Continuous real-time annotation fusion correction via rank-based spatial warping

Brandon M. Booth[1], Karel Mundnich[1], Shrikanth S. Narayanan[1]

1 Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA

Abstract

Human annotations are noisy and prone to influence from several factors including personal bias, task ambiguity, environmental distractions, health state, and more. These annotations, however, are of integral value in human behavior studies, and in design and evaluation of machine learning applications, especially those involving hidden mental states that cannot effectively be measured or assessed by other means. We propose a novel method for extending continuous real-time annotation fusion approaches to generate accurate ground truth estimates. We validate our approach in a mechanically simple but perceptually demanding psychophysical annotation experiment where an objective truth is known. Our method yields a ground truth in better agreement with the objective truth than state-of-the-art approaches and can be used to provide a more accurate fused annotation for real data.

1 Introduction

Automated human behavior estimation and prediction problems for hidden state constructs such as emotional state, engagement, productivity, or attention are notoriously difficult and typically approached using supervised machine learning. For these types of problems, self or expert annotations are often used to provide ratings for the target construct and establish a ground truth set of labels for machine learning. However, the human annotation process is noisy and produces several types of label artifacts due to many factors such as perception bias, interpretation ambiguity, and distractions, to name a few. The impact of these biases and cognitive effects on the annotations are magnified when the annotation task demands careful attention or vigilance over long sessions. A time-economical annotation protocol is therefore paramount in order to obtain labels of the target construct that are consistent and congruent. The usual strategy for combating these error sources involves gathering multiple annotations from different sources and fusing them to obtain a single ground truth. Aggregation of multiple annotations helps mitigate the effect of noise and a careful fusion method ideally suppresses individual biases and artifacts, but there is yet no consensus on a best-practice fusion approach.

In this work, we focus on a continuous-scale annotation scheme where expert annotators observe recordings of an experiment session and annotate the target construct in real-time. This annotation scheme has been used for mental constructs that evolve steadily over time like affective dimensions (arousal, valence, dominance) [1–5], and could be used for other mental states such as engagement, attention, productivity, or more generally whenever self-reports disrupt the natural flow of the target construct or introduce retroactive bias. Continuous-scale real-time annotation affords extra

freedom to the experts for subtle distinctions to be made while taking the full temporal context into account, but also is subject to further sources of noise and artifacts.

Prior work on continuous annotation has focused on ground truth estimation by modeling and removing general sources of lag, noise, and/or artifacts. One approach from [6] finds an optimal time shift for separate annotations to align them before fusion via per-frame averaging. This method corrects for variance in annotators' response times, but may perform poorly with adversarial annotations or changes in reaction lag over time (e.g. long annotation tasks). Dynamic time warping [7] is a well-known alignment solution that maximizes the agreement between annotators by handling variance in individual annotators' lag times, but also only corrects for temporal misalignments during fusion. Long short-term memory networks (LSTMs) have also been proposed recently that correct for asynchronous annotator lag when fusing annotations using additional contextual information [8]. This approach seems to elegantly handle lag differences between annotators when fusing, but has only been shown to achieve good objective truth approximations when the contextual representation is appreciably informative of the underlying construct.

Furthermore, canonical correlation analysis (CCA) [9] and correlated spaces regression (CSR) [10] focus on correcting systemic and consistent personal annotation biases by learning a projection function for a set of features that maximizes the projected features' correlation with the set of annotations. These two approaches correct for spatial annotation biases and require a separate set of features to be extracted from the stimulus that are in some combination representative of the changes in annotation signal values. Extensions of this tenet such as canonical time warping (CTW) [11], deep canonical correlation analysis (DCCA) [12], dynamic probabilistic canonical correlation analysis and time warping (DPCCA/DPCTW) [13], deep canonical time warping (DCTW) [14], and generalized canonical time warping (GCTW) [15] have also been studied and shown to perform well on various data sets. With the right set of features these methods can produce desirable fusions but often in human behavior studies these ideal features are as elusive as the estimation problems for which they are used, especially when annotation artifacts do not correlate with directly observable features of the stimulus.

Other methods use a shared latent state to model the ideal annotation signal and separate distortion states to learn annotators' biases and artifacts [16–18]. For analytical convenience and to reduce sample complexity, these Bayesian network models assume exponential family priors for spatial distortion modeling, which are not necessarily reflective of the types of errors expert annotators make. As volunteer-based crowd sourcing platforms such as Amazon's Mechanical Turk become more popular choices for large-scale annotation tasks, it becomes more difficult to model and fit a family of distortions to each annotator with confidence.

The fundamental problem confounding annotation fusion methods to date is the reliance on the average quality of continuous annotations. Several studies have shown that people are better at comparative ranking than absolute rating [3, 19, 20] suggesting that absolute continuous annotations may not exhibit coherence and self-consistency. In this paper we affirm this idea by presenting the results from simple continuous annotation experiments showing high levels of agreement between annotators and a lack of consistency in labeling over time. We then present a post-annotation correction method where additional relative rank information about interval subsets of the fused annotations are used to warp the result to better approximate the objective truths. Rather than focusing on evaluating performance in latent state experiments, we validate our approach on experiments that are reflective of the types of perceptual problems that occur in hidden state annotation tasks, but where objective truths are known *a priori*. We further the utility of our approach by testing the robustness of our technique to
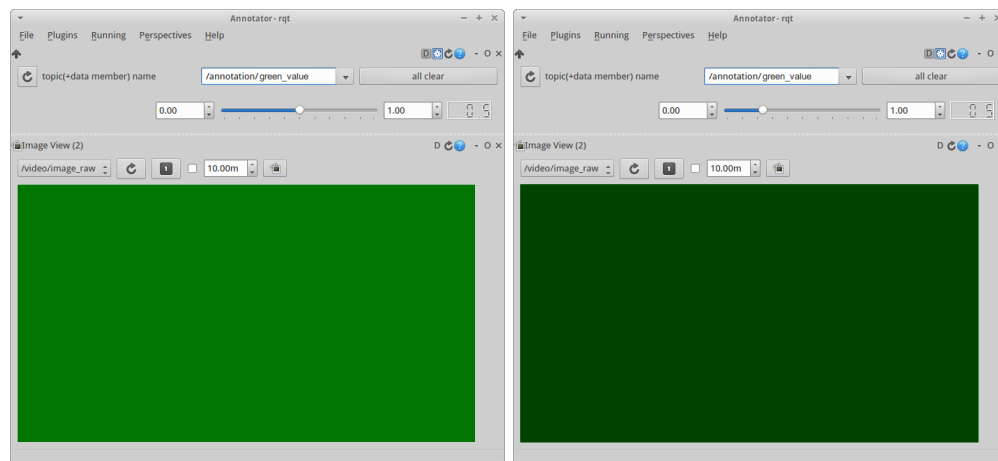
Fig 1. Snapshot of the user interface at different times during the green channel value annotation task. Annotators only adjusted the slider in sync with changes in the green video.

noisy and incomplete additional information. This method is complementary to other fusion procedures and can be used to obtain better ground truth approximations.

2 Experiment

We used a simple but perceptually challenging annotation task where the objective truth was known for our study. Ten annotators were asked to separately rate the intensity of the color green in real-time and on a continuous scale in two videos. The videos were less than five minutes in length, 864x480 resolution, and comprised entirely of solid color frames of green at varying green channel values in RGB color space. In each video, the green value was designed to change at different speeds and times while avoiding discontinuous changes. The intention was to make the annotation process mechanically easy with the simple interface provided and to help ensure the main annotation challenge laid in the translation of green intensity perception to annotation rating.

The annotation process occurred in real time where annotators adjusted an interface in tandem with perceived changes in the video. A slider widget representing a float value between zero (corresponding to black) and one (corresponding to full green) was displayed and annotators were instructed to watch the video and use a mouse to move the slider according to how green the video appeared. No further instructions or clarifications were given. The value of the slider was recorded for each video frame at 30Hz, and a picture of the interface is shown in Fig 1.

Fig 2 shows a plot of all ten annotations alongside the objective truth for both annotation tasks. This plot suggests that annotators were generally quite good at capturing large changes and trends, but had difficulties in several areas. First, most annotators tended to over-shoot the target value when annotating increases or decreases in value over a period of time thus suggesting they were fixated on annotating the rate of change rather than the actual rating. Annotators were sensitive to and captured the appropriate direction of change, but sometimes were unable to estimate the correct rate of change. Secondly, we note that approximately half of the annotators struggled to capture the lack of change in green value especially during the 100 to 150-second time interval in Fig 2b. One possible explanation is that the longer duration of this constant segment gave annotators time to realize their current rating did not match their perception and then adjust the value to match in spite of what was (not) occurring in

(a) **Ground truth for task A**

(b) **Real-time annotations**
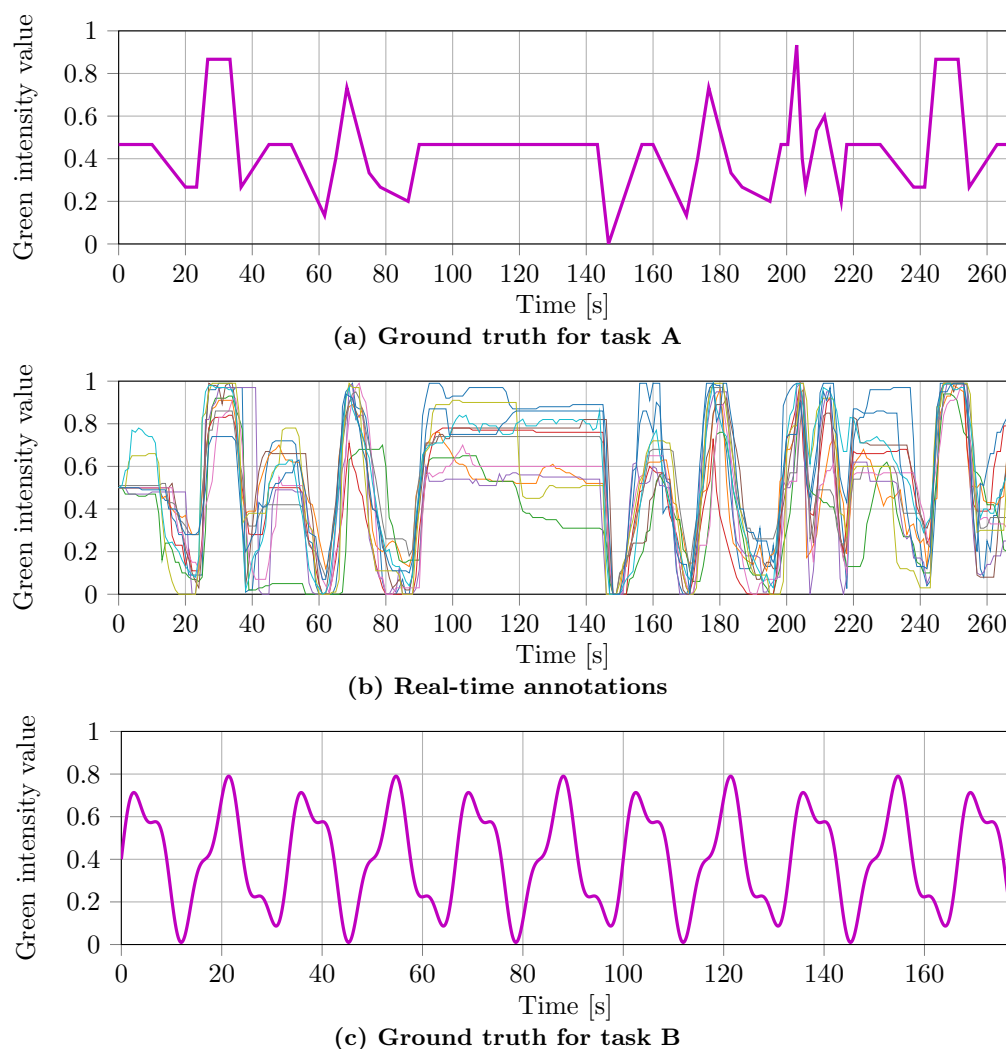

(c) **Ground truth for task B**

Fig 2. Annotations of green channel intensity. (a) Ground truth for the green intensity in RGB space for the annotated video. (b) 10 real-time annotations from different annotators.

the video. Lastly, we note that similar green values were annotated inconsistently over time. In particular, there was a significant difference in average annotation value per annotator between different time intervals where the green intensity was actually at a constant 0.5 value (see Fig 2b).

This last observation implies that even for this relatively simple annotation task, it is difficult for annotators to accurately capture the trends while preserving self-consistency over time. Given prior evidence that humans are better at ranking than rating [3, 19, 20] and our own observations from this study, it is reasonable to assume that in continuous real-time annotation, experts are more focused and perhaps even better at faithfully capturing trends and less able to accurately assess the true value at any point in time. We present a procedure for correcting these rating inconsistencies while preserving the more precise trend annotations.
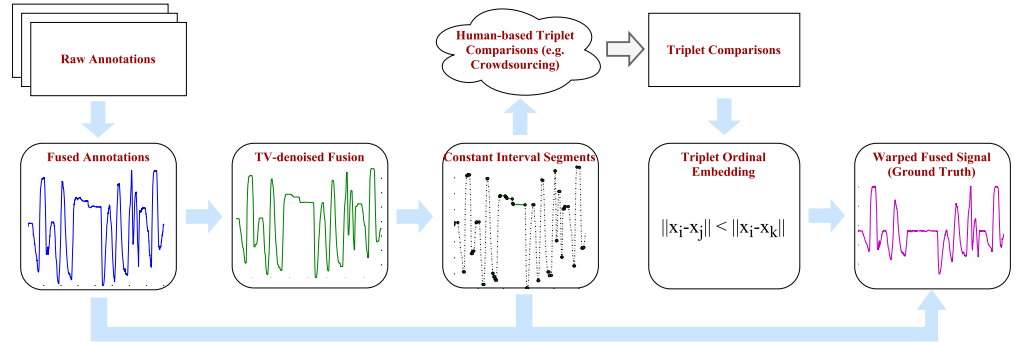
Fig 3. Proposed pipeline for ground truth correction given continuous human annotations.

3 Fused Annotation Warping for Ground Truth Estimation

We propose a method for warping fused annotations to establish a ground truth signal that has been corrected for various global inconsistencies, artifacts, and errors introduced during the real-time continuous human annotation process. The method leverages a recurring observation that annotators more successfully capture trends and less accurately represent exact ratings [3, 19, 20]. In our approach, additional information must be collected from annotators after the continuous annotation task. We leverage the structure of the fused annotations to identify segments of time in the video that can be treated as congruent units and thus reduce the amount of necessary additional information. Further significant reductions to the required supplementary information are discussed in the results section.

Our method is summarized as a sequence of steps shown in Fig 3. The first step fuses the raw annotations together to form a single time series. In this paper, this step simply time-aligns and averages all annotations to reduce systemic noise and limit the influence of random annotation artifacts. Total variation (TV) denoising is then used to approximate the fused signal as a piecewise-constant step function. Constant intervals are thereafter extracted from the denoised signal corresponding to time spans where annotators generally agree that the target construct does not noticeably change. Additional rank information is then procured from annotators to re-evaluate the proper sorting of the these time intervals with respect to the target construct. We collect comparison results among unique triplets of these constant intervals and employ an ordinal embedding technique to re-rank them. Finally, the average signal is warped piecewise-linearly so the corresponding constant intervals align with the embedding. These steps and their assumptions are described in detail in the corresponding sections below.

3.1 Lag Compensation and Averaging (Annotation Fusion)

The first step involves estimating an appropriate time shift for each annotation signal to align them with the video and compensate for lag due to human reaction times. Several methods have been proposed for this [6–8, 11–14] and in principle any choice works for this step. We use a simple per-annotator time shift (EvalDep) proposed by [6]. This method requires some feature sequences to be extracted from the video for alignment, so we provide the green value and its forward difference per frame. The average annotation lag is estimated at 1.6 seconds across annotators. After shifting each

annotation by its own lag estimate, we truncate the trailing frames so all annotations   154
are equal length and then average them in time. Note that the lag compensation is not   155
strictly necessary but yields better final results. Fig 4a shows the time-corrected   156
average annotation signal for one of the annotation tasks.   157

3.2   Total Variation Denoising   158

Total variation denoising has been successfully used to remove salt and pepper noise   159
from images while simultaneously preserving signal edges [21]. In our context, we want   160
to identify the set of nearly constant regions of the average annotation signal   161
corresponding to a lack of noticeable change in the target construct. TV denoising is   162
preferable to other smoothing processes both because it approximates the function as a   163
piecewise-constant step function as is desired, and also because it better preserves the   164
structure of the signal.   165

We use the TFOCS MATLAB library [22] to find a new sequence y_t that
approximates a given sequence x_t by minimizing:

$$y_t = \min_{y_t} \left[\sum_t ||x_t - y_t||_{\ell_2}^2 + \lambda \sum_t ||y_{t+1} - y_t||_{\ell_1} \right]$$

The parameter λ controls the influence of the temporal variation term and degree to   166
which y_t is approximately piecewise-constant. In general this parameter needs to be   167
tuned to produce a desirable sequence. For this study, we hand-tune λ and settle on a   168
value of 0.05. In principle, this parameter can be automatically selected based on other   169
criteria and heuristics, but we leave this endeavor for future work. Fig 4b shows an   170
example TV-denoised signal.   171

3.3   Constant Interval Extraction   172

A simple heuristic method is used to extract nearly constant intervals from the   173
TV-denoised signal. These time intervals become the targets for re-ranking in the next   174
step when an ordinal embedding is applied. In this step, a scan of the TV-denoised   175
signal is performed and the smallest set of (largest) intervals is found where each   176
interval satisfies two criteria: (1) the total height does not exceed threshold h, and (2)   177
the frame length of the interval is at least T frames. Fig 4b shows example extracted   178
approximately constant intervals for one of the annotation tasks.   179

For our experiment, we select $h = 0.005$ and $T = 18$ frames (for 30Hz videos). The   180
height threshold is chosen to be very small relative to the size of the annotation scale so   181
only very flat regions are considered. Because TV denoising does well at approximating   182
the signal as a piecewise-constant function, we find this step in the overall procedure is   183
not very sensitive to h. The T value is selected to match the duration of the fastest   184
change in the objective truth. In practice, this parameter could be approximated from   185
the average annotation signal or tuned manually, but should be set no smaller than the   186
equivalent of 0.25 seconds, which is roughly the average human reaction time. In the   187
future, we hope to obviate these parameters to make this approach more scalable and   188
robust.   189

3.4   Triplet Comparisons   190

In this step during an actual experiment, annotators are asked to view three extracted   191
video segments corresponding to each unique triplet of intervals. One video segment   192
serves as a reference and the other two as test candidates and the annotator is instructed   193
to select which of the two candidate video segments is most similar to the reference. For   194
the purpose of assessing the robustness of this approach to missing and conflicting   195
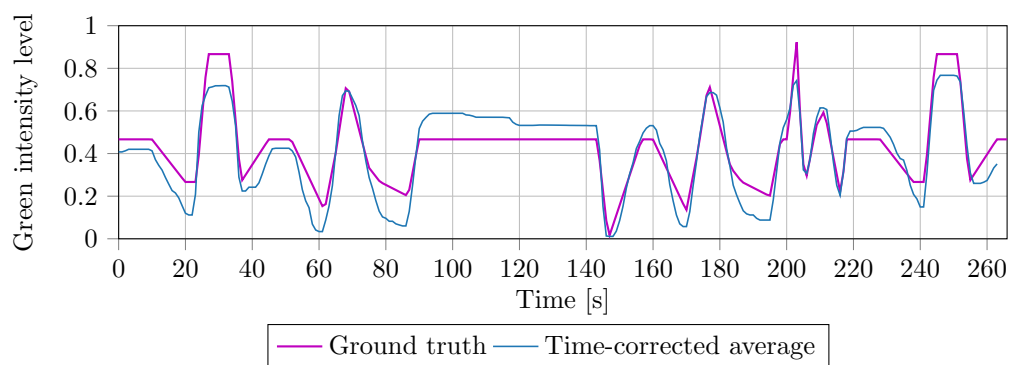
(a) Plot of time-corrected annotation average and objective truth.
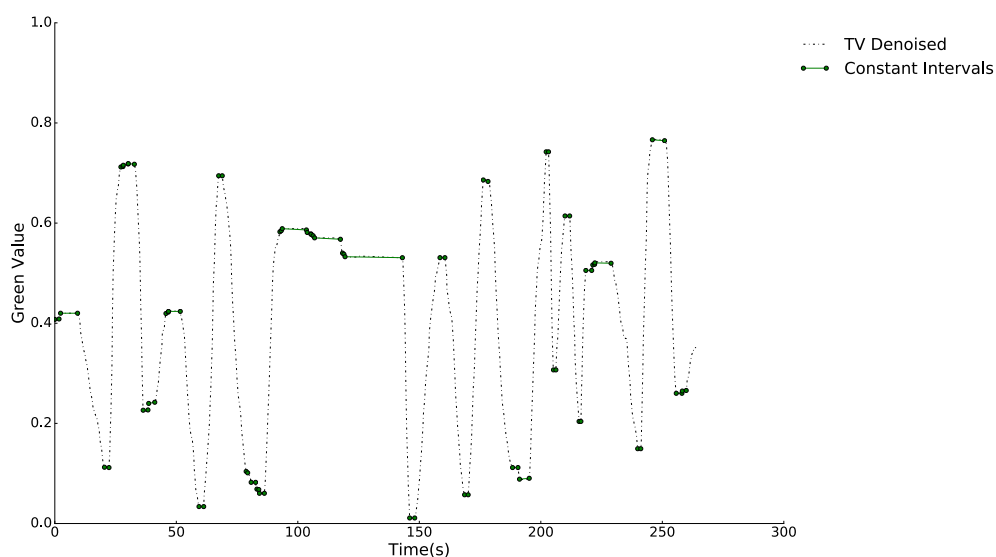
(b) Plot of TV-denoised fused annotation with extracted constant intervals.

Fig 4. Results at intermediate stages of the proposed method pipeline for Task A.

Plot (a) was generated from gt_mariooryad_tv.csv, but is only 264 samples long. Do you know why?

I'm unsure about the position of the legend in (a). This is the best option if we want the figures to be wider, but it could also be on the side.

comparison information, we simulate the comparison results using the objective truth as an oracle. Further analysis of these effects are explored in the results section.

3.5 Ordinal Embedding

Ordinal embedding problems attempt to learn a (typically lower dimension) embedding that preserves a similarity relationship between subsets of data points. Given a set of inputs $\mathcal{Z} = \{z_1, ..., z_n\}$ with each $z \in \mathbb{R}^m$ and a set of similarity relations on 4-tuples from $\mathcal{Z}$ of the form $s(z_i, z_j) < s(z_k, z_l)$ where $\{i, j, k, l\}$ is a 4-subset of $\{1, 2, ..., n\}$, the goal is to find a set $\mathcal{X} = \{x_1, ..., x_n\}$ with each $x \in \mathbb{R}^d$ such that:

$$||x_i - x_j|| < ||x_k - x_l|| \iff s(z_i, z_j) < s(z_k, z_l)$$

for some norm on $\mathcal{X}$. For our application, we are interested in the case $i = k$ where we have ordinal comparisons in the form of triplets (i.e. sample i is more similar to sample j than k). One reason to prefer this simplification of the general problem is that it reduces the cardinality of the set of all possible comparisons given $\mathcal{Z}$ and thus the amount of additional information we need.

A further reduction would be possible if we consider relationships of the form:

$$||x_i|| < ||x_k|| \iff ||z_i|| < ||z_k||$$

for unique index pairs $\{i, k\}$. This setup supposes that it is possible to directly assign a value to each sample z with respect to the target construct for the purpose of comparing two samples, but this may not always be possible. In cases where multiple conflicting or ambiguous criteria exist, as in the annotation of smile strength [18], such a scale may not exist or be too unintuitive for human annotators. So, for generality of this procedure, we choose the triplet comparison approach.

With triplets, the total number of comparisons that can be made when $|\mathcal{Z}| = n$ is $n\binom{n-1}{2}$ which scales $\mathcal{O}(n^3)$. Fortunately, there is considerable redundancy in the comparison information and only a small fraction is necessary to find an embedding close to the optimal embedding. Prediction error bounds have already been derived for a noisy formulation of this triplet embedding problem [23], nonetheless we expect that the number of comparisons must scale like $\mathcal{O}(dn\log(n))$ ($d = 1$ in our experiment).

In our method, ordinal embedding is used to reorder the constant intervals to make them more self-consistent and rank-aligned with the objective truth. Several triplet ordinal embedding solvers have been proposed [24–27]. We employ the t-stochastic triplet embedding (t-STE) approach [26] because, as the authors highlight, it aggregates similar points and repels dissimilar ones. We also favor this approach because it prefers the simpler explanation that two points in the embedding are identical when no evidence suggests otherwise (Occam's Razor principle). Fig 6 shows the embedding results for the extracted constant intervals that have been rescaled to the proper $[0, 1]$ range and computed using a complete set of triplet comparisons from the oracle. Note that the embedding only preserves the relative similarity relationships, so the embedding scale is expected to be off by a (unknown) monotonic transformation of the objective truth's scale.

3.6 Spatial Warping

In the final step, the fused annotation is spatially warped using the ordinal embedding results with the extracted constant intervals to rectify inconsistencies. First, within the time frame of each interval the annotation is corrected so its average over the interval is equal to the corresponding embedding value. Then, the annotation between two intervals is linearly scaled to align with the warped annotation at the neighboring

$$S_t = \begin{cases} \mathcal{E}_i - \frac{1}{|\mathcal{I}_i|} \sum_{s \in \mathcal{I}_i} y_s & \exists \mathcal{I}_i \in \mathcal{I} : t \subseteq \mathcal{I}_i \\ 0 & \text{else} \end{cases} \quad (1)$$

$$y'_t = \begin{cases} y_t + S_t & \exists \mathcal{I}_i \in \mathcal{I} : t \subseteq \mathcal{I}_i \\ \left(\frac{y_t - y_a}{y_b - y_a}\right)(y_b + \mathcal{S}_{b+}) + \left(\frac{y_b - y_t}{y_b - y_a}\right)(y_a + y'_{a-}) & \exists [a,b] = \mathcal{J}_j \in \mathcal{J} : t \subseteq \mathcal{J}_j \end{cases} \quad (2)$$

Fig 5. Equations for our proposed spatial warping method. Let $\mathcal{I}$ and $\mathcal{E}$ be the ordered and corresponding sets of non-overlapping time intervals and embedding values respectively. Define S_t for each interval $\mathcal{I}_i \in \mathcal{I}$ and $i \in \{1, 2, ..., |\mathcal{I}|\}$ to be the difference between interval i's average value and the corresponding embedding value. Let $t \in \{1, 2, ..., T\}$ be a time index, y_t denote the fused annotation signal, y'_t denote the warped signal value, and $\mathcal{J}$ be the set of time ranges between the intervals in $\mathcal{I}$ ($\{1, 2, ..., T\} \setminus \cup_i \mathcal{I}_i$). A subscript $+$ or $-$ denotes a time just before or after the associated time index. Edge cases where the time indices fall outside of $\{1, ..., T\}$ are handled using the average signal value at the boundary.

Maybe you can pose y'_t (or the constants associated) as the solution to an optimization problem?

intervals. We select a linear inter-interval warping function because it avoids distorting the signal. A formal definition is provided by Eqs (1,2) in Fig 5, and Fig 6 shows the results after applying this warping technique.

4 Results

Table 1 shows various agreement measures for different annotation fusion approaches and the objective truth in our perception experiment.

Three baseline annotation fusions are shown, one is a simple average of the expert annotations, one is the maximum likelihood estimation from a per-annotator distortion model [18], and the last is a time-aligned average using evaluator-dependent time shifts [6]. Our rank-based warping method is applied to each baseline using a full set of triplet comparisons from the oracle. In all cases except for one, the warping method achieves significantly better results ($p < 0.005$) when agreement is measured using Pearson correlation or normalized mutual information. In the one outlier case, the "distort" model already approximates the objective truth extremely well because the true signal is very smooth. In this case, the proposed warping method does not diminish the correlation considerably. Although it would seem that rank-based correlation metrics should show a substantial improvement, Spearman and Kendal's Tau correlations slightly decrease in some cases. This is primarily due to rank disagreements over the warped constant intervals, rather than disagreements at a large scale due to the ordinal embedding. The improved self-consistency over time of the warped signal combined with the general improvement in Pearson correlation demonstrate that the warped signal resulting from the proposed method is more suitable for use as a ground truth.

Lastly, we briefly address the robustness of the proposed warping method to incomplete triplet comparisons. Even for modest numbers of constant intervals, the number of triplets required for a complete set grows cubicly. Due to the large amount of redundancy in the triplet comparisons, only a small fraction is necessary for the warping method to approximate the objective truth well. Figure 7 shows a plot of the performance of this spatial warping approach for both incomplete triplet comparisons and partially adversarial comparisons. In our first green value perception experiment
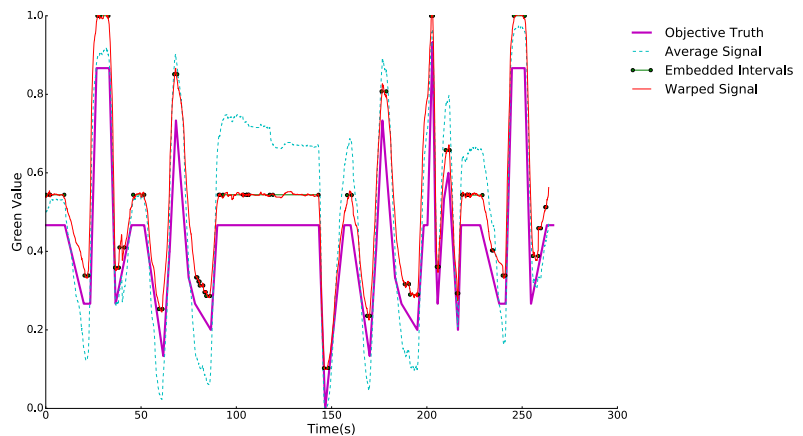
Fig 6. Plot of the objective truth signal, time-shifted average annotation signal, warped signal, and the 1-D embedding for extracted constant intervals for Task A. The spatially warped signal better approximates the objective truth and also achieves greater self-consistency over the entire annotation duration.

Table 1. Agreement measures for baseline and proposed warped fused annotation approaches

Task	Signal Type	Pearson	Spearman	Kendall's Tau	NMI
A	Simple Average	0.775	0.795	0.636	0.302
	Warped Average	0.811†	0.738	0.584	0.307
	Distort*	0.809	0.834	0.676	0.793
	Warped Distort	0.888†	0.839	0.695	0.794
	EvalDep** Average	0.906	0.946	0.830	0.484
	Warped EvalDep	0.967†	0.939	0.835	0.562
B	Simple Average	0.950	0.948	0.804	0.772
	Warped Average	0.964†	0.960	0.828	0.859
	Distort*	0.967	0.966	0.848	0.955
	Warped Distort	0.960	0.962	0.842	0.957
	EvalDep** Average	0.969	0.969	0.855	0.774
	Warped EvalDep	0.988†	0.987	0.906	0.862

All warped results use a complete set of ordinal comparisons from the oracle. NMI = normalized mutual information.

$\dagger$ - significant improvement ($p < 0.005$, using a Fisher z-transform) of the warped methods over the respective signal

* - method from [18]

** - method from [6]

(Task A), there are 40 extracted intervals and thus 29,640 possible comparisons. Assuming a uniform five percent triplet comparison error rate, significant improvement over the best tested baseline method is achieved with only two percent, or about 600, randomly selected triplet comparisons.
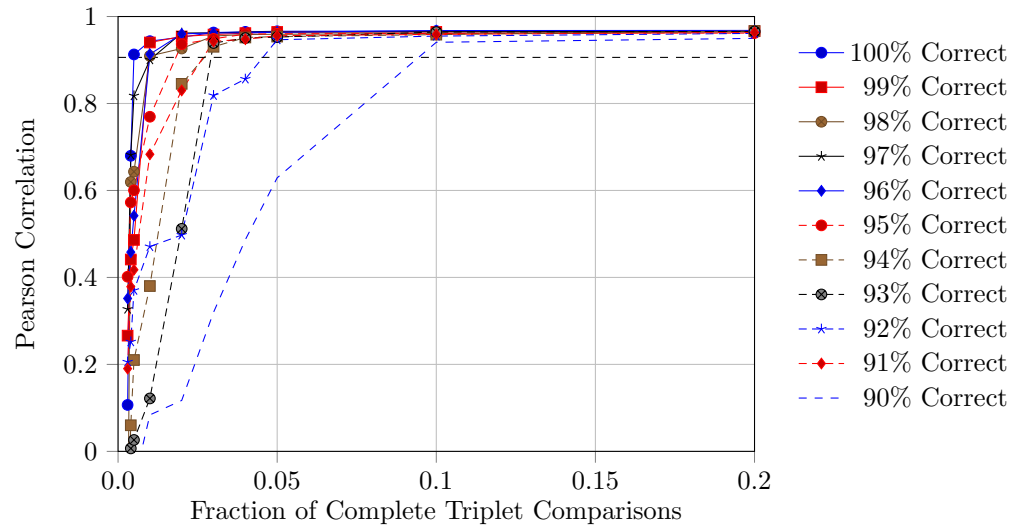
Fig 7. Plot showing Pearson correlation of the warped annotation (using Task A and the fusion from [6] for different percentages of the total number of possible triplet comparisons. Several plots are shown for varying levels of average annotation accuracy. A high correlation is possible with a small fraction of the triplets due to large amounts of redundancy in the comparisons. Dotted line represents the fused annotation correlation.

5   Future Work   267

There are several compelling research directions for expanding on this work which we   268
aim to address in future papers. The total variation denoising procedure requires careful   269
selection of an unintuitive tunable constant to achieve desirable results. Consideration   270
of other problem constraints, such as the quota of annotation resources available or the   271
required accuracy on predictions from the resulting ground truth, could be used to find   272
a sensible value for this parameter. The subsequent constant interval extraction step   273
also has two parameters that could be chosen automatically from the data given some   274
additional heuristics or constraints. Further analysis of this method's ability to produce 275
accurate ground truth estimates for more complex continuous annotation tasks, like 2-D   276
dimensional affect, is another exciting avenue. Larger reductions in the number of   277
required triplet comparisons may also be possible using adaptive sparse sampling   278
techniques and stochastic transitivity.   279

6   Conclusion   280

In this paper we propose a novel method for extending continuous real-time human   281
annotation fusion approaches to generate a more accurate ground truth. We leverage   282
the natural ability of annotators to provide accurate similarity comparisons and propose   283
a procedure for warping the fused annotation to better align with the target construct.   284
We test our approach in a mechanically simple but perceptually difficult annotation   285
experiment where an objective truth is known and show that our approach yields a   286
signal significantly more correlated with the objective truth even with the presence of   287
several annotation artifacts. We hope this method finds utility as a means for   288
establishing a more accurate ground truth in hidden state problems where no objective   289
truth is available *a priori*.   290

References

1. McKeown G, Valstar M, Cowie R, Pantic M, Schroder M. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Transactions on Affective Computing. 2012;3(1):5–17.

2. Metallinou A, Katsamanis A, Narayanan S. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. Image and Vision Computing. 2013;31(2):137–152.

3. Metallinou A, Narayanan S. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. IEEE; 2013. p. 1–8.

4. Ringeval F, Sonderegger A, Sauer J, Lalanne D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. IEEE; 2013. p. 1–8.

5. Valstar M, Gratch J, Schuller B, Ringeval F, Lalanne D, Torres Torres M, et al. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM; 2016. p. 3–10.

6. Mariooryad S, Busso C. Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators. IEEE Transactions on Affective Computing. 2015;6(2):97–108.

7. Müller M. Dynamic time warping. Information retrieval for music and motion. 2007; p. 69–84.

8. Ringeval F, Eyben F, Kroupi E, Yuce A, Thiran JP, Ebrahimi T, et al. Prediction of ashynchronous dimensional emotion ratings from audiovisual and physiological data. Pattern Recognition Letters. 2015;66:22–30.

9. Hotelling H. Relations between two sets of variates. Biometrika. 1936;28(3/4):321–377.

10. Nicolaou MA, Zafeiriou S, Pantic M. Correlated-spaces regression for learning continuous emotion dimensions. In: Proceedings of the 21st ACM international conference on Multimedia. ACM; 2013. p. 773–776.

11. Zhou F, Torre F. Canonical time warping for alignment of human behavior. In: Advances in neural information processing systems; 2009. p. 2286–2294.

12. Andrew G, Arora R, Bilmes JA, Livescu K. Deep canonical correlation analysis. In: Proceedings of the International Conference on Machine Learning; 2013. p. 1247–1255.

13. Nicolaou MA, Pavlovic V, Pantic M. Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations. IEEE transactions on pattern analysis and machine intelligence. 2014;36(7):1299–1311.

14. Trigeorgis G, Nicolaou MA, Zafeiriou S, Schuller BW. Deep Canonical Time Warping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 5110–5118.

15. Zhou F, De la Torre F. Generalized canonical time warping. IEEE transactions on pattern analysis and machine intelligence. 2016;38(2):279–294.

16. Audhkhasi K, Narayanan S. A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels. IEEE transactions on pattern analysis and machine intelligence. 2013;35(4):769–783.

17. Zhu T, Dunkley N, Behar J, Clifton DA, Clifford GD. Fusing continuous-valued medical labels using a Bayesian model. Annals of biomedical engineering. 2015;43(12):2892–2902.

18. Gupta R, Audhkhasi K, Jacokes Z, Rozga A, Narayanan S. Modeling multiple time series annotations based on ground truth inference and distortion. IEEE Transactions on Affective Computing. 2016;PP(99):1–1.

19. Yannakakis GN, Hallam J. Ranking vs. Preference: A Comparative Study of Self-reporting. In: Affective Computing and Intelligent Interaction: 4th International Conference. Springer; 2011. p. 437–446.

20. Yannakakis GN, Martínez HP. Ratings are overrated! Frontiers in ICT. 2015;2:13.

21. Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena. 1992;60(1-4):259–268.

22. Becker SR, Candès EJ, Grant MC. Templates for convex cone problems with applications to sparse signal recovery. Mathematical programming computation. 2011;3(3):165.

23. Jain L, Jamieson KG, Nowak R. Finite Sample Prediction and Recovery Bounds for Ordinal Embedding. In: Advances In Neural Information Processing Systems; 2016. p. 2703–2711.

24. Agarwal S, Wills J, Cayton L, Lanckriet GR, Kriegman DJ, Belongie SJ. Generalized Non-metric Multidimensional Scaling. In: Proceedings of the Artificial Intelligence and Statistics Conference; 2007. p. 11–18.

25. Tamuz O, Liu C, Shamir O, Kalai A, Belongie SJ. Adaptively Learning the Crowd Kernel. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). ACM; 2011. p. 673–680.

26. Van Der Maaten L, Weinberger K. Stochastic triplet embedding. In: IEEE International Workshop on Machine Learning for Signal Processing (MLSP). IEEE; 2012. p. 1–6.

27. Amid E, Ukkonen A. Multiview Triplet Embedding: Learning Attributes in Multiple Maps. In: Proceedings of the International Conference on Machine Learning; 2015. p. 1472–1480.