TRAPEZOIDAL SEGMENT SEQUENCING: A NOVEL APPROACH FOR FUSION OF HUMAN-PRODUCED CONTINUOUS ANNOTATIONS

Brandon M. Booth and Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab (SAIL), University of Southern California
brandon.m.booth@gmail.com, shri@sipi.usc.edu

ABSTRACT

Generating accurate ground truth representations of human subjective experiences and judgements is essential for advancing our understanding of human-centered constructs such as emotions. Often, this requires the collection and fusion of annotations from several people where each one is subject to valuation disagreements, distraction artifacts, and other error sources. This work proposes *trapezoidal segment sequencing*, a new method for fusing annotations into a single representation that, when used alongside a recently proposed signal warping pipeline for correcting annotation artifacts, produces accurate ground truths. We prove that annotations can be well approximated with trapezoidal signals and present results showing the proposed method performs competitively with state-of-the-art fusion methods on a data set where the true target signal being annotated is known. The main utility of the proposed approach is its ability to help segment individual annotations into interpretable regions where either changes or no perceived changes to the construct occur.

Index Terms— Trapezoidal signal regression, annotation fusion, triplet embedding, signal warping

1. INTRODUCTION

Human-produced annotations of human subjective experiences and judgement are important for machine learning and improving our understanding of many human-centered constructs, such as emotions and behaviors elicited by everyday stimuli. Many data sets, like SEMAINE [1], MOHNOB-HCI Tagging [2], DEAM [3], DEAP [4], and SEWA DB [5], represent individuals' emotional responses to stimuli using annotations provided by humans. The utility of annotations is not limited to emotional labeling and sees application in student engagement assessment [6] and sincerity perception [7], among many others. Human-produced annotations of human experience are the primary instruments for generating labels suitable for supervised machine learning, so it is paramount that the ground truth labels inferred from the annotations are as accurate as possible.

A typical strategy for limiting the influence of bias or artifacts on the ground truth representation is to collect a multiplicity of annotations of the construct of interest (e.g., emotional valence) from a variety of human annotators and then fuse them into a single representation. Producing an accurate ground truth label in this manner can be quite difficult due a variety of factors, such as temporal lag due to human reaction times, imprecise motor control, uncertainty or ambiguity in perception, or differences in construct valuation.

This paper focuses on one annotation scheme common to many data sets, and customary for dimensional affect annotation: continuous-scale, value-based, real-time annotation. An example of this style of annotation appears in the SEMAINE [1] database, which contains real-valued annotations of valence and arousal levels collected as annotators viewed video clips. Many methods have been proposed for fusing these kinds of annotations ,which can be characterized either as value projection and approximation methods or temporal correction methods. Correlated spaces regression [8], canonical correlation analysis [9], and neural network variants [10] are examples of value projection and approximation methods. These algorithms assume that the true construct value exists in a subspace spanned by features extracted from the stimulus (e.g., facial expressions, eye gaze, heart rate). Other methods such as dynamic time warping [11], and evaluator-dependent time shifting [12] attempt to correct for varying annotation delays due to human reaction time. Fusion is then sometimes achieved via frame-wise averaging after temporal alignment. More recent fusion methods combine both value projection and time correction such as generalized canonical time warping [13] and neural network variants [14]. Though each of these methods can improve the ground truth representation in some manner, none of them can account for annotation artifacts produced by distractions, perceptual uncertainty, or inconsistency in valuation over time.

Recently, a pipeline procedure was proposed [15] capable of correcting many kinds of artifacts, enhancing the valuation consistency, and improving the accuracy of the ground truth. The pipeline achieves this by leveraging a recurring observation that humans more reliably annotate perceived changes to the construct and are less capable of assigning accurate values at any given

The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No 2017-17042800005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

time [15]–[18]. Figure 1 shows a flow diagram of the different pipeline stages proposed in [15] and also includes an improved signal approximation method from [19]. The main shortcoming of this pipeline, which this work will address, is that it requires annotations to be fused first, using any existing fusion technique, before it corrects valuation and artifact errors. Operating on fused annotations means that important per-annotator information may be lost.

This work proposes a novel fusion method where each individual annotation is approximated by a trapezoidal signal and then converted to a sequence of values representing the direction of change in the construct per time frame. The trapezoidal signal representation allows each annotation to be segmented into regions of perceived change and no apparent change. We argue this segmentation strategy is a natural approach, given the growing amount of evidence that annotators more reliably capture changes or the lack thereof. A fusion compatible with the signal warping pipeline is then achieved by merging these sequences of construct changes.

The unique contributions of this work are (1) a novel fusion method using trapezoidal signal approximations, (2) a proof that trapezoidal signals can approximate any continuous function on a compact domain to arbitrary precision (i.e., an annotation signal), and (3) validation of the proposed fusion method using data from [15] where the true construct value is known *a priori*. Further benefits and implications of the proposed fusion technique are discussed. Open-source code implementing the proposed method and signal warping pipeline are made available at `https://github.com/brandon-m-booth/2019_continuous_annotations`.

2. TRAPEZOIDAL SIGNALS

The work in [19] introduces trapezoidal signals and uses them to approximate continuous-time, continuous-scale human annotations. Though this function class is shown to be helpful for improving annotation accuracy in that work, it is not made clear whether it is mathematically sensible to use this class to approximate annotations. We briefly introduce trapezoidal signals here and then proceed to demonstrate that these signals are universal annotation approximators.

2.1. Definition

Figure 2 shows two trapezoidal signals. The first is a typical example of a trapezoidal function. The second image shows a less prototypical trapezoidal signal using the broader definition from [19]: a piece-wise continuous function consisting of alternating sloped and constant line segments. The authors of [19] provide an algorithm for discovering the best-fit trapezoidal function for a set of data points using a pre-specified number of segments T. We make use of this algorithm in our proposed fusion method.

2.2. Universal Signal Approximation

Here we provide a formal proof that trapezoidal functions can approximate any human-produced annotation to arbitrary precision, represented as a continuous function on a compact domain. Note that in practice these annotations are sampled, so when T exceeds the number of data points, there is no residual error.

THEOREM: Let $f : [a,b] \to \mathbb{R}$ be a continuous function with compact support. There exists a sequence of trapezoidal functions $(\phi_n)_{n\in\mathbb{N}} : [a,b] \to \mathbb{R}$ such that $\phi_n \to f$ uniformly.

PROOF: Given some $\epsilon > 0$ and $x,y \in [a,b]$, $\exists\delta > 0$: $|x-y| < \delta \implies |f(x)-f(y)| < \epsilon$ because f is continuous. Let δ_0 be one such δ and let $n \in \mathbb{N}$. Partition the domain $[a,b]$ into open intervals (x_i,x_{i+1}) as follows: $\forall i \in \{0,1,...,n\} : x_i = a + i \cdot \frac{b-a}{n}$. Let ϕ_n be a sequence of trapezoidal functions such that $\forall i : \phi_n(x_i) = f(x_i)$. Within each interval (x_i,x_{i+1}), let $s = x_i + x_{i+1}$ and let:

$$\phi_n(x) = \begin{cases} f(x_i) & x_i < x \leq \frac{s}{2} \\ f(x_i) + \frac{(f(x_{i+1})-f(x_i))(2x-s)}{x_{i+1}-x_i} & \frac{s}{2} < x < x_{i+1} \end{cases}$$

$\phi_n(x)$ is a piece-wise continuous trapezoidal function and consists of two segments per interval, thus alternating constant and linear segments. By construction, $\forall x \in (x_i,x_{i+1})$ either: $f(x_i) \leq \phi_n(x) < f(x_{i+1})$ or $f(x_i) \geq \phi_n(x) > f(x_{i+1})$. From the intermediate value theorem, $\exists x' \in (x_i,x_{i+1}) : f(x') = \phi_n(x)$ since $\phi_n(x)$ is continuous in the same interval and shares the same boundary values. Pick some $N \in \mathbb{N}$ such that $N > \frac{b-a}{\delta_0}$. $\forall n \geq N : |x_{i+1} - x_i| < \delta_0$, therefore $|f(x_{i+1}) - f(x_i)| < \epsilon$. Given any $x \in [a,b]$ and $n \geq N$, we consider $|f(x) - \phi_n(x)|$ for two cases. If $x = x_i$ for some i, then $|f(x) - \phi_n(x)| = 0$. If $x_i < x < x_{i+1}$ for some i, then $|f(x) - \phi_n(x)| = |f(x) - f(x')|$ for some $x' \in (x_i,x_{i+1})$. So, $\forall n > N$, $|x - x'| \leq |x_{i+1} - x_i| < \delta_0 \implies |f(x) - f(x')| = |f(x) - \phi_n(x)| < \epsilon$. ■

3. ANNOTATION FUSION METHOD

The proposed method for fusing annotations combines the first two steps from the signal warping pipeline in Figure 1. First, the annotations are aligned to the stimulus using any time-alignment method. Each annotation is then approximated by a trapezoidal signal [19], and a trapezoidal segment sequence (TSS) is generated for each annotation from the regression. Finally, the sequences are merged to produce a single TSS that represents the fused annotations and can be easily consumed by the next stage in the pipeline (constant interval segmentation). Each of these steps is outlined in detail below.

3.1. Temporal Alignment

In order to remove delays introduced into the annotations by lag in human reaction time, the annotations are time-adjusted to better

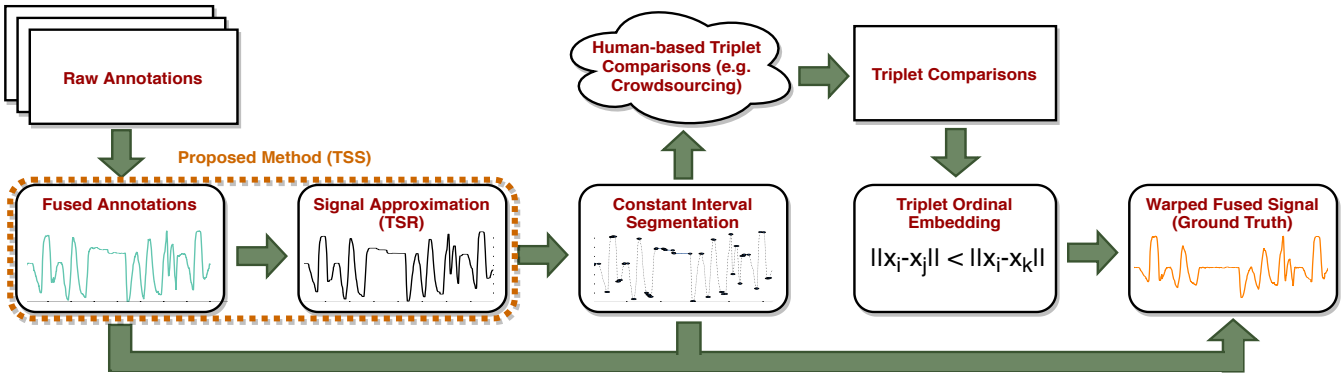

Fig. 1. Signal warping algorithm first presented in [15]. This work proposes a novel annotation fusion method that merges the fusion and signal approximation steps in this pipeline.

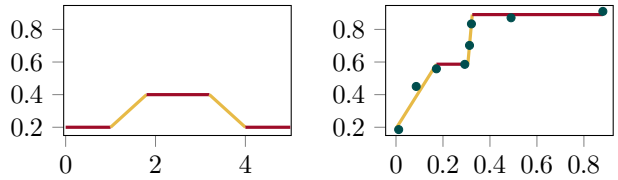

Fig. 2. The left image shows a prototypical trapezoidal signal. The right one illustrates a best-fit, four-segment trapezoidal signal to the data points using the broader definition of these types of signals from [19]. (Images reproduced from [19] with permission).

align them with the stimulus. This step is not strictly necessary, but since the ground truth that is ultimately produced may be used for modeling, it is valuable to perform this step. This has the added benefit of aligning the annotations to each other before performing fusion, which may improve the quality of the final ground truth. Some existing methods that serve this function are the *EvalDep* algorithm [12] and *Dynamic Time Warping* (DTW)[11].

3.2. Trapezoidal Signal Approximation

The *Trapezoidal Segmented Regression* (TSR) method from [19] is applied to each aligned annotation to produce a set of trapezoidal signal approximations. The TSR method requires an additional input T, which is the desired number of segments in the regression. As observed in [19], when T is large enough to capture the basic structure of any particular input signal, then further increases to T yield quickly diminishing returns in approximation accuracy, which makes estimation of the T parameter straightforward. In our experiments, T is roughly estimated by a human observer and then set to a value 1.2 times higher to ensure it is large enough. This parameter could be more precisely tuned using an *elbow* optimization technique, for example, but as we show in Section 4, this rough method works well and requires little time.

3.3. Trapezoidal Segment Sequencing

Each annotation's trapezoidal signal approximation is converted to a new time series where each frame is assigned a value from $\{-1,0,1\}$ based on whether the approximation shows a negative change (decreasing linear segment), no change (constant segment), or a positive change (increasing linear segment) respectively. We call this new time series the trapezoidal segment sequence (TSS).

3.4. Fusion

The TSS sequences from the annotators are pooled and majority voting is used per frame to create a single fused TSS. Ties involving constant segments are assigned a zero value; otherwise, the ties are resolved arbitrarily. In the next stage of the signal warping pipeline (see Figure 1), constant intervals can easily be extracted from the fused TSS by finding all sub-sequences of contiguous zeros.

4. VALIDATION

We test the proposed method using the green intensity annotation data set from [15]. In this data set, annotators were asked to separately rate the intensity of the color green in two videos which displayed only this color at varying intensities over time. We choose this data set for validation because the true green intensity is completely known *a priori*, unlike in many available emotion and human behavior data sets, which thus allows us to assess the ground truth accuracy.

4.1. Experiments

For both annotation tasks, Task A and Task B, in the green intensity data set, we first down-sample the annotations to 1Hz from their native 30Hz to expedite the trapezoidal signal approximation step. We generate a ground truth from these annotations using the *EvalDep* method [12] to serve as a baseline for comparison, which is shown in [15] to yield a good basis for comparison. We also generate a ground truth using the original fusion-first approach [15]

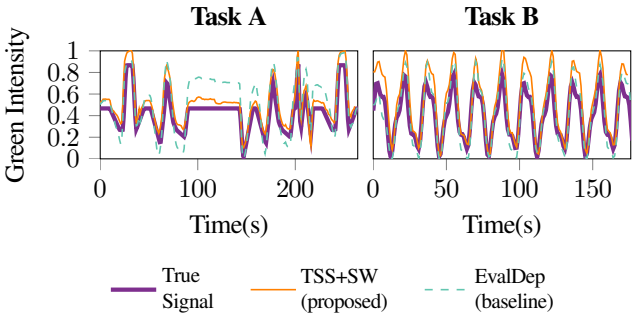

Fig. 3. Plots of the true target signal and ground truth signals produced by the baseline *EvalDep* fusion method and the proposed TSS with signal warping (SW) method.

and compare it to the proposed TSS method. The final stage of the signal warping pipeline (see Figure 1) requires segments from a fused signal to be warped according to the ordinal embedding (see [15] for details). While in principle we believe it would be best to produce these segments from a subset of annotations that align well with the embedding, we use the *EvalDep* fusion in these experiments because the annotations are generally all in agreement.

4.2. Results

Figure 3 plots the different ground truth representations and the true construct signal. Table 1 displays several agreement measures comparing each ground truth method to the true signal in both tasks.

Task	Method	Pearson	Spearman	Kendall's Tau	NMI
A	EvalDep (baseline)	0.90	0.93	0.80	0.88
	TSS+SW(proposed)	**0.96**	0.92	0.80	0.88
	Fusion-first SW	**0.97**	0.94	0.83	0.82
B	EvalDep (baseline)	0.96	0.96	0.83	1.0
	TSS+SW(proposed)	0.96	0.96	0.83	1.0
	Fusion-first SW	0.99	0.99	0.91	0.99

Table 1. Several agreement measures computed for various ground truth techniques showing the agreement between each method's ground truth and the true green intensity. Some values vary from [19] due to down-sampling. SW = signal warping algorithm. NMI = normalized mutual information.

5. DISCUSSION AND FUTURE WORK

Both of the signal warping methods in Table 1 achieve a similar or significantly higher Pearson correlation with the true green intensity signal than the baseline method. Aside from the mutual information score, the other measures show little to moderate improvements for these methods over the baseline, which is consistent with [15] and [19]. These experiments indicate that signal warping with TSS fusion is competitive with other methods.

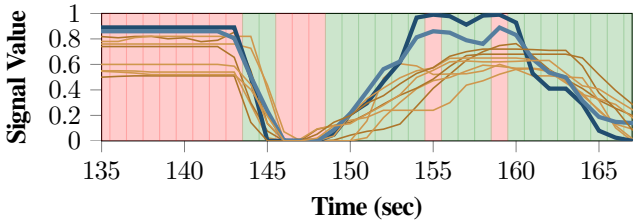

Fig. 4. Plot of Task A annotation signals (yellow and bold blue) and the fused TSS sequence (vertical bands of green or red). Red bands denote no change in the TSS (zero value) and green bands represent some change in the fused TSS (± 1).

The true potential benefits of TSS fusion lie in its rank-based encoding of annotations. It treats each annotation as a sequence of denoised construct value changes and thus allows the annotations to be fused or compared with respect to their local value differences rather than the values themselves. This means the TSS sequences are invariant to the potentially non-uniform construct scale differences between different annotators. Furthermore, the TSS representation enables researchers to study agreement between annotators over interpretable partitions in annotation space. For example, Figure 4 shows annotations from Task A plotted with vertical bands of color for each 1Hz segment representing the fused TSS. Red bands corresponds to zero TSS values (constant regions) and green ones denote non-zeros (increasing or decreasing regions). The first two contiguous subsequences of green show all annotations either decreasing or increasing, so the individual TSS representations would look very similar. The interval between 155s and 160s shows that two annotators (highlighted with bold blue lines) observed a phantom peak before a true peak near the 159s frame. Though this peak does not exist in Task A, the fact that two annotators observed it may reveal something about perception and may warrant further investigation. The TSS representation proposed in this work enables this type of interpretable local inquiry and analysis.

The proposed TSS method for approximating and fusing annotations opens several research avenues. One thread involves exploring different consensus measures defined over the TSS sequences per annotator. This would allow agreement to be measured during time spans between perceptually constant regions of the construct. These measures could be used to locally cluster the annotators or to assign confidence scores to the fused TSS depending on the proportion of annotations in agreement with the ordinal embedding results. Recently, rank-based learning schemes [20] and annotation tools, such as *RankTrace* [21], have been proposed where annotators are asked to rate construct changes instead of assigning values. When only the direction of annotated change is considered, these rank annotations produce TSS-like sequences. Understanding the subtle differences between annotator behavior in value-based versus rank-based annotations is an important subject for further research.

6. REFERENCES

[1] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent", *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2011.

[2] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging", *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[3] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music", *PloS one*, vol. 12, no. 3, e0173392, 2017.

[4] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals", *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2012.

[5] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild", *arXiv preprint arXiv:1901.02839*, 2019.

[6] B. M. Booth, A. M. Ali, S. S. Narayanan, I. Bennett, and A. A. Farag, "Toward active and unobtrusive engagement assessment of distance learners", in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, IEEE, 2017, pp. 470–476.

[7] B. W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. C. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language.", in *Interspeech*, vol. 2016, 2016, pp. 2001–2005.

[8] M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Correlated-spaces regression for learning continuous emotion dimensions", in *Proceedings of the 21st ACM international conference on Multimedia*, ACM, 2013, pp. 773–776.

[9] H. Hotelling, "Relations between two sets of variates", *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[10] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis", in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1247–1255.

[11] M. Müller, "Dynamic time warping", *Information retrieval for music and motion*, pp. 69–84, 2007.

[12] S. Mariooryad and C. Busso, "Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators", *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2015.

[13] F. Zhou and F. De la Torre, "Generalized canonical time warping", *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 279–294, 2016.

[14] G. Trigeorgis, M. A. Nicolaou, S. Zafeiriou, and B. W. Schuller, "Deep canonical time warping", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5110–5118.

[15] B. M. Booth, K. Mundnich, and S. S. Narayanan, "A novel method for human bias correction of continuous-time annotations", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 3091–3095.

[16] A. Metallinou and S. Narayanan, "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, pp. 1–8.

[17] G. N. Yannakakis and J. Hallam, "Ranking vs. preference: A comparative study of self-reporting", in *Affective Computing and Intelligent Interaction: 4th International Conference*, Springer, 2011, pp. 437–446.

[18] G. N. Yannakakis and H. P. Martınez, "Ratings are overrated!", *Frontiers in ICT*, vol. 2, p. 13, 2015.

[19] B. Booth and S. Narayanan, "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, Sep. 2019.

[20] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach", *IEEE Transactions on Affective Computing*, 2018.

[21] P. Lopes, G. N. Yannakakis, and A. Liapis, "Ranktrace: Relative and unbounded affect annotation", in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2017, pp. 158–163.