IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 26, 2020, accepted December 6, 2020, date of publication December 21, 2020, date of current version January 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046293

Weak Ground Truth Determination of Continuous Human-Rated Data

ANDREJ KOŠIR[1], (Senior Member, IEEE), GREGOR STRLE[1,2], AND MARKO MEŽA[1], (Senior Member, IEEE)

[1]User-adapted Communication and Ambient Intelligence Lab, Faculty of Electrical Engineering, University of Ljubljana, 1000 Ljubljana, Slovenia
[2]Research Centre of the Slovenian Academy of Sciences and Arts, 1000 Ljubljana, Slovenia

Corresponding author: Andrej Košir (andrej.kosir@fe.uni-lj.si)

This work was supported by the project *P2-0246 ICT4QoL—Information and Communications Technologies for Quality of Life*.

ABSTRACT The article presents a novel weak ground truth (WGT) determination procedure on continuous human-rated data. The notion of WGT is essential in cases where there is no direct empirical evidence for an observed construct and human annotations provide the most reliable means for determining the ground truth. The core idea behind the proposed procedure is to transform the ratings to reduce rater bias, maximize inter-rater agreement, and improve WGT. The procedure was evaluated on two behavioral datasets containing continuous annotations of several expressive dimensions. The results show that the procedure improves the size of WGT data by removing the disagreement originating from rater-specific distortions, such as rater mean and scaling bias. The entropy of residuals decreases after WGT optimization, meaning that more relevant information is retained. The average improvement of WGT data size is between 10.1 and 20.9 percentage points, depending on the respective dimension. However, in cases where the rater bias is small, the procedure does not substantially modify WGT. This indicates that the proposed optimization only removes rater biases derived from rater-specific distortions, while retaining and improving valid WGT. The proposed procedure is generalizable on any type and size of continuous or discrete numerical data where multiple raters are involved.

INDEX TERMS Bias removal, continuous data, inter-rater reliability and agreement, weak ground truth.

I. INTRODUCTION

Intelligent systems are increasingly capable of conducting more naturalistic (human-like) interaction with their users [1]–[3]. To this end, various elements of social intelligence are being employed that take into account the emotional engagement, attention, and fatigue of users. One example of this is advanced conversational systems where user engagement needs to be continuously evaluated to provide sustained communication [4]–[8]. To execute such evaluations and improve the interaction, a ground truth of user engagement needs to be obtained continuously and in real-time.

This presents a challenge. State-of-the-art intelligent systems cannot yet make reliable evaluations from behavioral data, especially when taking into account continuous time quantity [9]–[12]. In cases where there is no direct empirical evidence or ground truth for a target construct, human ratings still provide the most reliable means for determining the baseline for the construct or its weak ground truth (WGT).

However, WGT is highly sensitive to several factors that affect its reliability. Among these factors are types of observed phenomenon and latent constructs used in annotation, types of rating procedures, and, particularly, rater specific distortions due to individual differences among raters (e.g. level of expertise and domain knowledge, personality, perceptiveness) that contribute to rater biases and affect the quality of WGT. These issues are even more pressing for WGT determination of continuous data, which is a problem in many behavioral studies because proper statistical methods to offset rater bias are lacking.

The aim of this article is to propose a novel procedure for the determination of WGT on continuous human-rated data. The key contributions of the presented research are:

- WGT determination on continuous annotations where multiple raters are involved;
- WGT optimization and improvement of WGT data size;

The associate editor coordinating the review of this manuscript and approving it for publication was Giuseppe Desolda.

- rater bias removal (the proposed procedure performs the removal of the two most common types of rater bias: a) rater mean bias, which relates to differences in the mean ratings among individual raters, and b) scaling bias, which relates to raters' deviations from their own mean ratings).

Moreover, an online tool was developed and is freely available to researchers in order to assess the proposed procedure and calculate WGT for their data[1].

We build our case on the annotations of behavioral data from two well-known public datasets: the SEMAINE dataset of interactions between humans and artificial agents [13] and the CreativeIT dataset of expressive behaviors and natural interaction between human beings [14]. Both datasets present difficulties because the latent dimensions used in the annotation of behavioral data are highly subjective and thus susceptible to rater bias, and this makes determination of WGT challenging.

The article is organized as follows. An overview of related work on rater reliability and agreement statistics along with alternative approaches to WGT is presented in Section II). A novel WGT determination and bias removal procedure is presented in Section III along with materials used in the study. The results, a demonstration of the performance of the proposed approach on the two datasets, are presented in Section IV. The article concludes with Section V, which presents a discussion and possibilities for future work.

II. RELATED WORK

The problem of the lack of ground truth is prevalent in a wide range of domains [15], including, among others, behavioral studies [16], [17], medical studies [18], and computer vision [19]–[21]. In such cases, human annotation is treated as WGT and rater specific distortion tendencies are to be expected [15], [17]. Because of this, an important aspect of determining WGT is to assess reliability and agreement among raters in order to provide a measure of homogeneity and consensus in their ratings.

Reliability and agreement are two concepts often used interchangeably because they both provide insight into errors in measurement. However, the two concepts differ. Reliability is generally defined as the proportion of agreement between two measurements among or within raters. It is a measurement of raters' consistency and thus the variability among raters. In contrast, agreement is a measure of the degree to which the ratings of two or more raters are identical [22]. The reliability of agreement is measured in two ways: a) as the reliability of a rater over multiple occasions – the intra-rater reliability, and b) as the reliability of multiple raters on a single task – the inter-rater reliability. This second aspect is the primary focus of this article.

A comprehensive overview of statistical methods measuring the reliability of agreement is given by [23]–[26]. Several reliability measurement methods exist, from a simple percent agreement to more complex Kappa statistics. The percent agreement (number of agreement scores/total scores) is problematic as it does not account for the agreement made by chance (i.e. due to raters guessing) and can thus overestimate the inter-rater agreement. The Kappa statistic was introduced to control for this issue. The Kappa coefficient is a statistical measure of inter-rater agreement that is used to determine the agreement between two or more raters when the measurement scale is categorical. It takes into account the element of chance to measure "true" beyond-chance agreement [27].

Among the most known coefficients for measuring the reliability of agreement are Cohen's kappa [27], weighted kappa [28], Fleiss's kappa [29], Krippendorff's alpha [30], and the Intra-class correlation coefficient (ICC) [31], [32]. Cohen's kappa measures the coefficient of agreement between a pair of deliberately chosen raters on a nominal scale. Various extensions of Cohen's kappa exist, depending on the type of data and the number of raters. For example, Fleiss's kappa is an extension of Cohen's kappa that can measure the agreement among randomly selected multiple raters on a nominal scale [31].

The main drawback of the many statistical measurements of reliability, with the exception of Krippendorff's alpha and ICC, is that they don't differentiate between various kinds of disagreements [23]. Krippendorff's alpha is a widely-used reliability statistic because it provides a measure of reliability for several types of data (nominal, interval, ordinal, and ratio). It can also deal with incomplete data and account for disagreement by calculating inter-rater reliability as a ratio of observed disagreement versus expected disagreement [30], [33]. However, Krippendorff's alpha cannot measure the reliability of agreement on a continuous, time-dependent quantity [23], [26]. ICC, a special case of Krippendorf's alpha for continuous data, is commonly used for this purpose. ICC is based on an analysis of variance (ANOVA) models and measures the proportion of total variance accounted for between subject variation [32], [34].

Standard reliability statistics are based on the average of the ratings and aim to confound or remove disagreement, which is typically seen as annotation noise, or rater bias. This is more straightforward when dealing with discrete annotations, where raters agree or disagree on the chosen label, in contrast to continuous annotations. For continuous annotations, the ground truth is generally determined as a framewise mean of the ratings of various annotators [17]. There are several sources of rater disagreement and bias that can increase inter-rater variability, including time-shifting bias (represented by a rater-specific delay in annotations), rater mean bias (where raters annotate around different means and some raters' mean ratings are higher or lower than others), and scaling bias (the magnitude of deviations from the mean ratings) [16], [17], [35], [36]. Often ambiguity occurs in the ratings due to different interpretations of a construct observed by human raters, which, as several studies show, is particularly common in the annotation of behavioral data [15], [36]–[38].

[1] https://www.lucami.org/en/WGT/

Several approaches have been developed in order to deal with these issues. Rater-specific time delays can be addressed in post-processing by aligning the annotations before computing WGT [16], [17]. Because this procedure is very time consuming when conducted manually, several automated solutions have been developed, such as the use of framewise binning to group annotations [16]. More advanced approaches based on machine learning are being adopted to account for various types of rater-specific distortions, including rater bias and delay. These include a latent time warping process and generative probabilistic model based on dynamic Probabilistic Canonical Correlation Analysis (PCCA) [10] to solve the problems of temporal alignment and fusion of multiple annotations, other research models based on Expectation-Maximization algorithm [17], [39] and generalized additive mixed models [11], [40], [41] for fusing multiple continuous annotations.

Several novel approaches also combine machine learning with crowd-sourced data [12], [20], [21], [42]. For example, [42] combined machine generated labels derived from human generated data and deep convolutional neural networks (DCNN) to create a baseline for the classification of a dataset composed of approximately eight million videos. DCNN were also used by [12] to provide the baseline for annotations of affective dimensions (valence/arousal, liking/disliking), whereas [21] employed a probabilistic graphical annotation model to infer the underlying ground truth (as categorical distribution) and evaluate annotators' behavior and reliability.

In general, machine learning approaches require large amounts of training data and cannot be effectively applied to smaller datasets which are prevalent in behavioral studies. In such cases, alternative approaches to bias removal and WGT determination are more appropriate. These are generally based on the use of various combinations of truncated mean (or, alternatively, a weighted truncated mean) and correlation metrics in order to calculate and maximize the inter-rater agreement needed for WGT. The truncated mean approach is simple and, by omitting the lowest and highest ratings, can mitigate random effects, such as a rater's mind wandering and the temporary loss of attention. The weighted truncated mean approach also mitigates random effects but relies on different weight estimation techniques in order to calculate WGT.

To our knowledge, the only comparable alternatives to the WGT determination of continuous human-rated data proposed in this article are given by [16] and [35]. Both approaches are based on the truncated mean and use correlation metrics to account for variability among the raters. For example, [35] use correlation metrics (Pearson's correlation) to measure agreement among multiple raters on the CreativeIT dataset of behavioral annotations. To control for the variability of raters (the rater mean bias), a correlation threshold is used to remove highly inconsistent ratings and then the ground truth is computed with the mean ratings.

An alternative approach to WGT determination is proposed by [16]. WGT is produced by maximizing the inter-rater agreement based on the correlation and sign agreement statistics among pairs of two raters. Then the weighted truncated mean of agreement is calculated with these variables. In addition to WGT determination, [16] also propose a solution to automatically segment large sessions of audio-visual data for the purpose of machine learning approaches that cannot deal with unsegmented sequences. This approach was tested on a dataset of video annotations but could be generalized to other annotation tasks.

Neither of these two approaches addresses rater biases explicitly. [35]only partially address the problem of rater bias. The authors observe that raters often agree on relative but not absolute terms, but their approach does not distinguish between different types of bias. Because their approach is based on correlation statistics, it is affected by scaling bias, which is independent of rater mean bias. For each rating session, rater agreement is defined on the rater pairs with linear correlations greater than the preset cut-off threshold. The estimated value (or WGT, as used in this article) is then obtained by averaging the ratings of selected raters with no reported attempt of removing direct bias. The approach by [16] proposes solutions for ground truth determination and video segmentation of rated sessions (for the purpose of machine learning), and does not directly address various types of rater bias. However, this approach does make use of local normalization (for each coder and each session) in order to ''avoid propagating noise in cases where one of the coders is in large disagreement with the rest'' [16, p. 44]. The maximization of inter-rater agreement is implemented by maximizing the number of participating raters based on their pair comparisons, with the aim of producing WGT. By applying weighted averaging, this procedure may also indirectly account for rater disagreement by assigning lower weights to uncorrelated raters.

Our solution also utilizes the maximization of the inter-rater agreement, but on different grounds. It employs the weighted truncated mean approach and operates on a family of rating transformations that are used to directly remove individual rater biases. It employs a two-way randomized design and a single observation ICC where absolute value matters (see Subsection III-D). The proposed approach addresses the removal of rater bias by direct maximization of the inter-rater agreement on the transformed ratings, which is based on the inter-rater agreement calculated by the ICC for the whole rating session and not only for pairwise correlations.

It is difficult to conduct a direct comparison of the methods proposed by [35] and [16] using objective metrics. Since there is no ground truth, it is impossible to objectively compare the quality of WGT generated by these methods, or, for example, the effect size of bias removal. The two methods differ conceptually and do not directly address the removal of rater bias, which is the key advantage of the proposed procedure in terms of improving the quality of WGT.

The proposed procedure is straightforward, robust, configurable, and generalizable. Only the raw ratings are input into the procedure, without any need for sampling or segmentation of material. It can be applied to any task where the data types (rated quantities) are discrete or continuous numerical variables. The distribution of the data and its size is irrelevant. This makes the procedure straightforward to apply. The following sections discuss the proposed bias removal and WGT determination procedure in detail.

III. MATERIALS AND METHODS

A. MATERIALS

Materials were taken from two behavioral datasets widely used in affective computing research: the CreativeIT dataset [43] and the SEMAINE dataset [13]. Both datasets focus on expressive affective behavior and emotionally charged interaction, with annotation attributes that are largely subjective [14].

The CreativeIT is a multimodal dataset of theatrical improvisation [14], [43]. It was designed to study theatrical improvisation, affective and expressive behaviors, and natural human interaction, and has been extensively used in the domains of affective computing, emotion recognition, and annotation [35], [44]–[47]. The dyadic interactions performed by pairs of actors were recorded using video camera and motion capture technology.

The publicly available CreativeIT dataset contains eight recordings (each approximately an hour long) divided into 300 sessions containing the improvisations of 40 two-sentence exercises, with a total of 19 actors involved. Multiple raters (n=8), separated into three groups (actor, expert, novice), were used to annotate the data from the videos along the discrete dimensions for theatrical performance (naturalness, creativity) and the continuous emotional dimensions of valence, arousal (or activation), and dominance [14]. Each recorded session was annotated by three raters.

The SEMAINE audiovisual dataset was built in order to motivate research on systems that support sustained, emotionally-charged interaction with artificial agents [16], [35], [39], [43], [48]. The SEMAINE dataset contains recorded face-to-face emotional conversations between 150 participants (users) and an artificial agent (Sensitive Artificial Listener - SAL) [13]. It contains 959 audiovisual recording sessions lasting approximately five minutes each [13]. These sessions were annotated by two to eight raters and include the annotations of perceived emotions taken from audiovisual cues (acoustic cues and facial expressions) in the conversations, with several descriptors used for the five affective dimensions of Valence, Arousal/Activation, Power, Anticipation/Expectation, and Intensity.

To preserve the flow of affective interactions, the annotations in both datasets were recorded continuously using the Feeltrace annotation tool [49]. The Feeltrace tool enables continuous annotations over time by using a mouse to mark the annotation (as a value ranging from -1 to 1) in two-dimensional affective space. The valence and arousal dimensions are annotated in the Feeltrace interface as a pair of orthogonal dimensions in two-dimensional affective space. For the dimensions that need to be traced individually as a single scale (such as dominance, intensity, power, anticipation), the Feeltrace interface was slightly adapted to allow for annotation on a single (one-dimensional) scale [13], [14].

Materials used in the presented study include a subset of the SEMAINE dataset and the entire publicly available CreativeIT dataset. The CreativeIT dataset contains the annotations of Valence, Activation (or arousal), and Dominance dimensions over 300 sessions, rated by three raters per session. The subset of the SEMAINE dataset includes only the dimensions with at least 15 sessions per dimension rated by three or more raters, resulting in a total of 406 sessions. The threshold on the number of sessions for each dimension was applied in order to obtain a balanced representation of the dimensions used in WGT optimization. The final set of descriptors for the five core dimensions of the SEMAINE dataset used in the study includes: Agreeing, Thoughtful, Gives Information, Gives Opinion, Activation, Intensity, Power, Valence, and Amusement. The terms *arousal* and *activation* are used interchangeably in the literature (see [50], [51]), but we use *activation* because this term is reported in the results of both studies [13], [14]. The number of sessions for each of the selected dimensions is shown in Table 1.

B. OVERVIEW OF WGT DETERMINATION PROCEDURE

The core idea behind the proposed WGT determination procedure is to calculate the weighted truncated mean of the ratings and define WGT on session intervals where inter-rater agreement is high enough. We transform individual ratings in order to decrease the rater bias (errors), maximize the inter-rater agreement, and increase the size of WGT data.

A pipeline for the WGT determination procedure is presented in Figure 1. We treat the problem of inter-rater agreement estimation and the problem of WGT determination as a single problem. Our reasoning is as follows. As each step of the rating procedure could represent a potential source of rater bias, we first analyze and define a class of parameterized transformations of the ratings. These transformations are defined according to the factors that caused the disagreement originating from the rater bias. Next, we maximize inter-rater agreement for the transformed ratings in order to obtain optimal parameters of rating transformations. We argue that the corrected (optimally transformed) ratings are a more realistic representation of inter-rater agreement than the original values because of the removed rater bias. Moreover, as demonstrated later (Section IV-D), the proposed optimization procedure only removes the disagreement originating from rater bias and preserves the true disagreement among raters. We compute WGT based on the weighted truncated mean of the transformed ratings where ICC is above the predefined cut-off threshold.

TABLE 1. Comparison of WGT Data Size Before and After Optimization for Each Dimension and Dataset. The Table Shows Number of Sessions for Each Dimension (N), the Min and Max Values for WGT, Average WGT Data Size Before and After Optimization (rWGT vs. rWGT_opt), and Average Improvement (impr. pp) of WGT Data Size in Percentage Points After Optimization.

SEMAINE dataset								
Dimension	N	rWGT	min	max	rWGT_opt	min	max	impr. pp
Gives Info	42	51.4%	0	88.5%	61.6%	0	91%	10.2
Gives Opinion	28	32%	0	76.2%	43.6%	0	87.1%	11.6
Thoughtful	15	24.8%	0	83.3%	45.8%	4.1%	83.8%	20.9
Activation	60	2%	0	22.3%	12.3%	0	51.9%	10.2
Intensity	60	5.5%	0	39.7%	22.9%	0	56.1%	17.5
Power	60	5.4%	0	37.3%	15.5%	0	73.9%	10.1
Valence	60	12.7%	0	53.8%	25.5%	0	86.8%	12.8
Amusement	39	38.3%	6.4%	85.2%	55.1%	20.7%	86.4%	16.8
Agreeing	42	29.6%	0	83.6%	47.6%	7%	84.2%	18
CreativeIT dataset								
Dimension	N	rWGT	min	max	rWGT_opt	min	max	impr. pp
Activation	100	24.8%	0	86.9%	42.9%	0	91.4%	18.1
Dominance	100	16.3%	0	81.5%	35.0%	0	85.1%	18.7
Valence	100	20.3%	0	82.0%	35.5%	0	87.1%	15.2

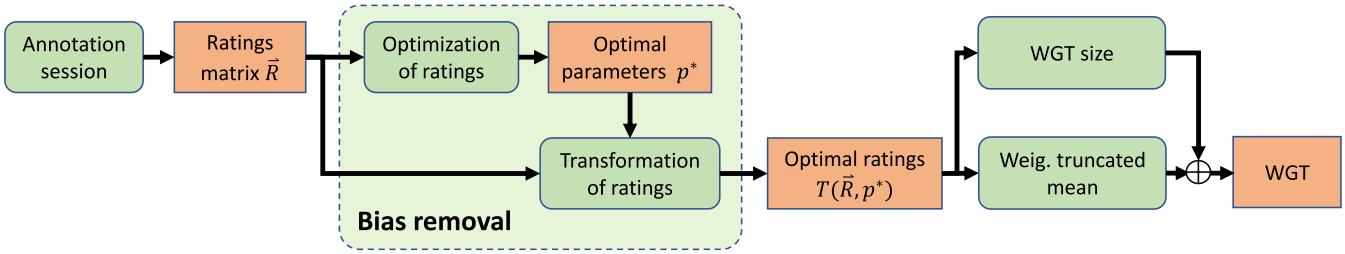


FIGURE 1. Pipeline of WGT determination procedure. The annotation session results generate a ratings matrix $\vec{R}$. The output of the optimization of the ratings maximizing the inter-rater agreement is the optimal set of the transformation parameters p^* that is used to produce optimal ratings $T(\vec{R}, p^*)$. The individual rater biases (mean and scaling bias) are removed during the optimization and transformation of the ratings (within the dashed-line rectangular). The weighted truncated mean of these optimal ratings yields WGT whereas the minimal reliability criteria is used to determine WGT data size.

The inter-rater reliability is estimated in order to decide whether the rating procedure was successful and the obtained ratings can be reliably used for WGT. For this purpose, a cut-off threshold value must be defined to account for a level of agreement adequate to produce WGT. The cut-off threshold value depends on the coefficient used to estimate it, and is relatively independent of the domain of measurement. WGT obtained from the weighted truncated means of the ratings is then determined only where the achieved inter-rater agreement exceeds a predefined cut-off threshold value.

The following sections present WGT determination procedure in detail.

C. INTER-RATER RELIABILITY AND AGREEMENT

The proposed approach measures the inter-rater reliability using ICC and assuming the proper ANOVA model. We use a two-way randomized design and single observation ICC where the absolute value is relevant (see [24]). Rated quantity is denoted by y, a vector of their values by $\vec{y} = [y_j]_j$ where $j = 1, \ldots, n$ is the time index and y_j its value, and n is the number of ratings provided by a single rater. The raters are denoted by their indices $i = 1, \ldots, N$ (N is the number of raters in a single rating session). The rating of a rater i of the quantity y_j is denoted by r_{ij}. A vector of the ratings for rater i is $\vec{r}_i$ and the matrix of the ratings $N \times n$ is denoted by a capital letter $\vec{R} = [r_{ij}]_{ij}$ (the rating vectors are rows). The estimation of the rated quantity y_j at a time index j is denoted by $\hat{y}_j$ and the vector of its estimations by $\hat{\vec{y}}$. We call this estimation the weak ground truth (WGT).

We denote the inter-rater agreement of the rating matrix $\vec{R}$ by $\rho(\vec{R})$ and the reliability of these ratings by $\gamma(\vec{R})$.

The inter-rater agreement is defined as

$$\rho(\vec{R}) = \frac{\#(\text{not stat. diff. ratings})}{\#(\text{all ratings})}, \tag{1}$$

where # stands for the "number of". To count the values in the nominator correctly, we apply statistical hypothesis testing. We follow the reasoning given in [52] where the authors clearly distinguish between reliability and agreement while acknowledging the close connection between the two concepts. They take into account only statistically significant disagreements among raters, as opposed to all disagreements. To do so, they employ the concept of the Reliable change index (RCI), which relates to inter-rater reliability γ. From this index, we derive the critical difference between the two ratings

$$\Delta = z_\alpha \sqrt{2} s_1 \sqrt{1 - \gamma}, \tag{2}$$

where $z_\alpha = 1.96$ at assumed risk level $\alpha = 0.05$, s_1 is the rating's standard deviation, and γ is the inter-rater

reliability (here measured with ICC). The two ratings r_a, r_b are significantly different at risk level α if, and only if, their difference is larger than the critical difference $\Delta < |r_b - r_a|$. For this step, we reason that the raters are in better agreement the lower the critical difference between the two test scores Δ. According to Eq. (2) above, the optimization task of maximizing the inter-rater agreement argmax $\rho(\vec{R})$ is equivalent to the maximization of the reliability argmax $\gamma(\vec{R})$, see Eq. (4). Since inter-rater agreement and reliability have the same maximums, we maximize the inter-rater reliability because the implementation of the optimization procedure is simpler for reliability than for inter-rater agreement.

D. PARAMETERIZED RATING TRANSFORMATION

WGT determination is formulated as an optimization problem (Eq. (4)), maximizing the inter-rater agreement in the optimization space of the rating transformation parameters $\vec{p}$.

We identify two types of rater disagreement: a) true disagreement among the raters, and b) disagreement originating from rater bias. Clearly, to better estimate WGT, true disagreement should be preserved whereas the disagreement based on rater bias should be removed. For this purpose, we implement a parameterized rating transformation $T(r, \vec{p})$ in order to remove two common types of rater bias: a) mean bias (differences in mean ratings between individual raters), and b) scaling bias (the range or the magnitude of deviations from the mean ratings for an individual rater).

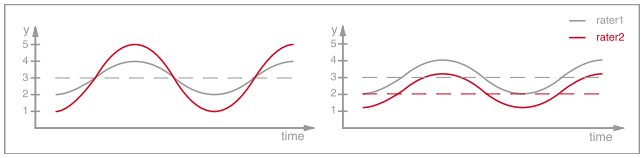


FIGURE 2. Conceptual representation of rater biases with discrete ratings depicted as continuous curves. Left: scaling bias or magnitude of deviations from the mean ratings. Right: rater mean bias (the differences between rater mean ratings) where raters rate around different means.

First, we address scaling bias. Assume that two raters rate around level 3. For the same observed event, the ratings of the first rater vary between levels 2 and 4, and the ratings of the second rater vary between 1 and 5. They may agree on the observations but they disagree on the range, and this generates a scaling bias (see Figure 2, left). We remove the scaling bias by introducing scale parameter a (see Eq. 3). Second, we address mean bias. For the same observed event, if two raters rate around different means but with similar amplitude (e.g., one rater rates around level 3 and the other around level 4), they may agree, but the difference in their means generates rater mean (or offset) bias (see Figure 2, right). We remove the mean bias by introducing the additive parameter denoted by b (see Eq. 3).

The inversion of offset and scaling bias is obtained with a linear transform (see Eq. (3)). Altogether, we parameterize the rating transformations as

$$T(\vec{r}, (a, b))(t) = a \cdot \vec{r}(t) + b, \tag{3}$$

where $\vec{p} = (a, b)$ is a set of parameters.

E. REMOVING RATER BIAS BY MAXIMIZING INTER-RATER AGREEMENT

The transformation of ratings is an inverse transformation of the rater biases. For instance, if one rater has a consistently positive bias, the transformation will remove this bias by adding a negative value. The optimization problem (see Eq. 4) is defined where $\vec{p}_i = (a_i, b_i)$ is a pair of the rating transformation parameters for the i-th rater. The first parameter a_i represents scaling bias and the second parameter b_i represents offset bias (i.e., the transformation maps the rating r_i as $T(r_i, a_i, b_i) \mapsto a \cdot r + b$). These pairs (a_i, b_i) are combined into a vector of transformation parameters $\vec{p} = (\vec{p}_i)_i$, $i = 1, \ldots, N$, which is a tuple of tuples of the rating transformation parameters. For example, if $N = 5$ raters, the vector $\vec{p}$ has 10 entries. The regularization term $\beta\|\vec{p}\|^2$ prefers smaller transformations over larger ones, with $\|\vec{p}\|$ being the Euclidean distance and the regularization parameter $\beta = 0.1$ working well in practice.

The unconstrained optimization task is presented by the following equation

$$(\vec{p}^*) = \text{argmax}_{\vec{p}} \gamma(T(\vec{r}_i, \vec{p}_i)_i) + \beta\|\vec{p}\|^2 \tag{4}$$

The optimal transformation parameters $\vec{p}^*$ define the corrected ratings $T(\vec{r}_i, \vec{p}_i^*)$ for each rater $i = 1, \ldots, N$. The initial solution $(a_i, b_i) = (1, 0)$, representing the identity transformation (no change to the ratings), leads to a stable optimization result. Note that the maximized inter-rater reliability $\gamma(T(\vec{r}_i, \vec{p}_i^*)_i)$ is understood as the achieved inter-rater agreement, and not the original one $\gamma(\vec{r}_i)$.

F. WGT DETERMINATION

WGT $\hat{y}(t)$ of the quantity $y(t)$ is determined using the transformed ratings $T(\vec{r}_i, \vec{p}_i^*)$ where the transformation parameters $\vec{p}_i^*$ are set to minimize rater bias and maximize inter-rater agreement see Eq. (4).

After bias removal, WGT determination proceeds with the following formula:

- the weighted truncated mean is applied in order to leave out the most deviated ratings;
- a higher weight is assigned to the ratings in higher agreement (based on the assumption that higher agreement among raters contributes more to WGT).

Weighting (w_i) is assigned using a leave-one-out method by measuring the individual rater's i reliability versus the overall reliability among raters $\vec{R}_i = \vec{R} \setminus \vec{r}_i$ (with i-th row removed). Individual contributions are then normalized to obtain the weights $w_i = \gamma(\vec{R}_i)/\sum_k \gamma(\vec{R}_k)$. The weighted truncated mean is defined as

$$\hat{y}_j = \frac{1}{N-2} \sum_{i=1+a/2}^{N-a/2} w_i \cdot \text{sort}(T(\vec{r}_i, \vec{p}_i^*)), \tag{5}$$

where a is a number of values left out (applicable values are even values $a = 0, 2, 4$ less than $N - 1$) and sort stands for the ratings sorted according to their values. This is the the final step of the WGT determination procedure. WGT is determined only from the truncated means (Eq. 5) where optimized reliability exceeds the cut-off threshold, that is $\gamma(T(r_i, \vec{p_i}^*)) > T_\gamma$. The $D_{WGT} = \{k : \gamma(T(r_i, \vec{p_i}^*))[k] > T_\gamma\}$ represents the size of WGT.

ICC is computed at each step within the predefined time interval of a sliding window. At any given time point, ICC is estimated within the sliding window in a time-local manner. This local ICC value is then computed against the cut-off threshold in order to evaluate if ICC at this particular time point is high enough to compute WGT or not. The length of the interval for the sliding window is chosen as a compromise between the following two competitive features: 1) high time dynamics (where a shorter sliding window is preferable), and 2) short confidence intervals of ICC estimation (where a longer sliding window is preferable).

IV. RESULTS

The proposed WGT determination procedure was tested on the SEMAINE and CreativeIT datasets. WGT is determined where the inter-rater agreement is above the ICC threshold of 0.2 and the time interval of the sliding window is set to 40s. The relatively low value of the cut-off threshold for ICC is due to low inter-rater agreement typically found in the annotation of behavioral data and also to the ICC formula appropriate in this setting. The length of the interval for the sliding window was chosen as a compromise between the two competitive features mentioned in the previous section, the high time dynamics and the short confidence intervals of ICC estimation.

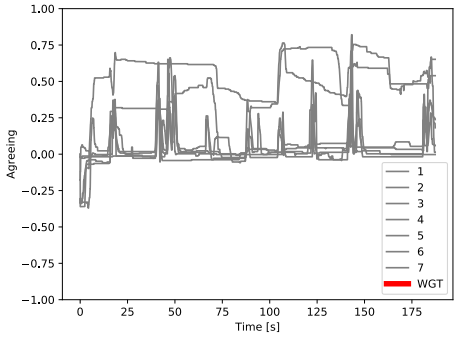
In the subsequent section, we report on the following aspects:

A. the effect of optimization on the size of WGT data: by removing the rater bias we expect to improve the size of WGT data (see Section IV-A);
B. the effect of optimization on the ratings' min-max intervals: the reduction of the min-max interval indicates the reduction of scatter among the transformed ratings (see Section IV-B);
C. the effect of optimization on the entropy of residuals: the size of residuals indicates whether the optimization procedure truly eliminates rater bias from the ratings where bias is present (see Section IV-C);
D. the effect of optimization on rater disagreement to verify that the proposed optimization preserves the true disagreement among raters (see Section IV-D).

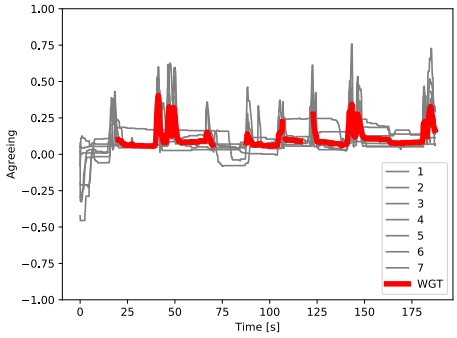
A. THE EFFECT OF OPTIMIZATION ON WGT DATA SIZE

The optimization procedure is expected to improve the inter-rater agreement and the size of WGT data. Moreover, by removing the disagreement originating from rater bias, we also expect to obtain a more reliable measure of WGT.

The overall effect of optimization is shown in Table 1. The WGT optimization procedure improved the size of WGT data in almost all cases, however, the improvement varies in different dimensions. For instance, the largest average improvement (20.9 percentage points) is measured in the "Thoughtful" dimension (from the SEMAINE dataset), with WGT data size increasing from 24.8% before to 45.8% after optimization. The smallest average improvement of WGT data size is 10.1 percentage points in the "Power" dimension (from the SEMAINE dataset), increasing from 5.4% before to 15.5% after optimization. Optimization was not always successful. In a fraction of cases from the CreativeIT dataset (3% of the samples), optimization actually reduced the original WGT data size with a median (IQR) reduction of -1.4 percentage points (-3.9 to -0.7 percentage points). We suspect this was due to time-shifting bias where rater-specific delays in annotation did not generate sufficient inter-rater agreement for the predefined WGT interval.



(a) Before the optimization



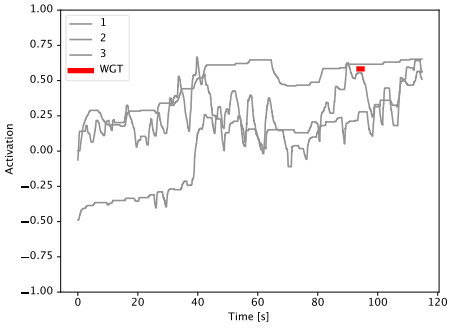
(b) After the optimization

FIGURE 3. SEMAINE dataset: An example of optimization effect with a significant improvement of WGT data size (red line). The session was rated by 7 raters (grey line).

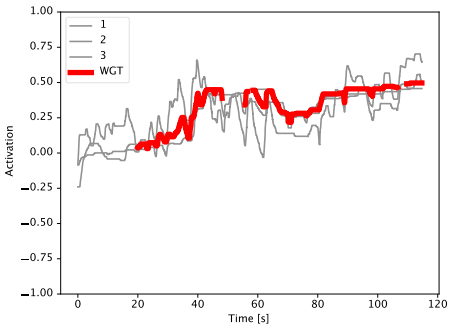
Time graphs (Figures 3–7) provide further details on how optimization affects individual dimensions in terms of increased WGT data size. The selected examples compare WGT data size before and after optimization and illustrate the optimization effect in terms of the maximum (Figure 3), median (Figure 5), and minimal (Figure 7) improvement. The results are shown for selected sessions and for different dimensions of both datasets, with the number of raters varying in each example from the SEMAINE dataset. Examples of

TABLE 2. The Optimization Effect on the Size of the Min-Max Interval for Each Dimension and Dataset. The Table Shows the Number of Sessions for Each Dimension (N), the Min-Max Interval for WGT Before and After Optimization (rMinMax vs. rMinMax_opt), the Min and Max Values of the Interval, and the Average Improvement (impr.) of the Interval Size (Lower is Better) After the Optimization.

SEMAINE dataset								
Dimension	N	rMinMax	min	max	rMinMax_opt	min	max	impr.
Gives Info	42	1.098	0.356	1.693	0.819	0.207	1.581	-0.279
Gives Opinion	28	1.08	0.306	1.528	0.81	0.267	1.528	-0.27
Thoughtful	15	0.495	0.305	0.687	0.282	0.182	0.412	-0.213
Activation	60	0.746	0.46	1.124	0.528	0.178	1.124	-0.219
Intensity	60	0.665	0.314	1.037	0.37	0.18	1.037	-0.296
Power	60	0.696	0.324	1.108	0.481	0.214	1.108	-0.215
Valence	60	0.471	0.262	0.952	0.313	0.102	0.952	-0.158
Amusement	39	0.551	0.238	0.992	0.383	0.218	0.642	-0.169
Agreeing	42	0.492	0.138	0.744	0.338	0.106	0.65	-0.154
CreativeIT dataset								
Dimension	N	rMinMax	min	max	rMinMax_opt	min	max	impr.
Activation	100	0.540	0.219	4.596	0.333	0.127	4.596	-0.207
Dominance	100	0.457	0.125	0.951	0.279	0.001	0.951	-0.178
Valence	100	0.370	0.065	0.960	0.226	0.041	0.960	-0.145

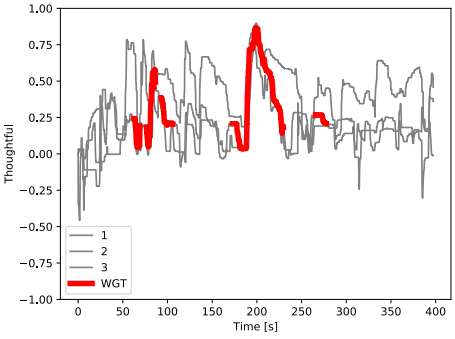


(a) Before the optimization

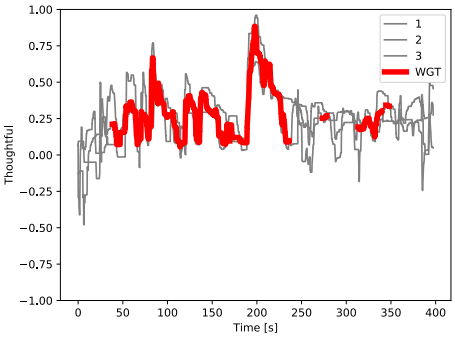


(b) After the optimization

FIGURE 4. CreativeIT dataset: An example of the optimization effect with a significant improvement of WGT data size (red line). The session was rated by 3 raters (grey line).



(a) Before the optimization



(b) After the optimization

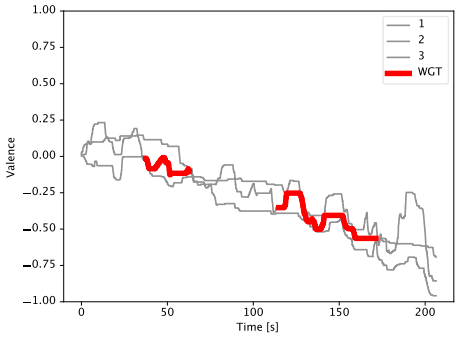
FIGURE 5. SEMAINE dataset: An example of the optimization effect with a median improvement of WGT data size (red line). The session was rated by 3 raters (grey line).

significant WGT optimization effect are shown in Figures 3 and 4. Initially, WGT could not be determined due to the inter-rater reliability being below the ICC threshold for the entire session. After WGT optimization, a significant improvement in WGT data size was obtained in several cases (see Figures 3 and 4, and refer also to Table 1).

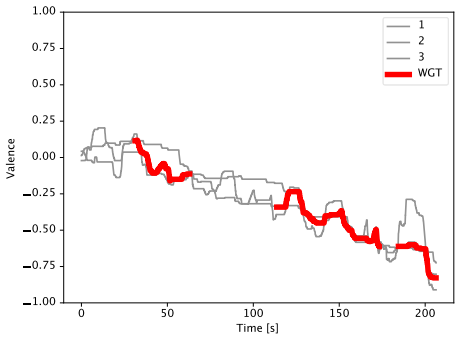
B. THE EFFECT OF OPTIMIZATION ON THE SIZE OF THE RATINGS' MIN-MAX INTERVALS

This subsection illustrates how the proposed optimization procedure impacts the size of the ratings' min-max intervals. The min-max interval is calculated by subtracting the lowest rating value from the highest (for example, the min-max interval of the three ordered ratings a, b, c is $[a, c]$ and its size is $c - a$). A smaller min-max interval means a lower scatter of the transformed ratings.

Table 2 shows the effect of optimization on the size of the min-max intervals for all sessions by dimension. The reduction of the min-max interval indicates an improved WGT estimation. Optimization was successful for all dimensions,

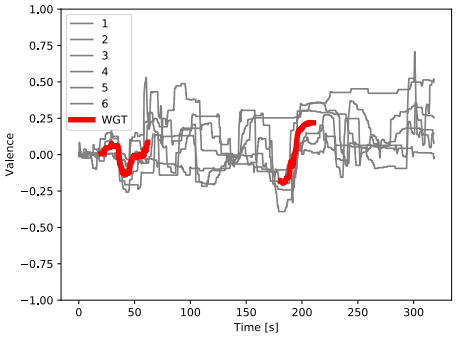


(a) Before the optimization

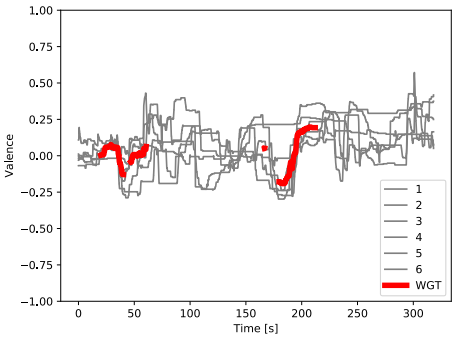


(b) After the optimization

FIGURE 6. CreativeIT dataset: An example of the optimization effect with a median improvement of WGT data size (red line). The session was rated by 3 raters (grey line).



(a) Before the optimization



(b) After the optimization

FIGURE 7. SEMAINE dataset: An example of the optimization effect with a minimal improvement of WGT data size (red line). The session was rated by 6 raters (grey line).

with the size of min-max intervals after optimization reduced from between 15.4 and 29.6 percentage points.

The time graphs in Figures 9 and 11 show the best and worst examples of the optimization effect on the size of the min-max interval. In the best case, there is a significant reduction in the size of the min-max interval. However, in the worst case, the reduction is close to zero and has no practical impact.

C. THE EFFECT OF OPTIMIZATION ON THE ENTROPY OF RESIDUALS

The entropy of residuals is another aspect of measuring the optimization effect on WGT data size. The rating residuals of WGT are computed before ICC thresholding. The residuals are computed for each rater's vectors $\vec{r}_i$, $i = 1, \ldots, N$ from a set of N rating vectors. These N rating vectors are input into the weighted truncated mean. Eq. (5) is used to obtain WGT ($\hat{y}$) where the residuals are defined as

$$\vec{s}_i = \vec{r}_i - \hat{y}, \qquad i = 1, \ldots, N, \tag{6}$$

(see notations in Subsec. III-C).

As an indicator of the improvement achieved by the proposed procedure, we report on the entropy of residuals before and after optimization. As shown in Table 3, the entropy of residuals decreases after optimization. A lower entropy of residuals means more relevant information is retained in the ratings. The improvement is small but consistent across all dimensions.

D. THE EFFECT OF OPTIMIZATION ON RATER DISAGREEMENT

We further examine how the proposed optimization procedure affects WGT in terms of the type of rater disagreement. This is done in order to verify that the proposed optimization removes only the disagreement originating from rater bias and not true disagreement among raters. In cases where the rater bias is relatively small, the procedure should not substantially modify the value of WGT.

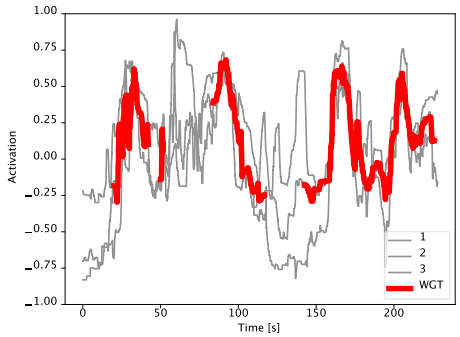
We tested this hypothesis on WGT and min-max intervals. As shown in Figure 13, optimization reduced the size of the min-max intervals (compare the green and the magenta lines). However, comparing the segments with the valid (above the cut-off threshold) WGT before optimization (the yellow line) and after optimization (the red line), we do not observe any significant changes in WGT value for the respective segments. This points to the optimization procedure being sensitive to rater bias, with true disagreement among raters being preserved. A similar effect can be observed in Figures 9–12.

V. DISCUSSION AND FUTURE WORK

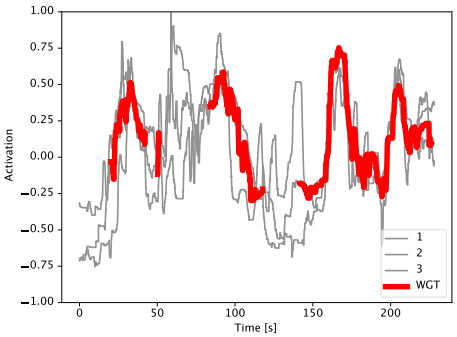
The notion of WGT is essential in cases where there is no ground truth for an observed construct. The problem of

TABLE 3. The Entropy of Residuals Before and After Optimization for Each Dimension and Dataset. The Table Shows the Number of Sessions for Each Dimension (N), the Min-Max Values of Residuals, the Values of Residuals Before and After Optimization (rResid vs. rResid_opt), and the Average Improvement (impr.; Lower is Better) Due to the Optimization Procedure.

SEMAINE dataset								
Dimension	N	rResid	min	max	rResid_opt	min	max	impr.
Gives Info	42	0.505	0.296	0.662	0.375	0.163	0.584	-0.131
Gives Opinion	28	0.532	0.362	0.696	0.380	0.146	0.679	-0.145
Thoughtful	15	0.249	0.168	0.312	0.152	0.098	0.200	-0.097
Activation	60	0.261	0.170	0.389	0.185	0.069	0.389	-0.076
Intensity	60	0.233	0.114	0.356	0.130	0.071	0.356	-0.103
Power	60	0.250	0.125	0.416	0.171	0.082	0.373	-0.078
Valence	60	0.168	0.098	0.268	0.110	0.039	0.251	-0.058
Amusement	39	0.238	0.140	0.360	0.166	0.106	0.237	-0.071
Agreeing	42	0.217	0.109	0.354	0.152	0.079	0.292	-0.065
CreativeIT dataset								
Dimension	N	rResid	min	max	rResid_opt	min	max	impr.
Activation	100	0.374	0.129	9.355	0.265	0.068	9.355	-0.109
Dominance	100	0.231	0.067	0.523	0.129	0.007	0.523	-0.099
Valence	100	0.207	0.093	0.554	0.109	0.025	0.300	-0.097



(a) Before the optimization



(b) After the optimization

FIGURE 8. CreativIT dataset: An example of the optimization effect with a minimal improvement of WGT data size (red line). The session was rated by 3 raters (grey line).

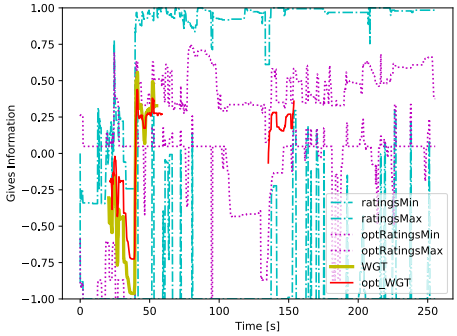


FIGURE 9. SEMAINE dataset: An example of the optimization effect with a significant reduction of the min-max interval size and an improvement of 1.08. The figure shows the non-optimized (cyan line) and the optimized (magenta line) min-max ratings as well as the WGT data size before optimization (yellow line) and after (red line).

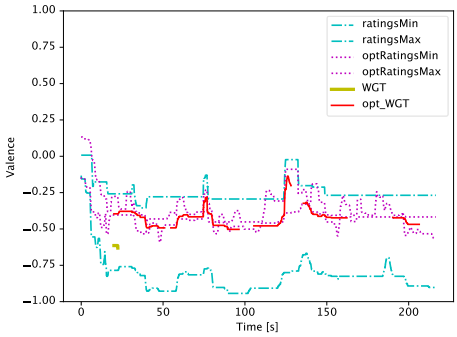


FIGURE 10. CreativeIT dataset: An example of the optimization effect with a significant reduction of the min-max interval size and an improvement of 0.463. The figure shows the non-optimized (cyan line) and the optimized (magenta line) min-max ratings as well as WGT data size before optimization (yellow line) and after (red line).

the lack of ground truth is prevalent in a wide range of domains. In such cases, human annotation is treated as WGT and rater-specific distortion tendencies (or biases) are to be expected. Because of this, an important aspect of determining WGT is assessing reliability and agreement among raters in order to provide a measure of homogeneity and consensus in their ratings. However, traditional reliability statistics do not properly address the disagreement originating from rater bias. The challenge of WGT determination is even more pressing for continuous data as these introduce several types of biases, including rater-specific delays in annotation (time-shifting bias), rater mean bias (where different raters annotate

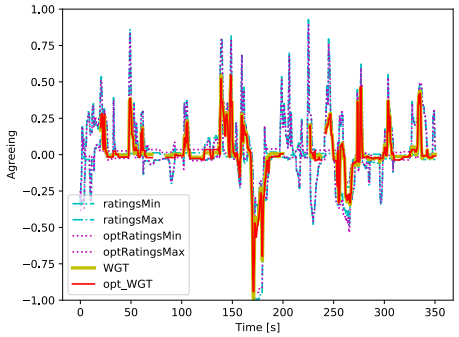


FIGURE 11. SEMAINE dataset: An example of the optimization effect with a minimal reduction of the min-max interval size. The figure shows the non-optimized (cyan line) and the optimized (magenta line) min-max ratings as well as WGT data size before optimization (yellow line) and after (red line).

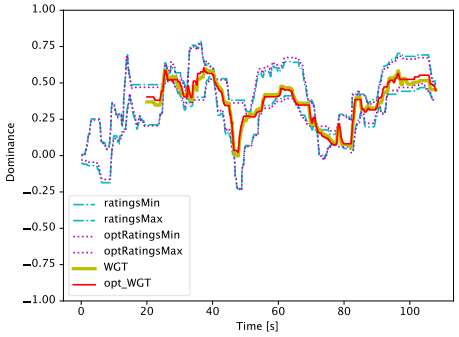


FIGURE 12. CreativeIT dataset: An example of the optimization effect with a minimal reduction of the min-max interval size. The figure shows the non-optimized (cyan line) and the optimized (magenta line) min-max ratings as well as WGT data size before optimization (yellow line) and after (red line).

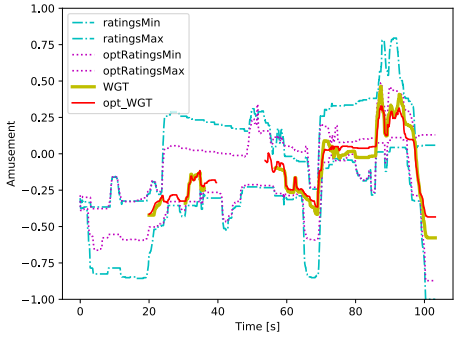


FIGURE 13. SEMAINE dataset: The effect of the optimization on rater disagreement. The comparison is given for WGT data size before optimization (yellow line) and after (red line), and for the min-max intervals before optimization (cyan line) and after (magenta line).

around different means), and scaling bias (the magnitude of deviations from the mean ratings).

As presented in Section II, there are several competitive methods for determining WGT. Standard reliability statistics are based on the averaging of ratings in order to remove disagreement and does not differentiate between true rater disagreement and disagreement due to rater bias. On the other hand, machine learning approaches require large amounts of training data and cannot be effectively applied to smaller datasets. In such cases, alternative approaches, commonly based on using various combinations of truncated mean (or, alternatively, a weighted truncated mean) and correlation metrics, are more appropriate.

It is difficult to conduct a direct comparison of these methods using objective metrics. Since there is no ground truth, it is impossible to objectively compare the quality of WGT generated by these methods, or, for example, the effect size of the bias removal. The two methods that are comparable with our approach differ conceptually (cf. [16], [35]) and do not directly address the removal of rater bias, which is the key advantage of the proposed procedure in terms of improving the quality of WGT.

The proposed procedure determines WGT from multiple raters by combining their ratings about some time-varying quantity into a valid WGT only when sufficient inter-rater agreement has been achieved. This is done using the following parameters: 1) the measurement of the inter-rater agreement with ICC on a global and time-local scale; 2) the removal of rater bias and optimization of global ICC through the (shifting and scaling) transformation of ratings; 3) the time-local determination of valid WGT only where the inter-rater agreement is above the cut-off threshold, and; 4) the optimization of WGT using the weighted truncated mean where higher weights are assigned in a time-local manner to raters that contributed the most to local ICC within a sliding window.

There are several advantages to the proposed WGT determination procedure. It is robust, configurable, and generalizable to any type or size of continuous or discrete numerical data where multiple raters are involved. It utilizes a maximization of the inter-rater agreement in order to improve WGT. It operates on a family of rating transformations aimed at removing the two most common types of rater bias (rater mean bias and scaling bias) and preserves true disagreement among the raters.

The proposed WGT determination procedure was tested on two widely-used behavioral datasets of continuous annotations, the SEMAINE dataset and CreativeIT dataset. The results show that, in most cases, the WGT optimization improved the size of the WGT data, as well as the inter-rater reliability and agreement, by removing rater bias. The average improvement of WGT data size was between 10.1 - 20.9 percentage points for the SEMAINE dataset and 15.2 - 18.7 percentage points for the CreativeIT dataset depending on the dimension. The entropy of residuals also decreased after optimization for both datasets, meaning more relevant information was retained. This improvement was small but consistent. The results also show the procedure is sensitive to the type of rater disagreement – it preserves true rater disagreement and removes disagreement originating from rater bias.

We are making an online tool available to researchers to apply the proposed procedure to their data.[2] The tool removes

[2]https://www.lucami.org/en/WGT/

the two types of rater biases (mean and scaling bias) and generates optimized WGT along with an estimation of inter-rater reliability and agreement.

The drawback of the proposed procedure is that it removes rater mean and scaling bias, but not time-shifting bias. In practice, there are often rater-specific delays in annotation, and rater biases may also drift over time. This might have contributed to the decrease of WGT data size found in a fraction of the CreativeIT samples (3%), where, after WGT optimization, the size of original WGT data was reduced (by a median of -1.4 percentage points). In order to address these issues, our future work will focus on the development of a time-local bias removal procedure that will detect and remove rater biases affected by drift.

REFERENCES

[1] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: A survey," *Int. J. Social Robot.*, vol. 5, no. 2, pp. 291–308, Jan. 2013.

[2] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robot. Auto. Syst.*, vol. 63, pp. 22–35, Jan. 2015.

[3] C. Burr, N. Cristianini, and J. Ladyman, "An analysis of the interaction between intelligent software agents and human users," *Minds Mach.*, vol. 28, no. 4, pp. 735–774, Dec. 2018.

[4] C. Roda and J. Thomas, "Attention aware systems: Theories, applications, and research agenda," *Comput. Hum. Behav.*, vol. 22, no. 4, pp. 557–587, Jul. 2006.

[5] C. Roda and J. Thomas, Eds., "Attention aware systems," *Comput. Hum. Beaviour*, vol. 22, no. 4, 2006.

[6] R. Chua, D. J. Weeks, and D. Goodman, "Perceptual-motor interaction: Some implications for human-computer interaction," in *The Human-Computer Interact. Handbook*, J. A. Jacko and A. Sears, Eds. Trenton, NJ, USA: Lawrence Erlbaum Associates, 2008, pp. 23–34.

[7] C. Peters, G. Castellano, and S. de Freitas, "An exploration of user engagement in HCI," in *Proc. Int. Workshop Affective-Aware Virtual Agents Social Robots - AFFINE*, 2009, pp. 9:1–9:3.

[8] M. Tkalčič, B. De Carolis, M. de Gemmis, A. Odić, and A. Košir, "Emotions personality personalized services," in *Human-Computer Interaction Series*. Amsterdam, The Netherlands: Elsevier, 2016.

[9] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. Face Gesture*, Mar. 2011, pp. 827–834.

[10] M. A. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1299–1311, Jul. 2014.

[11] D. Dupré, N. Andelic, G. Morrison, and G. McKeown, "Dynamic analysis of automatic facial expressions recognition 'in the wild' using generalized additive mixed models and significant zero crossing of the derivatives," *Proc. 32nd Int. BCS Hum. Comput. Interact. Conf., HCI 2018*, 2018.

[12] J. Kossaifi, B. W. Schuller, K. Star, E. Hajiyev, M. Pantic, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, and A. Toisoul, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 1, 2019, doi: 10.1109/TPAMI.2019.2944808.

[13] 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," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan. 2012.

[14] S. N. A. Metallinou, C. Lee, C. Busso, and S. Carnicke, "The USC CreativeIT database: A multimodal database of theatrical improvisation," in *Proc. Multimodal Corpora, Adv. Capturing, Coding Analyzing Multimodality (MMC)*, Valletta, Malta, May 2010, pp. 1–4.

[15] M. Schaekermann, E. Law, A. C. Williams, and W. Callaghan, "Resolvable vs. Irresolvable ambiguity: A new hybrid framework for dealing with uncertain ground truth," in *Proc. 1st Workshop Hum.-Centered Mach. Learn. SIGCHI*, 2016.

[16] M. Nicolaou, H. Gunes, and M. Pantic, "Automatic segmentation of spontaneous data using dimensional labels from multiple coders," in *Proc. Workshop Multimodal Corpora, Adv. Capturing, Coding Analyzing Multimodality*, M. Kipp, J.-C. Martin, P. Paggio, and D. Heylen, Eds. Kaiserslautern, Germany: German Research Center for AI (DFKI), May 2010, pp. 43–48.

[17] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, "Modeling multiple time series annotations as noisy distortions of the ground truth: An expectation-maximization approach," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 76–89, Jan. 2018.

[18] T. Ronchetti, C. Jud, P. M. Maloca, S. Orgül, A. T. Giger, C. Meier, H. P. Scholl, R. K. M. Chun, Q. Liu, C. H. To, B. Považay, and P. C. Cattin, "Statistical framework for validation without ground truth of choroidal thickness changes detection," *PLoS ONE*, vol. 14, no. 6, pp. 1–17, 2019.

[19] P. Smyth, M. C. Burl, U. M. Fayyad, and P. Perona, "Knowledge discovery in large image databases: Dealing with uncertainties in ground truth," in *Proc. 3rd Int. Conf. Knowl. Discovery Data Mining*, 1994, pp. 109–120.

[20] G. Srivastava, J. A. Yoder, J. Park, and A. C. Kak, "Using objective ground-truth labels created by multiple annotators for improved video classification: A comparative study," *Comput. Vis. Image Understand.*, vol. 117, no. 10, pp. 1384–1399, Oct. 2013.

[21] J. Li, S. Ling, J. Wang, and P. Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3339–3347.

[22] J. Kottner, L. Audigé, S. Brorson, A. Donner, B. J. Gajewski, A. Hróbjartsson, C. Roberts, M. Shoukri, and D. L. Streiner, "Guidelines for reporting reliability and agreement studies (GRRAS) were proposed," *J. Clin. Epidemiology*, vol. 64, no. 1, pp. 96–106, Jan. 2011.

[23] M. Banerjee, M. Capozzoli, L. McSweeney, and D. Sinha, "Beyond kappa: A review of interrater agreement measures," *Can. J. Statist.*, vol. 27, no. 1, pp. 3–23, Mar. 1999.

[24] A. von Eye and E. Y. Mun, *Analyzing Rater Agreement: Manifest Variable Methods*. London, U.K.: Psychology Press, 2014, p. 202.

[25] K. L. Gwet, *Handbook Inter-Rater Reliability: Definitive Guide to Measuring Extent Agreement Among Raters*. Gaithersburg, MD, USA: Advanced Analytics, LLC, 2010.

[26] E. Cho, "Making reliability reliable: A systematic approach to reliability coefficients," *Organizational Res. Methods*, vol. 19, no. 4, pp. 651–682, Oct. 2016.

[27] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, p. 37, 1960.

[28] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.," *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968.

[29] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

[30] K. Krippendorff, "Content analysis: An introduction to its methodology," in *Content Analysis: An Introduction to its Methodology*. Newbury Park, CA, USA: Sage, vol. 2013, pp. 18–96.

[31] J. L. Fleiss and J. Cohen, "The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability," *Educ. Psychol. Meas.*, vol. 33, no. 3, pp. 613–619, 1973.

[32] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979.

[33] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Commun. Methods Measures*, vol. 1, no. 1, pp. 77–89, Apr. 2007.

[34] J. J. Bartko, "The intraclass correlation coefficient as a measure of reliability," *Psychol. Rep.*, vol. 19, no. 1, pp. 3–11, Aug. 1966.

[35] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations," *Lang. Resour. Eval.*, vol. 50, no. 3, pp. 497–521, Sep. 2016.

[36] C. Hedge, G. Powell, and P. Sumner, "The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences," *Behav. Res. Methods*, vol. 50, no. 3, pp. 1166–1186, Jun. 2018.

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

[38] L. Aroyo and C. Welty, "Truth is a lie: Crowd truth and the seven myths of human annotation," *AI Mag.*, vol. 36, no. 1, pp. 15–24, Mar. 2015.

[39] 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. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 769–783, Apr. 2013.
[40] G. J. McKeown and I. Sneddon, "Modeling continuous self-report measures of perceived emotion using generalized additive mixed models," *Psychol. Methods*, vol. 19, no. 1, pp. 155–174, 2014.
[41] A. Bender, A. Groll, and F. Scheipl, "A generalized additive model approach to time-to-event analysis," *Stat. Model.*, vol. 18, nos. 3–4, pp. 299–321, Jun. 2018.
[42] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," 2016, *arXiv:1609.08675*. [Online]. Available: http://arxiv.org/abs/1609.08675
[43] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.
[44] P. M. Muller, S. Amin, P. Verma, M. Andriluka, and A. Bulling, "Emotion recognition from embedded bodily expressions and speech during dyadic interactions," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 663–669.
[45] C.-M. Chang, B.-H. Su, S.-C. Lin, J.-L. Li, and C.-C. Lee, "A bootstrapped multi-view weighted kernel fusion framework for cross-corpus integration of multimodal emotion recognition," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 377–382.
[46] M. Atcheson, V. Sethu, and J. Epps, "Using Gaussian processes with LSTM neural networks to predict continuous-time, dimensional emotion in ambiguous speech," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 718–724.
[47] Z. Huang and J. Epps, "An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech," *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 653–668, Oct. 2020.
[48] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image Vis. Comput.*, vol. 65, pp. 23–36, Sep. 2017.
[49] 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 *Proc. ISCA Tutorial Res. Workshop (ITRW) Speech Emotion*, 2000, pp. 1–6.
[50] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
[51] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant.," *J. Personality Social Psychol.*, vol. 76, no. 5, pp. 805–819, 1999.
[52] M. Stolarova, C. Wolf, T. Rinker, and A. Brielmann, "How to assess and compare inter-rater reliability, agreement and correlation of ratings: An exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs," *Frontiers Psychol.*, vol. 5, p. 509, Jun. 2014.

ANDREJ KOŠIR (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Ljubljana, in 1999. Since 2014, he has been a Full Professor with the Faculty of Electrical Engineering, University of Ljubljana, and has also been the Head of the User-adapted Communications and Ambient Intelligence Lab. His research interests include user modeling and personalization (user models and recommender systems), user interfaces (machine learning method design and statistical analysis), and social signal processing.

GREGOR STRLE received the B.A. degree in philosophy and the M.A. degree in information science from the University of Ljubljana, in 2002 and 2008, respectively, and the Ph.D. degree in cognitive sciences from the University of Nova Gorica, in 2012. He is currently a Research Fellow with the Faculty of Electrical Engineering, University of Ljubljana, also with the Scientific Research Centre of Slovenian Academy of Sciences and Arts, and also a Research Member of the User-adapted Communications and Ambient Intelligence Lab. His research interests include cognition, social intelligence, affective computing, and human–computer interaction.

MARKO MEŽA (Senior Member, IEEE) received the B.Sc and Ph.D. degrees in electrical engineering from the University of Ljubljana, in 2001 and 2007, respectively. Since 2013, he has been an Assistant Professor with the Faculty of Electrical Engineering, University of Ljubljana, and also a Research Member of the User-adapted Communications and Ambient Intelligence Lab. His research interests include medical and social signal processing using machine learning and datamining.