# Continuous-time Annotation Fusion via Rank-based Signal Warping

Brandon M. Booth, Karel Mundnich

January 23, 2017

## 1 Introduction

Many machine learning tasks pertaining to human behavior involve supervised learning from noisy labels that are taken as the ground truth or "gold standard". Whether used for classification or regress, having a set of labels that congruently and accurately describe the samples is of paramount importance. In certain hidden mental state problems, such as attention, engagement, or productivity, attempts to have participants momentarily introspect and provide labels interrupts the normal flow of their experience and can lead to noisy labels especially for longer sessions. In these cases, other humans are used as judges for the target construct and often several annotations are collected from a group of expert annotators to establish a ground truth. Nevertheless, these annotations suffer from a variety of artifacts: reaction lag, personal bias, input noise, human input error, global and local inconsistencies, and adversarial annotation.

In this work, we focus primarily on a continuous and real-time annotation scheme where annotators view a recording of a session (i.e. a video of a participant during a trial) and annotate the target construct in real-time. Other schemes are possible, including offline and/or discrete annotation, but we believe continuous annotation in real-time is the most natural because it allows annotators to make subtle distinctions in state changes and also because it allows them to account for all of the context leading up to a particular moment. These affordances lead to further sources for noise and artifacts: preference for local consistency and quality reduction over time due to high vigilance demands. We aim to address these types of artifacts for continuous and real-time annotations with a correction procedure.

Prior work on continuous annotations focuses on ground truth estimation via handling certain sources of noise or artifacts. Several works tackle the problem of correcting for varying lag in each annotator's reaction time by aligning the signals [**?**, **?**, **?**, **?**] and some methods also simultaneously model annotator biases and attempt to remove them to recover the intended annotation signal [**?**]. Newer research uses an LSTM to try to model and correct for these noise sources [**?**]. Typically, the ground truth is used in a machine learning framework for prediction, and a few methods integrate the annotation aggregation into the machine learning stage and learn a projection function for the set of features that during training maximizes the correlation of the projection with the set of annotations [**?**, **?**].

Need to distinguish our work from annotation bias compensation works

Nevertheless, the annotation of continuous-time variables, and specifically, affective or emotional dimensions in multimedia has recently gained traction. Current methods for fusing continuous annotations attempt to correct for known artifacts [**?**], assuming a shared latent space among annotators [**?**, **?**]. Some methods attempt to correct for lag in reaction time to changes in the stimulus. These methods range from simple shifts in time to dynamic time warping where varying-length time windows are shifted different amounts of time [**?**]. These approaches require some representation of the target stimulus and often use several features extracted from the stimulus for time alignment. The obvious downside to these approaches is that the alignment may fail if changes in the extracted features to not correspond to perceivable changes in the stimulus.

Other fusion methods model the different kinds of artifacts introduced by different types of annotators (cite Rahul).

Other fusion methods use event detection to locate points of interest in time and use windows between these boundaries for fusion (maybe? cite?).

Several studies have shown that people are better at ranking compared to rating [**?**].

Include background research and citations. Include Yannakaki's paper [**?**].

The problems for which modern fusion approaches are adapted are primarily hidden state estimation problems such as the engagement level of a student viewing an online lecture. Because for these types of problems the true value of the hidden state is not known, fusion approaches are typically validated using the resulting fused signal in a machine learning context to show that the model is better able to predict the target. The problem with this type of validation is that it conflates any potential success of the fusion approach with the success of the learning model making it difficult to tell whether the fusion approach actually leads to a better representation of the object truth signal (e.g. the actual engagement value of a student). To this end, we conduct our own study where annotators are instructed to provide continuous annotations in real-time where an objective truth signal is known *a priori*. We are careful to pick a task where differences in each annotator's prior biases and perception lead to disagreements and artifacts in each annotation signal.

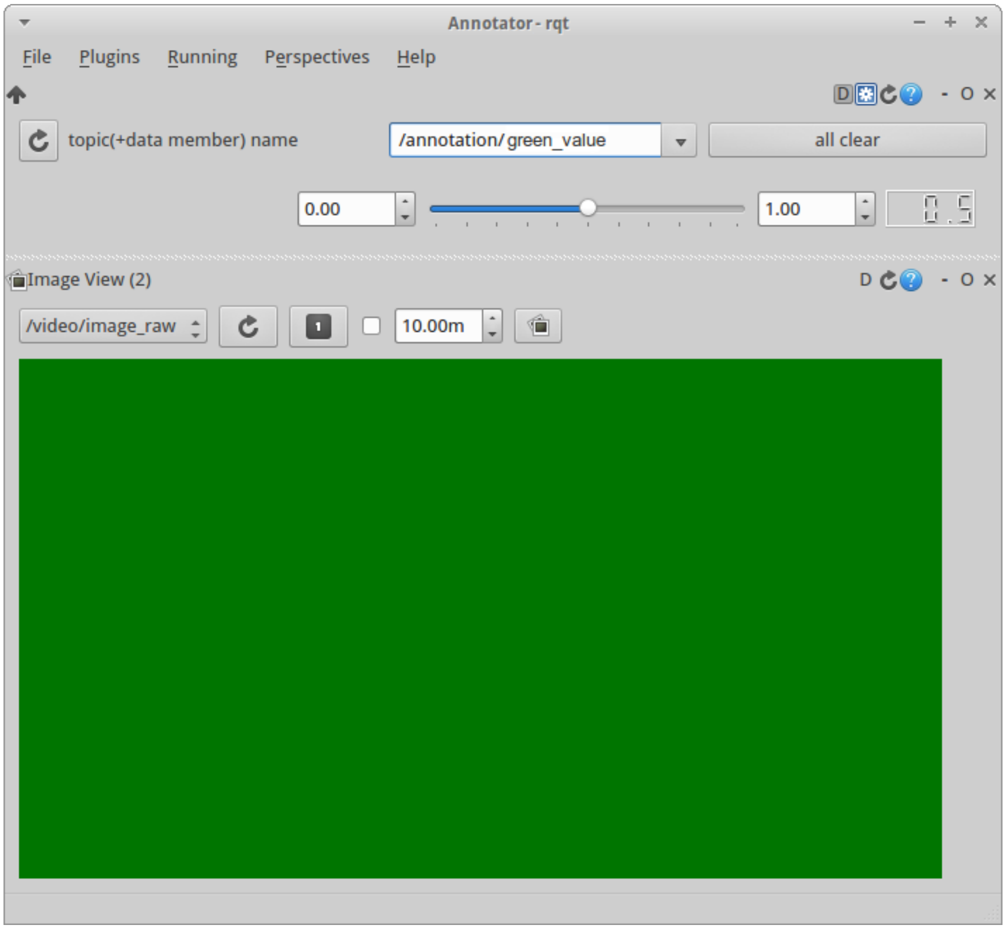
## 2 Experiment and Results

Ten annotators were asked to annotate the level of "green-ness" in a video. The video was 4:27 minutes in length, 864x480 resolution, and comprised entirely of solid color frames of green at varying green channel values in RGB color space. The green value was designed to change both slowly and rapidly at different times and frequencies and also in an unpredictable manner, but did not contain any sudden oscillations. The slow varying and smooth changes were chosen because they mimic the often slow changes in other hidden mental state problems (engagement, productivity, etc. (cite?)). Avoiding sudden oscillations means the annotators were not required to make any "twitch" reactions during the annotation process which would have added unnecessary artifacts.

The annotation process occurred in real time. A slider widget representing a float value between zero and one was displayed and annotators used a mouse to move the slider as they watched the video. The value of the slider was recorded for each video frame. Figure 1 shows the interface.

Figure 2 shows a plot of all ten annotations alongside the objective truth. This plot suggests that annotators are generally quite good at capturing large changes and trends for this task, but have difficulties in several areas. First, most annotations tend to over shoot the target value when annotating increases or decreases in value over a period of time. This is perhaps due to the lack of a reference color for various values of green thereby forcing annotators to estimate rate of change and focus on annotating a consistent rise or fall in the value. By this logic, we consider the tendency to overestimate rather than underestimate merely a coincidence in this example, though it may be illustrative of a psychological trend. Secondly, we note that approximately half of the annotators struggle to capture the lack of change in the stimulus especially near the 100 to 150 second time interval. We posit this artifact is mainly due to the annotators having a few moments to stop focusing on capturing the value changes and refocus on accuracy. In these cases it seems the annotators noticed the value of the slider did not match their perception of the green value and thus changed the slider value to attempt to move it to the correct location. Lastly, we note that similar green values are annotated inconsistently over time. In particular, there is a significant difference in average annotation value per annotator between the sections where the stimulus is at a 0.5 constant value.

This last observation suggests that even for this relatively simple annotation task, it is difficult for annotators to accurately capture the trends while preserving self-consistency over time. It seems reasonable to conclude that annotators are more focused, and perhaps even better at, faithfully capturing trends and



topic(+data member) name /annotation/ green_value

0.00 1.00 0.5

/video/image_raw 10.00m

Figure 1: Snapshot of the user interface for the green channel value annotation task

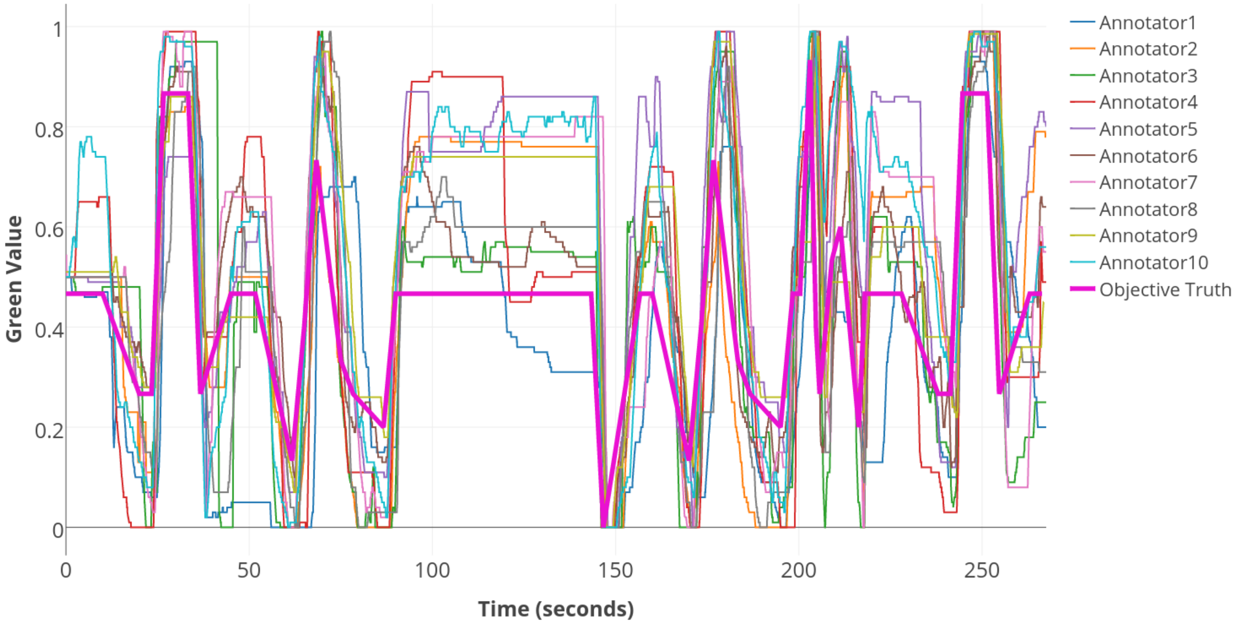


Figure 2: Annotations of green channel values in the video alongside the true value

less able to accurately assess the true value at any point in time. Our primary aim in this paper is to propose a procedure to correct for these global inconsistencies and artifacts.

## 3 Ground Truth Estimation

We present a method for annotation fusion for the purposes of establishing a ground truth signal that corrects for global inconsistencies, artifacts, and errors introduced during the real-time continuous annotation process. The following list names each step of the procedure.

1. Annotation average
2. Total variation denoising (TV denoising)
3. Constant interval extraction
4. Triplet Comparisons
5. Non-metric multi-dimensional scaling (NMDS)
6. Signal warping

The first step averages all annotations to limit the influence of systemic noise such as mouse jitter and sampling aliasing. Total variation (TV) denoising is used to approximate the average signal as a piecewise-constant step function. Nearly constant intervals are extracted from the denoised signal and correspond to time spans where annotators generally agree that the hidden objective truth did not change noticeably. At this point, we focus our efforts to fix-up the signal by asking for additional information from the annotators to help re-evaluate the proper order of the target with respect to the stimulus during these time intervals.

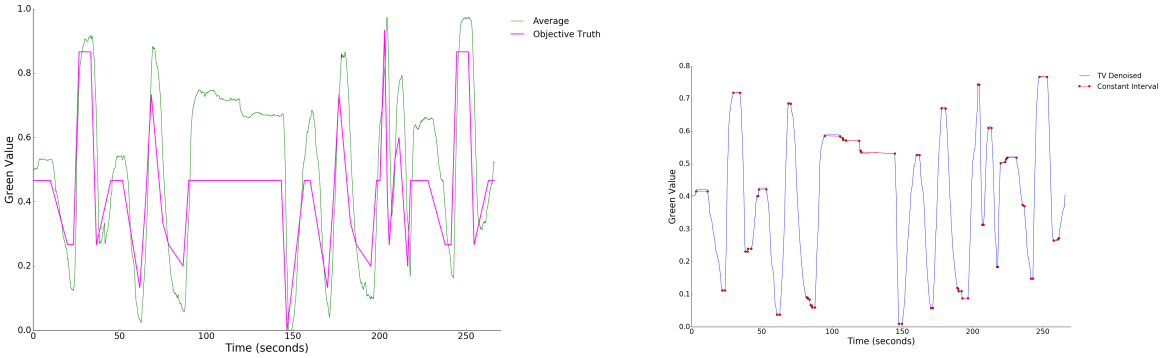


Figure 3: Plot of annotation average and objective truth on the left and TV denoised average signal with extracted constant intervals on the right

We collect ordinal relationships between triplets of these constant intervals from annotators and employ a 2D ordinal embedding technique to re-rank the intervals. Finally, the original average signal is warped linearly to more closely match the re-ranked piecewise-constant step function. Each of these steps and their assumptions are described in detail in the corresponding section below.

## 3.1 Total Variation Denoising

Total variation denoising has been successfully used to remove granular noise from images while simultaneously preserving signal edges. Our aim at this step of the procedure is to identify regions in the average annotation signal where the stimulus is not changing noticeably. We focus on obtaining a set of these time intervals under the assumption that annotators already capture the changes in the stimulus quite well and we only need to make adjustments to the constant portions. TV denoising is preferable to other smoothing processes both because it explicitly attempts to model the function as a piecewise-constant step function as is desired, and also because it minimizes smoothing near transition boundaries which maximizes the size of each constant interval.

We use the Matlab TFOCS library to find a new signal $y_t$ that approximates the signal $x_t$ and minimizes:

$$\min_{y_t} \left[ \sum_t (x_t - y_t)^2 + \lambda \sum_t |y_{t+1} - y_t| \right]$$

The parameter $\lambda$ controls the influence of the temporal variation term and encourages signals that look more piecewise-constant. In general this parameter needs to be tuned to produce a desirable TV signal. For this study, we casually hand-tuned $\lambda$ and settled on a value of 0.05. In principle, this parameter can be automatically selected based on other criteria, but we save this endeavor for future work. Figure 3 shows the averaged annotation signal and the TV denoised signal.

## 3.2 Non-metric Multidimensional Scaling

Non-metric multidimensional scaling (NMDS), also known as ordinal embedding, is a technique for positioning points in a metric space that satisfy ordinal constraints of the form:

$$||x_i - x_j|| < ||x_k - x_l||$$

where $i, j, k, l$ is a 4-subset of $1, 2, ..., |\mathcal{D}|$ and $\mathcal{D}$ is the set of all samples of $x \in \mathcal{R}^d$. For our application, we are interested in the case where $i = k$ 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 NMDS problem is that it reduces the cardinality of the set of all possible comparisons given a data set $\mathcal{D}$ and thus the amount of additional information we need.

It is worth mentioning a further reduction is possible if we consider relationships of the form $||x_i|| < ||x_k||$ for unique values of $i$ and $k$. This setup supposes that it is possible to directly assign a value to each sample $x$ for the purpose of comparing two samples, but this may not always be possible for example in cases where multiple conflicting or ambiguous criteria may exist for valuation (e.g. annotation of behavioral "weirdness"). So, for generality of this procedure we choose the triplet comparison approach.

Even after electing to use triplet comparisons, the total number of comparisons that can be made when $|\mathcal{D}| = n$ is $n * \binom{n-1}{2}$ which scales $\mathcal{O}(n^2)$ with the number of data samples. Fortunately, there is considerable redundancy in the comparison information and not all are necessary to find an approximately optimal embedding. Concentration bounds have already been derived for this triplet comparison problem and we know that a perfect embedding is possible with $\mathcal{O}(dn \log(n))$ triplet comparisons.

We employ this ordinal embedding technique to reorder the constant intervals and make then more self-consistent and hopefully aligned with the ground truth. Normally, we would solicit the triplet comparisons from the annotators by asking them to view three video segments corresponding to each unique set of intervals. One video would serve as a reference and the other two as the tests and the annotator would select which of the two test video segments is most similar to the reference. Our purpose in this study is to assess the robustness of this approach to missing and conflicting additional comparison information, so we simulate the comparison results. An analysis is presented in the results section, but Figure **??** shows the embedding results for our extracted constant intervals that have been rescaled to the proper $[0, 1]$ range.

### 3.3 Signal Warping

## 4 Results

Show correlations chart with perfect information

Show results with imperfect information (incomplete triplets)

Show results with conflicting triplets and complete information

Show results with missing information and incomplete triplets (simulation of adversarial annotator, ISN'T THIS THE BYZANTINE GENERALS PROBLEM?)

## 5 Discussion

Why is this better? Even with perfect triplet comparison information there is no way for NMDS to reconstruct the scale. NMDS does guarantee that there exists some monotonic transformation of the output that would yield the original signal. If the goal is to treat the warped signal as a ground truth and then machine learn with it, then there is no need to learn this monotonic transformation function because the "distorted" ground truth signal is still self-consistent (i.e. the warped signal is approximately a dilation of the true signal).

# 6 Conclusion

# 7 Future Work

Finding the right value for the tunable TV denoising constant. Perhaps choosing a parameter that creates N constant intervals where concentration bounds guarantee fewer than M type I/II errors given X annotators and P probability of adversarial ordinal comparisons.

Adjusting the window size for the constant interval extraction and the maximum allowable slope for the nearly constant portions.

# 8 Acknowledgements

This research is supported by... Dr. Narayanan.