large, multimodal CreativeIT dataset.

Despite the many challenges, it is important to keep in mind the potential of this emerging domain to bring researchers a better understanding of how humans continuously assess and respond to emotional stimuli. It could also enable technologies such as naturalistic Human-Computer Interfaces (HCI) that can continuously process a variety of multimodal information from the user(s) as they unfold, monitor the users' internal state and respond appropriately when needed. For example, one could imagine educational computer applications with audiovisual sensing capabilities, that would continuously assess the user's engagement and frustration levels and accordingly modify the educational material. Similarly, health-related applications could continuously monitor a subject's stress and anxiety levels and potentially give useful feedback to the user. Such applications are part of the emerging Behavioral Signal Processing domain that explores the role of engineering in developing health-oriented methods and tools [36]. Finally, affect-sensitive virtual agents in gaming applications could continuously sense and interpret verbal and non-verbal cues of the user in order to estimate enjoyment and satisfaction. Such technologies would bring HCI closer to producing a human-like experience, and would have large impact in domains such as entertainment, education, security and healthcare.

REFERENCES

[1] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, pp. 273–294, 1977.

[2] H. Schlosberg, "Three dimensions of emotion," *Psychology Review*, vol. 61, pp. 81–88, 1954.

[3] D.G. Dastidar M. Grimm and K. Kroschel, "Recognizing emotions in spontaneous facial expressions," in *Proc. of Int. Conf. on Intelligent Systems And Computing (ISYC)*, 2006.

[4] A. Metallinou, M. Woellmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. of Affective Computing*, vol. 3, pp. 184–198, 2012.

[5] A Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," in *Proc. of Affective Computing and Intelligent Interaction (ACII)*, 2007.

[6] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan, "Primitives based estimation and evaluation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[7] D. Wu, T. Parsons, E. Mower, and S. S. Narayanan, "Speech emotion estimation in 3d space," in *Proc. of IEEE Intl. Conf. on Multimedia & Expo (ICME) 2010*, 2010.

[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 Workshop on Speech and Emotion, 2000*, 2000, pp. 19–24.

[9] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmuller, "Emujoy: Software for continuous measurement of perceived emotions in music," *Behavior Research Methods*, vol. 39, pp. 283–290, 2007.

[10] D.S. Messinger, T. Cassel, S. Acosta, Z. Ambadar, and J.F. Cohn, "Infant smiling dynamics and perceived positive emotion," *Journal of Nonverbal Behavior*, vol. 32, pp. 133–155, 2008.

[11] R. Cowie and M. Sawey, "GTrace - General trace program from Queen's, Belfast," http://www.dfki.de/ schroed/feeltrace/, 2011.

[12] E. Douglas-Cowie, N. Campbell, and R.P. Cowie, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, pp. 3360, 2003.

[13] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent.," *IEEE Trans. of Affective Computing, Special issue of Resources for Affective Computing.*, vol. 6, 2011.

[14] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," in *Workshop on Multimodal Corpora, LREC*, 2010.

[15] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *Proc. of ICASSP*, 2011.

[16] 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 *Proc. of LREC*, 2006.

[17] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking.," in *Proc. of ICASSP 2011*, 2011.

[18] M.A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, pp. 92–105, 2011.

[19] M. Woellmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. of Interspeech*, 2008.

[20] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 867–881, 2010.

[21] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing (IMAVIS), Special Issue on Continuous Affect Analysis, in press*, 2012.

[22] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. On Multimedia*, vol. 7, pp. 143–154, 2005.

[23] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV," *IEEE Signal Processing Magazine*, pp. 90–100, 2006.

[24] C. Busso, M. Bulut, C-C Lee, A.Kazemzadeh, E. Mower, S. Kim, J. Chang, S.Lee, and S.Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[25] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag german audio-visual emotional speech database," in *In Proc. of the IEEE Intl. Conf. on Multimedia and Expo (ICME)*, 2008.

[26] Yi-Hsuan Yang and Homer H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, 2011.

[27] K. Audhkhasi and S. Narayanan, "A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels.," *IEEE Trans. on Pattern Analysis and Machine Intelligence.*, 2012.

[28] H. Meng, A. Kleinsmith, and N. Bianchi-Berthouze, "Multi-score learning for affect recognition: the case of body postures," in *Proc. of ACII*, 2011.

[29] M. Nikolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic cca for analysis of affective behaviour," in *Proc. of the 12th European conference on Computer Vision (ECCV)*, 2012.

[30] I. Sneddon, G. McKeown, M. McRorie, and T. Vukicevic, "Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour," *PLoS ONE*, vol. 6, 2011.

[31] P.E. Shrout and J.L. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, pp. 420–428, 1979.

[32] C. Busso and S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Proc. of InterSpeech*, 2008.

[33] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[34] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk : A new source of inexpensive, yet high-quality, data?," *Perspectives on Psychological Science*, vol. 6, 2011.

[35] C. Callison-Burch, "Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk," in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2009.

[36] S. Narayanan and P. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of IEEE, in press*.