Fifty Shades of Green: Towards a Robust Measure of Inter-annotator Agreement for Continuous Signals

Brandon M. Booth
brandon.m.booth@gmail.com
Signal Analysis and Interpretation Lab, USC
Los Angeles, CA

Shrikanth S. Narayanan
shri@ee.usc.edu
Signal Analysis and Interpretation Lab, USC
Los Angeles, CA

ABSTRACT

Continuous human annotations of complex human experiences are essential for enabling psychological and machine-learned inquiry into the human mind, but establishing a reliable set of annotations for analysis and ground truth generation is difficult. Measures of consensus or agreement are often used to establish the reliability of a collection of annotations and thereby purport their suitability for further research and analysis. This work examines many of the commonly used agreement metrics for continuous-scale and continuous-time human annotations and demonstrates their shortcomings, especially in measuring agreement in general annotation shape and structure. Annotation quality is carefully examined in a controlled study where the true target signal is known and evidence is presented suggesting that annotators' perceptual distortions can be modeled using monotonic functions. A novel measure of agreement is proposed which is agnostic to these perceptual differences between annotators and provides unique information when assessing agreement. We illustrate how this measure complements existing agreement metrics and can serve as a tool for curating a reliable collection of human annotations based on differential consensus.

CCS CONCEPTS

• **Human-centered computing** → *User models.*

KEYWORDS

human annotation; continuous ratings; intercoder agreement

ACM Reference Format:
Brandon M. Booth and Shrikanth S. Narayanan. 2020. Fifty Shades of Green: Towards a Robust Measure of Inter-annotator Agreement for Continuous Signals. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3382507.3418860

1 INTRODUCTION

Continuous human annotations have been widely employed in human behavioral machine intelligence tasks as a means to obtain a ground truth representation of human experience of a given stimulus. As online annotation tools [17] and methods for generating consistent ground truth representations from human annotations [3, 5] have recently been developed, continuous human annotation is becoming an accessible research tool for understanding the dynamics of human experiences and judgments. The affective computing community, in particular, is making use of these types of annotations to produce data sets of annotated human experience and expressed emotions. Some examples include databases for expressed emotion prediction [6, 11, 16, 20, 23, 24], movie emotion portrayal [13], and student engagement assessment [4], among other areas.

Assessing reliability of the annotations, or the soundness and reproducibility of an annotation process across studies [2], is a key step during analysis. For subjective constructs where no notion of the true dynamics exists, reliability is often presumed when the annotations are collected independently and exhibit sufficient agreement [12], for example using the Cohen's κ measure. Agreement measures are also employed for identifying and removing outlying or potentially adversarial annotations when researchers are interested in the majority consensus about a construct's dynamics. Quite often, the same agreement measurement tools are used for annotation selection and reliability assessment, and when annotations are selected for analysis based on their consensus, a selection bias is introduced that may reduce the generalizability of the results to other studies. Sometimes similar measures are used to assess the validity of a collection of annotations, which is a measure of the representativeness of the annotations to the true construct dynamics. This work will focus on inter-rater agreement measures as a tool for finding consensus among a subset of human-produced interval-scale continuous annotations for the purposes of annotation curation prior to analysis. One of our aims is to highlight shortcomings of common existing agreement metrics and to present a new agreement measure, based on differential analysis, which offers unique complementary information that is potentially beneficial for curation.

Various works have observed that errors in human annotation are not random [5, 18, 27]. Booth et al. [5] include a brief list of observed common mistakes made by annotators in a controlled interval-scale continuous annotation experiment, including one observation that annotators capture trends more reliably than they accurately assign values. This statement is partially supported by philosophical and psychological perspectives on the fundamental nature of human experience [18, 27]. Yannakakis et al. [25] provide a summary and exposition of research and arguments in favor of the underlying ordinal character of emotional experience and how it may impact human-produced continuous annotations. Some researchers [5, 18, 26] have found it beneficial to treat interval annotations as ordinal ones, but so far the arguments for doing so have been its utility and potential correspondence with the underpinnings of human experience. This work aims to help build a foundation for this

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418860

assumption through the analysis of interval-scale continuous human annotations in an experiment where the true target signal is known[1].

The unique contributions in this work are:

(1) Analysis of the accuracy of interval-scale continuous annotation output and its differentials
(2) Manifested evidence that annotator perception is modeled by a noisy monotonic distortion function
(3) Analysis of existing agreement measures and proposal of a new metric offering complementary information

2 BACKGROUND

A variety of techniques have been used to assess agreement among interval-scale continuous human annotations. Several works [1, 6, 16, 18] use Cronbach's α to assess inter-annotator agreement, and the authors of [1] employ it in a crowd-sourced annotation scenario to remove annotations which disagree with the majority cluster. Other databases for emotion and affect recognition [10, 18, 24, 28] use Pearson correlation along with different thresholding schemes to remove low agreement annotations and try to assess reliability. Yannakakis et al. employ Krippendorff's α to measure inter-annotator agreement for both ordinal and interval human emotion labels [26]. Cohen's κ coefficient thresholding is used in [8] to form a consensus among annotators and discard outliers. Nicolaou et al. introduce a metric called signed agreement (SAGR) for assessing the accordance of emotional valence annotations in the SAL-DB database [20]. The RECOLA data set employs mean squared error (MSE), Pearson correlation, Cronbach's α, and Cohen's κ for assessing inter-annotator agreement but does not use them to exclude outliers [23].

We include each of these methods in our analysis in Section 4, and we also include Kendall's τ and Spearman's correlation to examine rank-based measures. Though correlation-based methods should not be used to assess inter-annotator agreement [12], they continue to see use in practice. To the best of our knowledge, this is the first work closely examining annotator behavior in interval (i.e., value-based) continuous-time annotation tasks in order to evaluate existing agreement measures. Development of ordinal interpretations of annotations have been conducted [18, 25, 26], but not for the purpose of establishing a foundation for measuring agreement.

3 CHARACTERIZING INTERVAL-SCALE CONTINUOUS HUMAN ANNOTATION

Before we examine the assumptions, benefits, and drawbacks of different agreement measures, we first characterize the intra-annotator variability and noise present in interval-scale continuous human annotations. To avoid making assumptions about the relative correctness of annotations of ambiguous stimuli, we analyze data from [5] where annotators were asked to rate the intensity of varying shades of green in videos. We briefly describe this data set next then present an analysis of annotator accuracy and consistency. Finally, we demonstrate that annotators asked to perform interval (value-based) annotations incidentally provide reliable ordinal information.

[1]Source code for all analysis in this work is available from https://github.com/brandon-m-booth/2020_annotator_agreement

3.1 Experiment: Annotating Shades of Green

In [5], ten annotators were asked to view videos of solid shades of green varying over time and to annotate the intensity of the green color shown in real time. Two videos, of only a few minutes in length each, were annotated separately by each annotator on a bounded [0,1] scale using a computer mouse and slider user interface widget. Annotations were recorded at 10Hz and down-sampled to 1Hz for analysis. Figure 1 plots all annotations in both tasks against the true target signal. The intra-class correlation measure ICC(k) [15] (one-way random, average score), was used to assess the agreement between annotators and produced a 0.97 score, suggesting extremely high agreement. The benefits of drawbacks of this and other metrics are revisited in Section 4. We use the 1Hz annotations and corresponding true green intensity time series for both annotation tasks in our analysis.

3.2 Qualitative Analysis

To facilitate fair comparison between the samples comprising each annotation, we first try to align them using two methods: an evaluator-dependent constant temporal shift (*EvalDep*) [14] and dynamic time warping (DTW) using a symmetric Sakoe-Chiba step pattern to limit the warping distortion [9, 19]. Both methods are effective at correcting the annotations for human lag time, but DTW's allowance for repeated values when locally stretching a signal leads to a distortion in its structure (i.e., derivatives). We proceed using the *EvalDep* method where each annotation is shifted a constant amount to align with the true signal, which averages around 1.6 seconds in both tasks. In other experimental scenarios with unknown true signals, this temporal adjustment could be based on mutual alignment of annotations to each other or to features extracted from the stimulus (see [29] for one example).

Figure 2 shows scatter plots of the true intensity of the green color per frame against the annotated values for each annotator. An idealized annotator with no delay, no perceptual bias, and no incidentally or motorically added noise would produce a scatter plot matching the straight dashed line. It is apparent from the scatter patterns in these figures that each annotator deviates from this ideal in a unique manner. The vertical *spread* of scattered points at any true green intensity value represents the precision with which each target value is annotated. Most annotations show an increased spread in the mid-to-upper range ([0.5-1]) indicating that this range of values is annotated more inconsistently than, for example, values near zero (black color).

The general shape of the scattered points for each annotator forms either an arc or S-curve, so we opt to fit a cubic polynomial to the data. These regressions vary in exact shape and location for each annotator and seem to characterize their individual perceptual biases. Interestingly, though cubic regressions are capable of representing non-monotonic S-curves, all of the fitted cubic regressions lie in their monotonic parametric regime.

Figure 3 shows scatter plots of the forward difference in the true intensity of the green color per frame against the forward difference of the annotated values per annotator. As before, an idealized annotator would have a scatter plot precisely following the dashed line. To help examine whether annotators locally preserve ordinal relationships as mentioned in [5, 17, 24–26], we partition the

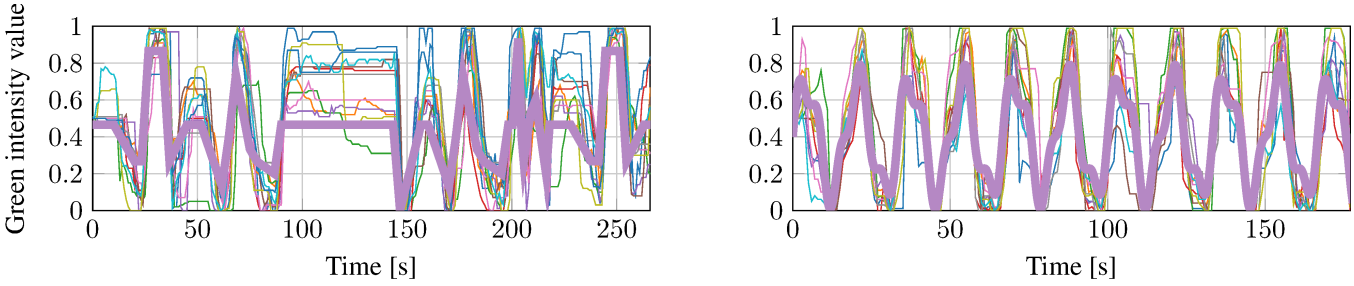
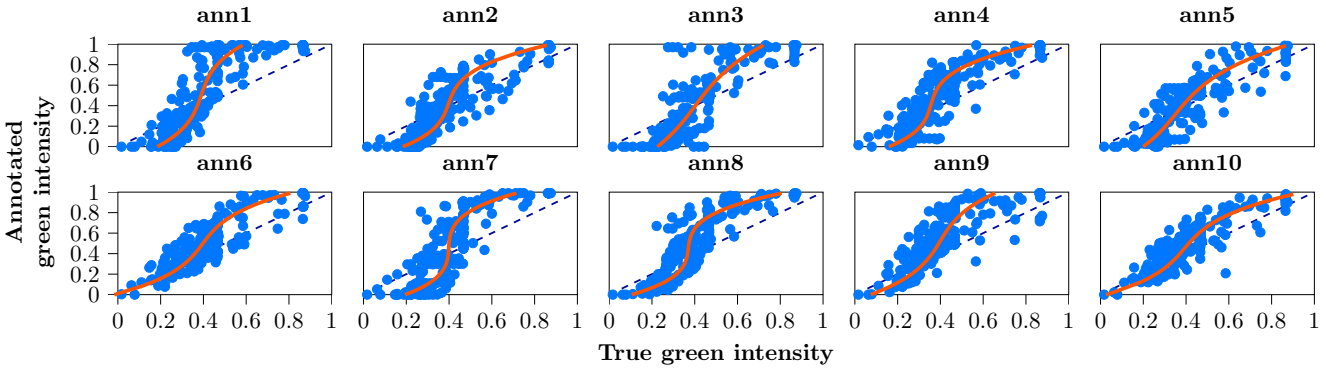
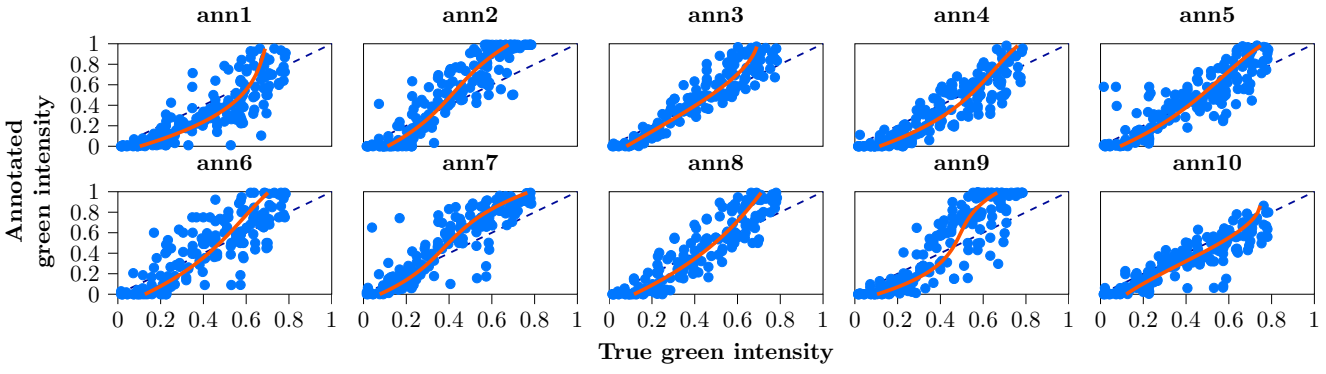


Figure 1: Plots of the true target signal (bold) and real-time interval-scale continuous annotations of green intensity from ten annotators in two separate tasks (left: Task A, right: Task B)



(a) Task A

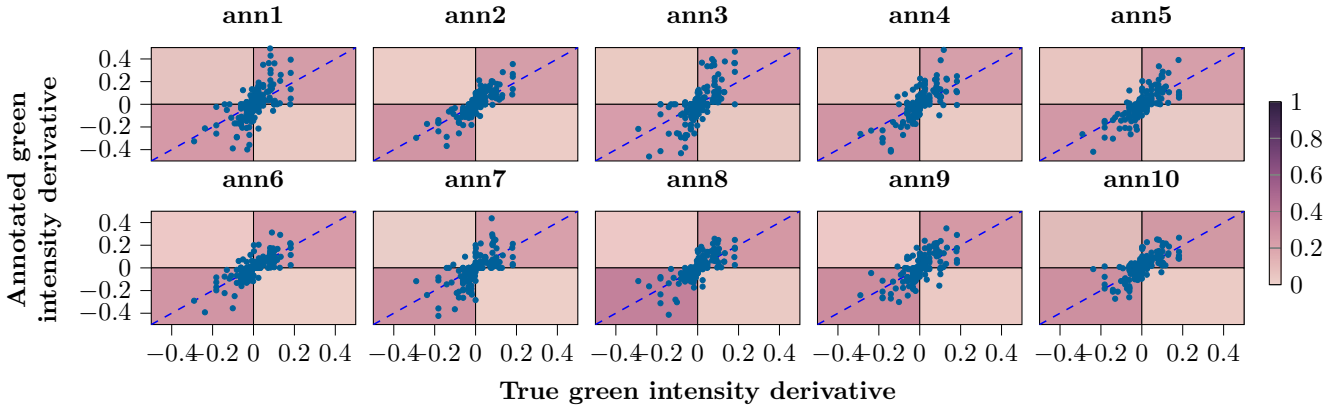


(b) Task B

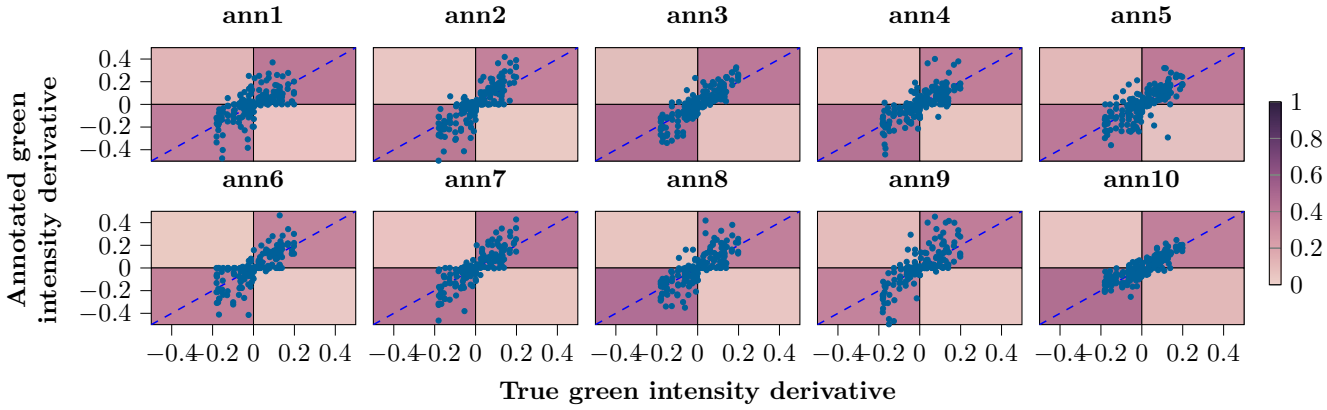
Figure 2: Scatter plots of the true green intensity values against the annotated values for each annotator in both tasks. The dashed line show where the points would lie for a perfect annotator. The orange curves are cubic polynomial regressions of the scattered points.

points about zero to compare frame-wise agreement in trend direction. The plots have been partitioned into quadrants and colored according to the percentage of total points within each quadrant (numbered counter-clockwise starting from the upper right section). Points in quadrants I/III correspond to samples where the true signal derivative increases/decreases and the annotation slope also increases/decreases. Quadrants II and IV contain samples where an annotation's trend differs from the true trend. Together, these quadrants can be viewed as a confusion matrix, and when the points are binned in this manner, it is clear that the trend error rate is relatively low. There does not appear to be a similar partitioning scheme of the scatter plots of annotation values in Figure 2 which can bin the data across all annotators in a meaningful way.

Finally, Figure 4 plots the annotations over time where blue line segments correspond to samples where the forward difference matches the sign of the true target signal's forward difference and the red segments represent samples which do not match. Most of the mismatches occur near transitions where the derivative of the signal



(a) Task A



(b) Task B

Figure 3: Scatter plots of the forward difference of the true green intensity values against the forward difference of annotated values for each annotator in both tasks. The dashed line shows where the points would lie for an idealized annotator. Four quadrants are color coded according to the percentage of each plot's points falling within them.

changes (i.e., near the peaks, valleys, and plateaus). Unless annotator reaction times are perfectly consistent, one would naturally expect a few samples to disagree near these transitions. We claim in Section 5 that an agreement measure should not heavily penalize these types of annotation errors.

In summary, through the value-based and differential frame-wise comparison of annotations in this experiment, we have observed two key aspects of interval-scale continuous human annotation: (1) annotators approximately preserve rank ordering when annotating values (as shown by the monotonicity of the cubic regression), and (2) annotators reliably capture the direction (increasing or decreasing) of trends, but are less accurate in both annotation valuation and trend magnitude.

3.3 Quantitative Analysis

Next we formalize our observations about interval-scale continuous human annotation quality. Let $T(t)$ represent a time series of values corresponding to the true construct target at time t. Let $P_i(z)$ be a unique noisy perceptual distortion function for annotator i for an observed stimulus value z (e.g., a shade of green). Our two observations about human annotator behavior can be formally expressed as follows:

(1) Annotators preserve rank ordering when annotating values:

$$\frac{\mathrm{d}P_i}{\mathrm{d}z} \gtrapprox 0 \tag{1}$$

(2) Annotators capture trend directions reliably:

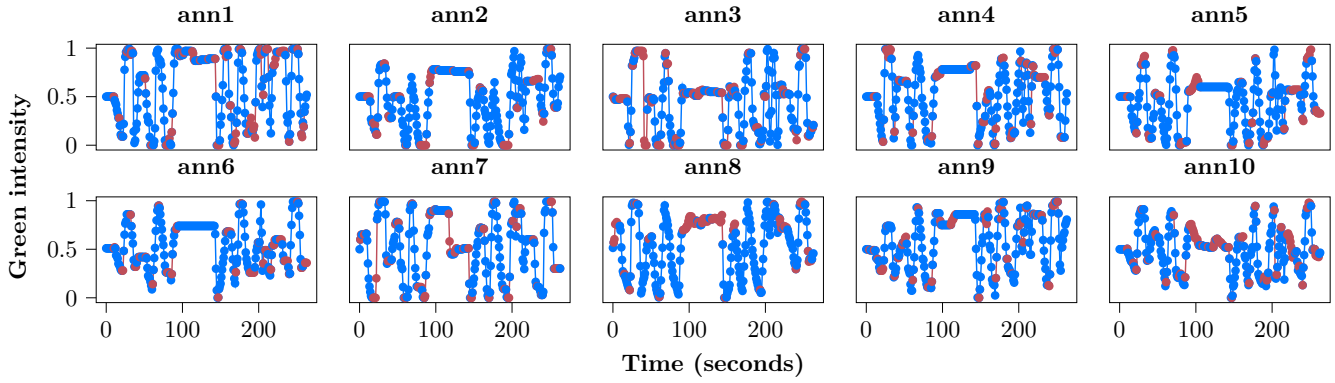
$$\operatorname{sgn}\left[\frac{\mathrm{d}T}{\mathrm{d}t}\right] \approx \operatorname{sgn}\left[\frac{\mathrm{d}}{\mathrm{d}t} P_i[T(t)]\right] \tag{2}$$

Using the chain rule on Equation 2 and notation $T'(t) = \frac{dT}{dt}$, we can rewrite it as:

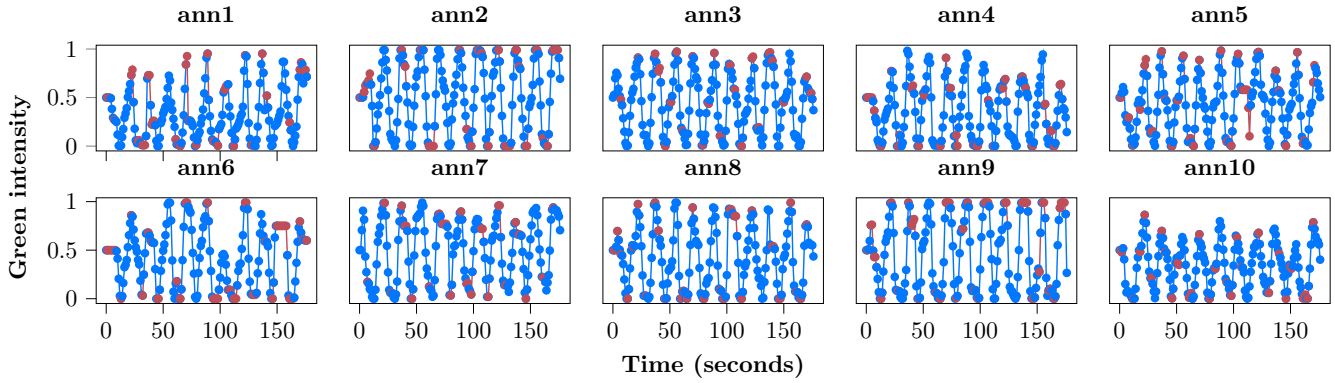
$$\operatorname{sgn}[T'(t)] \approx \operatorname{sgn}\Big[P_i'[T(t)]T'(t)\Big] = \operatorname{sgn}\Big[P_i'[T(t)]\Big] \cdot \operatorname{sgn}[T'(t)]$$

If we add back Equation 1, we deduce that these two terms are equal since $\operatorname{sgn}(P_i'[T(t)]) = 1$, which demonstrates that these two observations are mathematically consistent and complementary.

Each observation also suggests an avenue for assessing agreement. If one could somehow measure the degree to which an annotator preserves monotonicity (Equation 1), it could be utilized as a measure



(a) Task A



(b) Task B

Figure 4: Plots of the aligned annotations over time. Blue points and line segments correspond to samples where the forward difference matches the sign of the true target signal's forward difference. Red points and line segments correspond to mismatches.

of agreement. Unfortunately, $T(t)$ is unknown for most problems of interest. Instead, we can exploit our approximation in Equation 2 for two independent annotators i and j to state:

$$\operatorname{sgn}\Big[P_i'[T(t)]T'(t)\Big] \approx \operatorname{sgn}\Big[T'(t)\Big] \approx \operatorname{sgn}\Big[P_j'[T(t)]T'(t)\Big]$$

$$\implies \operatorname{sgn}\Big[P_i'[T(t)]\Big] \approx \operatorname{sgn}\Big[P_j'[T(t)]\Big]$$

If we further assume that $T(t)$ is a continuous function, we can approximate $P_i'[T(t)]$ using a forward difference:

$$\operatorname{sgn}\Big[P_i[T(t+\Delta)] - P_i[T(t)]\Big] \approx \operatorname{sgn}\Big[P_j[T(t+\Delta)] - P_j[T(t)]\Big] \quad (3)$$

for some small $\Delta > 0$. In practice, Δ will be the distance between two adjacent annotation samples in time, which may not be very small, but the approximation is valid provided that the sampling rate is high enough for the construct dynamics. We exploit this key relationship in Section 4.2 to define a new agreement measure.

4 MEASURING AGREEMENT

In this section, we examine the metrics mentioned in Section 2 that have been used for measuring inter-annotator agreement, and we evaluate the benefits and drawbacks of each one based on our observations about human annotation and errors in the shades of green experiment.

4.1 Analysis of Existing Measures

Pearson correlation is not agnostic to the effects of a monotonic transformation of the sample data. As a simple example, we take the true target signal from TaskA and apply a distortion mimicking one of the "best" annotators. Annotator #7 has the lowest differential error rate on that task with an F1 score of 0.96 and a Matthews correlation coefficient (MCC) of 0.92. We apply the cubic regression from Figure 2a for *ann7* to the true signal, then compute the Pearson correlation with the unwarped true signal, which yields a value of 0.90. Thus, if we assume the cubic regression models this annotator's perceptual bias, the best possible Pearson correlation score that can be achieved is 0.9, even though *ann7* arguably captures the shape and structure of the true annotation better than this (using the F1 score and MCC for rough comparison). A second drawback of this metric is that it requires some variance to be present in the annotation. It cannot be used to measure agreement over subsets of the annotation in time where no variability is present or in situations where a legitimate annotation for a stimulus would contain no fluctuations.

Concordance correlation coefficient (CCC) is defined as:

$$\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2}$$

for annotations x and y. It is based on the Pearson correlation measure and further punishes annotations which have sufficiently different mean values or standard deviations. Again, borrowing the cubic regression model of perceptual bias for *ann7* and comparing the true signal to itself after a simulated distortion, we obtain a CCC value of 0.43. Different perceptual distortions would cause pairwise agreement values obtained using CCC to be difficult to compare.

Both **Spearman correlation** and **Kendall's τ** coefficients measure correlation from the rank order of a pair of annotations. As long as the perceptual distortion function (approximated by a cubic regression in Figure 2) is monotonic, then these metrics will remain unchanged. However, one major drawback is their sensitivity to prolonged errors in ranking. Figure 5 shows two example annotations with very similar structure but a difference in the constant annotated value past 13s. This minor difference substantially affects the rankings and the Spearman and Kendall's τ correlations of these two signals are -0.119 and -0.194 respectively.

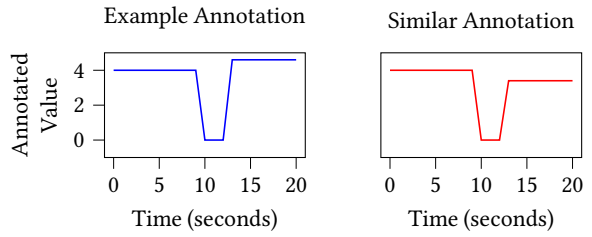


Figure 5: Example annotations showing similar structure but negative Spearman and Kendall's τ correlations.

Mean squared error (MSE) measures metric deviation between two annotations and thus is naturally impacted by any monotonic perceptual distortion. Any differences in annotator valuations are further magnified by the square of the error, meaning this metric is highly unforgiving of value differences in spite of structural similarity.

Signed agreement (SAGR) is a unique measure of similarity initially proposed by Nicolaou et al. [20] for assessing the correspondence of dimensional emotional valence annotations. The measure is intended for signals on a [-1,1] scale where the origin is a conceptual reference point for the annotations (e.g., a neutral expression with no positive or negative valence). This measure discards the magnitude of the value assigned to each frame and captures how often the annotators agree on the positive or negative nature of an emotional expression in each frame. For constructs where there exists a natural mid-point, this type of agreement measure over the annotated values is interesting and warrants further investigation. However, for many annotation tasks, such as the intensity of the shade of green or emotional arousal, there is no middle reference point and this metric is not interpretable or directly applicable.

Cohen's κ is computed by $\kappa = 1 - \frac{1-p_0}{1-p_e}$ where p_0 is the observed agreement among all annotators and p_e is a measure of the random chance of agreement. In order to estimate the probabilities of agreement between interval-scale continuous annotations, probabilistic models of annotators must be assumed or estimated empirically (e.g., via binning). Either approach requires a supposition about the annotation process which may be difficult to validate in small-scale experiments. The correction this metric utilizes, however, to measure the relative agreement between two annotations above random chance is useful and we revisit it in Section 4.2.

Cronbach's α, **intra-class correlation coefficient** (ICC), and **Krippendorff's α** are group-level metrics which are not intended to be applied to subsets of the annotations (such as pairs). These metrics are useful for measuring the agreement amongst entire collections of annotations, but are less useful for identifying adversarial ones (e.g., from crowd sourcing) during curation.

4.2 Signed Differential Agreement

Based on our observations about annotator behavior in Section 3.2, we propose that an ideal measure of continuous human annotation agreement for data set reliability assessment and curation purposes should be invariant to unique perceptual biases, and it should focus on capturing consensus in signal structure and shape rather than valuation. In order to achieve this, we propose an ordinal agreement measure called *Signed Differential Agreement* (SDA) for measuring agreement over construct dynamics in ordinal-scale continuous annotations and which we claim also functions for interval-scale annotations because of the high degree of ordinal consensus shown in Figure 3 and captured by Equation 3.

For notational brevity, let $x,y \in \mathbb{R}^N$ represent the two annotation time series $P_i[T(t)]$ and $P_j[T(t)]$ respectively where $t \in \{1,...,N\}$. Equation 3 becomes:

$$\text{sgn}(x_t - x_{t-1}) \approx \text{sgn}(y_t - y_{t-1})$$

At this point, we could define an agreement measure that is agnostic to human perceptual effects using any loss function with this equation. We elect for a sum-of-delta approach and define SDA as:

$$\text{SDA} = \frac{1}{N-1}\sum_{t=2}^{N}\delta[\text{sgn}(x_t - x_{t-1}),\text{sgn}(y_t - y_{t-1})]$$
$$\delta(p,q) = \begin{cases} 1 & p = q \\ -1 & p \neq q \end{cases} \quad (4)$$

This measure gives an equal weight to every annotation sample and thereby avoids heavily penalizing short periods of disagreement. This function's range is [-1,1] like many other agreement measures and also similarly interpretable. A value of 1 indicates complete agreement, -1 indicates complete disagreement, and 0 means the two annotations are mutually uncorrelated.

Many human annotation data sets contain unaligned annotations due to annotator lag or varying temporal resolution due to asynchronous input. If resampling is infeasible, [26] suggests that aggregation over small (three to five-second) windows may be appropriate. In this scenario, majority voting could be used to determine the prevailing sign of the differentials within each time window before applying the delta function when computing SDA.

Additionally, with this agreement defined as a similarity measure over annotation samples, it possible to add a correction for chance

agreement for cross-corpora comparison [7], similar to Cohen's Kappa measure:

$$\kappa_{\text{SDA}} = 1 - \frac{1-p_o}{1-p_e}$$

where the observed probability of agreement p_o is measured using SDA with its δ replaced by a Kronecker delta function (δ_K) and where the probability of chance agreement is approximated empirically:

$$p_e = \frac{1}{(N-1)^2} \sum_{i \in \{-1,0,1\}} \left[\sum_{t=2}^{N} \delta_K[\text{sgn}(x_t - x_{t-1}), i] \cdot \sum_{t'=2}^{N} \delta_K[\text{sgn}(y_{t'} - y_{t'-1}), i] \right]$$

This is one possible formulation of a chance-corrected SDA measure, with other formulations involving alternative estimation algorithms for p_o and p_e. As we indicate in Section 4.1 for some of the existing agreement measures, this corrected SDA formula is undefined for signals with no variability. Therefore, this style of chance correction may be undesirable in domains where annotations have a fair likelihood of being completely constant.

5 AGREEMENT MEASURE COMPARISON

To demonstrate the advantages of using the proposed SDA agreement metric, we employ it in a few examples cases. First, we use it in a simulated annotation setup and demonstrate that SDA captures structural agreement while other existing measures do not. Second, we apply SDA to real human annotations to show that the metric offers complementary information in practice and that it can be used to help identify potentially poor or adversarial annotations.

5.1 Simulated Task Comparison

Figure 6 shows two hypothetical annotations of the same stimulus.

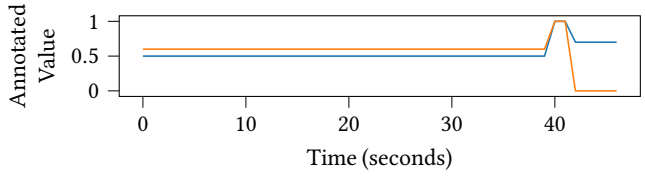


Figure 6: Hypothetical annotations of the same stimulus.

In both cases, these annotations capture a *spike* event near 40s followed by a decrease in value. The primary difference between the two is the rate of decrease near the end. In a real human annotation scenario, this disparity is realistic and could be due to perceptual differences or more simply an input error during the annotation recording process. Either way, this difference, which accounts for less than 10% of the total annotation duration, has a profound impact on existing agreement measures. Table 1 shows pairwise and group-level agreement measures for this example. SAGR is excluded because there is no presumed conceptual reference at the midpoint, and Cohen's κ is computed using 10 equally-spaced bins.

Table 1: Pairwise and group-level agreement measures applied to the simulated annotation example

Agreement Measure	Simulated Task Value
Pearson	-0.10
Spearman	-0.36
Kendall's τ	-0.38
CCC	-0.08
Cohen's κ	0.61
SDA	**1.0**
Cronbach's α	-0.18
ICC(k)	-0.15
Krippendorff's α	-0.50

We contend that these annotations are sufficiently similar, and if they were part of a larger collection of annotations containing outliers, they should be deemed equally valid (or invalid) during annotation curation. Though a contrived example, these annotations are representative of the types of human errors observed in [5] and the results indicate that SDA is the only measure that outputs a number close to 1.0, virtually guaranteeing that the two annotations would be treated equally (e.g., by a clustering algorithm). The large disparity in agreement measures also confirms our prior analysis concluding that other metrics do not easily or directly capture the structural similarity in annotations.

5.2 Comparison using Human Annotations

Next we compare SDA to existing measures in real human annotation experiments, starting with the shades of green study. Figure 7 shows the Pearson correlation and κ_{SDA} pairwise agreements for the temporally aligned task A annotations. Notably, κ_{SDA} yields a substantially smaller value for unique pairs of annotations compared to its perfect self-similarity on the main diagonal. Since this metric is sensitive to temporal alignment and annotator lag, even with lag correction, it captures short durations of disagreement when the trend changes (refer to Figure 4). What is most important about the κ_{SDA} measure in this case is the uniformity of the values for unique annotators rather than the exact agreement value [2, 21], which indicates a similar level of structural agreement for all annotations.

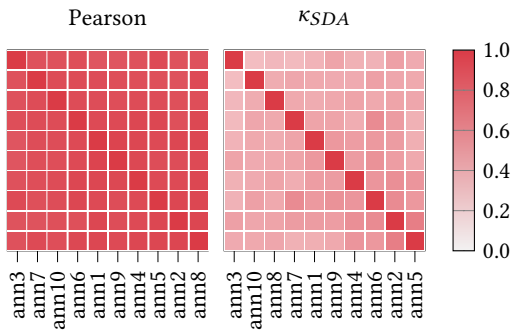
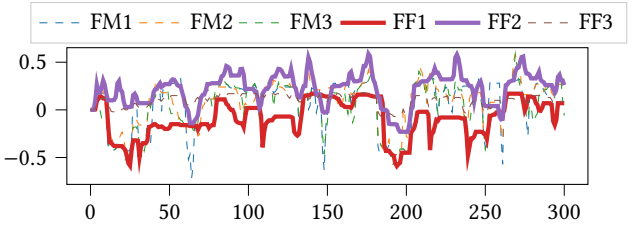


Figure 7: Pairwise agreement measures for task A in the shades of green experiment; rows permuted via agglomerative clustering.

To show how SDA performs on other real data, we examine annotations from the RECOLA data set [23]. RECOLA contains interval-scale continuous human annotations of spontaneous dyadic affective interactions between French speakers, and it has been highly utilized by researchers, in particular for semi-annual emotion prediction challenges (AVEC) [22]. Each interaction is rated continuously by a fixed set of six annotators for emotional arousal and valence on a bounded $[-1,1]$ scale. The authors employ Pearson correlation, MSE, and the percentage of positive samples (similar to SAGR) to assess annotator agreement. For reasons mentioned in Section 4, we exclude MSE and focus our analysis of the benefits of SDA in relation to the other types of agreement. Because SDA is particularly sensitive to misalignment, we apply DTW to temporally warp the annotations using a symmetric Sakoe-Chiba step pattern to constrain the maximum temporal distortion to 3 seconds.

Figure 8a shows pairwise agreement matrices for Pearson, SAGR, and κ_{SDA} across all six annotators for their annotations of emotional arousal in one dyadic interaction from the data set. The corresponding annotations are plotted in Figure 8b with the two traces from annotators FF1 and FF2 drawn in solid bold. The magnitude and uniformity of both the Pearson and SAGR agreement matrices suggests that the annotations co-vary similarly. The κ_{SDA} measure, however, highlights more subtle differences between the annotations and indicates that annotations FF1 and FF2 are structurally uncorrelated. This is apparent from the two annotations shown in Figure 8b, which agree on the trend in arousal sometimes, but disagree about half of the time. This example demonstrates that SDA offers unique information when measuring human annotation agreement and can be helpful when curating data sets.

(a) Pairwise agreement measures in matrix form; rows permuted via agglomerative clustering.



(b) DTW-shifted annotations.

Figure 8: Example annotations of emotional arousal from the RECOLA data set. Annotators FF1 and FF2 (bold solid lines) structurally disagree about half the time.

6 DISCUSSION AND FUTURE WORK

The examples and analysis in Section 5 show that the proposed *Signed Differential Agreement* and chance-corrected κ_{SDA} measures yield interpretable and unique information about human annotation agreement. When combined with other agreement metrics, this additional information can help mitigate the risk of improper inclusion of adversarial annotations or exclusion of sensible ones. Thus, we believe κ_{SDA}, or other measures derived from the approximate equality of signed derivatives across annotators (Equation 3), are useful tools offering supplemental information to facilitate annotation selection. Since SDA effectively ignores some of the structured biases present in independent annotation (e.g., overshooting values as observed in [5]), it may also be an effective reliability measure for reproducibility and machine learning purposes [21].

The results pertaining to the shades of green experiment support the notion that annotators incidentally output reliable ordinal annotations when asked to perform interval annotations while also providing approximately rank-consistent valuations. This suggests that there is more valuable information to be obtained from interval annotations than simply an ordinal interpretation. In part, this idea conflicts with views expressed by Yannakakis et al. [25], suggesting that both interval and ordinal reliability suffer during interval annotation tasks due to cognitive loading. Admittedly, the shades of green experiment is cognitively less demanding compared to other tasks probing expressed human states and traits, for example, annotation of emotional expressions, which may explain why both interpretations are meaningful. A deeper investigation of the impact of cognitive loading on annotation accuracy in both interval and ordinal annotation scenarios is an interesting subject for further research.

Other items for future research include a deeper examination of cross-subject differences in the S-curve distribution of annotation samples from Figure 2. The unique shapes of these noisy monotonic regressions may aid in unmasking and clustering differences in perceptual patterns between individual annotators. Developing alternative temporal alignment strategies for human annotations may also prove beneficial. DTWs allowance for repeated values adds noise to the SDA agreement measure, so a fast DTW-like alternative which preserves gradient directions would be preferable.

7 CONCLUSION

We closely examine interval-scale continuous human annotations in an experiment where a true target signal is known. Evidence suggests that annotator perception can be modeled as a noisy monotonic distortion function. We propose *Signed Differential Agreement*, an agreement measure exploiting the monotonicity of annotators' perceptual biases, and we show that it provides additional insights, in both simulated and real human annotation experiments, when measuring consensus to establish reliability. We conclude that SDA is a tool to consider alongside existing agreement measures when curating continuous human annotation data sets.

REFERENCES

[1] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2017. Developing a benchmark for emotional analysis of music. *PloS one* 12, 3 (2017), e0173392.

[2] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.

[3] Brandon Booth and Shrikanth Narayanan. 2019. Trapezoidal Segmented Regression: A Novel Continuous-scale Real-time Annotation Approximation

Algorithm. In *In proceedings of Proceedings of the 8th International Conference on Affective Computing Intelligent Interaction* (Cambridge, UK).

[4] Brandon M Booth, Asem M Ali, Shrikanth S Narayanan, Ian Bennett, and Aly A Farag. 2017. Toward active and unobtrusive engagement assessment of distance learners. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 470–476.

[5] Brandon M Booth, Karel Mundnich, and Shrikanth S Narayanan. 2018. A novel method for human bias correction of continuous-time annotations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3091–3095.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[7] Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *arXiv preprint cmp-lg/9602004* (1996).

[8] Laurence Devillers, Roddy Cowie, Jean-Claude Martin, Ellen Douglas-Cowie, Sarkis Abrilian, and Margaret McRorie. 2006. Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches.. In *LREC*. 1105–1110.

[9] Toni Giorgino et al. 2009. Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software* 31, 7 (2009), 1–24.

[10] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 65 (2017), 23–36.

[11] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Bjorn Schuller, Kam Star, et al. 2019. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *arXiv preprint arXiv:1901.02839* (2019).

[12] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research* 30, 3 (2004), 411–433.

[13] Nikos Malandrakis, Alexandros Potamianos, Georgios Evangelopoulos, and Athanasia Zlatintsi. 2011. A supervised approach to movie emotion tracking. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2376–2379.

[14] Soroosh Mariooryad and Carlos Busso. 2015. Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators. *IEEE Transactions on Affective Computing* 6, 2 (2015), 97–108.

[15] Kenneth O McGraw and Seok P Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods* 1, 1 (1996), 30.

[16] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, 1 (2011), 5–17.

[17] David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. PAGAN: Video Affect Annotation Made Easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 130–136.

[18] Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. IEEE, 1–8.

[19] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

[20] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2010. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In *Proc. of LREC Int. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*. Citeseer, 43–48.

[21] Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics* 34, 3 (2008), 319–326.

[22] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. 2018. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*. 3–13.

[23] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

[24] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*. 3–10.

[25] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. 2018. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing* (2018).

[26] Georgios N Yannakakis and Hector P Martinez. 2015. Grounding truth via ordinal annotation. In *2015 international conference on affective computing and intelligent interaction (ACII)*. IEEE, 574–580.

[27] Georgios N Yannakakis and Héctor P Martínez. 2015. Ratings are overrated! *Frontiers in ICT* 2 (2015), 13.

[28] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. 2017. Aff-wild: Valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 34–41.

[29] Feng Zhou and Fernando De la Torre. 2016. Generalized canonical time warping. *IEEE transactions on pattern analysis and machine intelligence* 38, 2 (2016), 279–294.