Fusing Annotations with Majority Vote Triplet Embeddings

Brandon M. Booth
SAIL
University of Southern California
Los Angeles, CA
brandon.m.booth@gmail.com

Karel Mundnich
SAIL
University of Southern California
Los Angeles, CA
mundnich@usc.edu

Shrikanth Narayanan
SAIL
University of Southern California
Los Angeles, CA
shri@ee.usc.edu

ABSTRACT

Human annotations of behavioral constructs are of great importance to the machine learning community because of the difficulty in quantifying states that cannot be directly observed, such as dimensional emotion. Disagreements between annotators and other personal biases complicate the goal of obtaining an accurate approximation of the true behavioral construct values for use as ground truth. We present a novel majority vote triplet embedding scheme for fusing real-time and continuous annotations of a stimulus to produce a gold-standard time series. We illustrate the validity of our approach by showing that the method produces reasonable gold-standards for two separate annotation tasks from a human annotation data set where the true construct labels are known *a priori*. We also apply our method to the RECOLA dimensional emotion data set in conjunction with state-of-the-art time warping methods to produce gold-standard labels that are sufficiently representative of the annotations and also that are more easily learned from features when evaluated using a battery of linear predictors as prescribed in the 2018 AVEC gold-standard emotion sub-challenge. In particular, we find that the proposed method leads to gold-standard labels that aid in valence prediction.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**;

KEYWORDS

Annotation fusion, triplet embeddings, inter-rater agreement

ACM Reference Format:
Brandon M. Booth, Karel Mundnich, and Shrikanth Narayanan. 2018. Fusing Annotations with Majority Vote Triplet Embeddings. In *2018 Audio/Visual Emotion Challenge and Workshop (AVEC'18), October 22, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3266302.3266312

1 INTRODUCTION

Accurate prediction of human behavior and mental states from time series data is of great interest to academic and industrial researchers and is one problem where advancement directly enables technological innovation to promote social wellness. Accurately estimating latent human states such as emotion or attention continues to be a difficult research problem however. One of the most common and promising avenues for approaching it involves training machine learning models to learn a mapping from human behavioral features (e.g. physiologic data, audio, facial expressions, posture, etc.) to a set of labels rating the latent construct of interest. These labels are typically derived from several human annotations and thus the accuracy depends not only on the quality of each individual annotation but also the methods used to fuse them into a single set of labels.

Many researchers have proposed algorithms for fusing annotations to generate a set of labels for use as ground truth in machine learning. Several authors propose using an average of the individual annotations after first performing time alignment to remove artifacts produced by annotation latencies that vary across annotators. Marioooryad et al. [8] demonstrate an improvement in classification performance when utilizing gold-standard labels produced via averaging after first aligning the annotations in time with a uniform shift computed per annotator based on mutual information. Dynamic time warping (DTW) is another popular time-alignment method which monotonically warps time to maximize alignment [9]. Some methods like canonical correlation analysis [3] and correlated spaces regression [10] learn how to warp the fused annotation space so the fusion is more correlated with its associated features. Many combined time and space warping methods have also been proposed such as canonical time warping [16], generalized time warping (GTW) [14], and deep canonical time warping [12].

More recent work explores alternative strategies for computing a gold-standard. Lopes et al. [7] show that in some cases the gradient in an annotation is more informative than the annotation value, which can be exploited to produce a better ground truth. Booth et al. [1] hypothesize that changes in the annotation value over time (trends) are more meaningful than the values and propose a fusion approach using additional comparative information collected from humans to produce a gold-standard.

This list of prior work in annotation fusion is not exhaustive, as many other viable time-alignment and fusion approaches have been and continue to be developed. In aggregate, these algorithms approach annotation fusion from many different angles, but the accuracy of any fused annotation fundamentally cannot be measured when the underlying construct is a latent mental or behavioral state.

In this paper, we present a triplet embedding algorithm for annotation fusion where the annotations produced by distinct annotators vote on the similarity of the target construct between each unique triplet of frames. The model assumes that comparisons between

AVEC'18, October 22, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5983-2/18/10…$15.00
https://doi.org/10.1145/3266302.3266312

annotation values at distinct times are meaningful. Thus it presumes that annotators are able to continuously rate the stimulus in real time with consistency such that the same approximate value is assigned whenever the same assessment is made at two distinct points in time, after time-alignment. Some anecdotal evidence in [1] suggests that consistency may not be preserved during continuous real-time annotation, but we aim to show that making this simplifying assumption still produces quality annotation fusions.

We preliminarily test our idea on a data set presented by Booth et al. [1] where a truth signal is known *a priori* to show that our majority vote triplet embedding approach yields a sensible fused annotation. We then apply the method to the RECOLA dimensional emotion data set [11] as part of the 2018 AVEC gold-standard emotion sub-challenge to produce better gold-standard labels for dimensional emotion. Our findings suggest that our proposed gold-standard labels can improve emotional valence prediction from the associated features while having a negligible impact on arousal prediction.

2 TRIPLET EMBEDDINGS

Let $z_1, \ldots, z_n$ be the items that we want to represent with points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^m$, respectively. The z_i items do not necessarily lie in a metric space, but we assume we have some form of a dissimilarity measure $d(z_i, z_j)$ among them. This dissimilarity may be a perceptual model that may not be mathematically defined, such as the dissimilarity of items' arousal or valence as perceived by an annotator. We are provided a set $\mathcal{T}$ of possibly noisy and incomplete triplet comparisons such that

$$\mathcal{T} = \{(i,j,k) \mid i \neq j \neq k \neq i,\ d(z_i, z_j) < d(z_i, z_k)\} \tag{1}$$

which we can use to construct an embedding $\boldsymbol{X} \in \mathbb{R}^{n \times m}$. Each row $\boldsymbol{x}_l^T$ of $\boldsymbol{X}$ represents a point in m-dimensional Euclidean space for each z_l, respectively. In this paper, we consider the indices $l = \{1, \ldots, n\}$ as frame offsets in a regularly sampled time series, so that $\boldsymbol{x_l}$ is the value that the signal takes at time index l.

2.1 Stochastic Triplet Embeddings

Various different techniques have been proposed to find the embedding $\boldsymbol{X}$. For an extensive list, please refer to [6]. In this paper, we use t-Student stochastic triplet embeddings (tSTE) because of its proven empirical performance [5] and suitability for recovering 1-dimensional embeddings. As the authors highlight in [13], tSTE aggregates similar points and repels dissimilar ones leading to simpler embedding solutions.

tSTE defines a value p_{ijk} that a certain triplet $(i,j,k) \in \mathcal{T}$ is satisfied under a stochastic selection rule:

$$p_{ijk} = \frac{\left(1 + \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1 + \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{\alpha}\right)^{-\frac{\alpha+1}{2}} + \left(1 + \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_k\|_2^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}, \tag{2}$$

where α regulates the thickness of the tails of the t-Student kernels in p_{ijk}. The goal is to maximize the log-probabilities through all the triplets in $\mathcal{T}$ to find the embedding $\boldsymbol{X}$:

$$\max_{\boldsymbol{X}} \sum_{(i,j,k) \in \mathcal{T}} \log p_{ijk} \tag{3}$$

The difficulty of solving this problem lies in the fact that using t-Student kernels produces an objective function that is non-convex (a sum of quasi-concave functions $\log(p_{ijk})$). In tSTE, this optimization problem is solved using gradient descent with random initializations.

When tSTE is applied to one-dimensional time series, it can produce single frame signal spikes in the embedding corresponding to a local optimum in the non-convex likelihood function. Since we aim to use the embedding directly for annotation fusion, we need to avoid these types of high frequency spikes in the solution. In our experiments, these sub-optimal solutions occur when the random initial signal already contains high frequencies, meaning that $\boldsymbol{x}_i$'s are initialized far away from their desired location. In practice, we find that initializing the optimization routine with an educated guess (i.e. a signal similar to the desired embedding $\boldsymbol{X}$) avoids these kinds of local optima.

Each obtained embedding $\boldsymbol{X}$ is invariant to monotonic transformations of the relative metric distances between points (i.e. rotations and scaling) [4, 6]. In the case of a one-dimensional embedding ($m = 1$), only the shift and scale need to be corrected (using an affine transformation). In scenarios where the target signal is known *a priori*, isotonic regression can be employed to learn a monotonic transformation which optimally maps the embedding to the target. We propose using triplet embeddings for annotation fusion where there is no known truth, so we orient and scale the embedding by applying an affine transformation so that its mean squared error is minimized with respect to the simple average of all annotations.

3 MAJORITY VOTING TRIPLET EMBEDDINGS

In this paper, we explore the idea that although the values provided during continuous real-time annotation may not be directly reliable, their relative pairwise distances are reliable throughout an annotation scheme for the majority of the annotators. Therefore, for all annotators, we hypothesize that we can obtain the correct direction of the triplet relation

$$d(z_i, z_j) \overset{?}{\leqslant} d(z_i, z_k) \tag{4}$$

by taking all annotators' opinions for any $(i,j,k) \in \mathcal{T}$. Here, $d(z_i, z_j)$ represents the unknown true dissimilarity of the construct being annotated between time i and time j (note that i and j are not necessarily instants in time, but refer to short time windows). This leads to a natural weighted majority vote construction to decide which direction in Eq. 4 is correct.

We consider all possible triplets (i,j,k) and assign weights to each individual annotators response:

$$\begin{aligned}
\text{Decision of } \mathcal{A}_1 &: \ d_1(z_i, z_j) < d_1(z_i, z_k) \in \mathcal{T} \Rightarrow w_1^{<} \\
\text{Decision of } \mathcal{A}_2 &: \ d_2(z_i, z_j) < d_2(z_i, z_k) \in \mathcal{T} \Rightarrow w_2^{<} \\
&\vdots \\
\text{Decision of } \mathcal{A}_{r-1} &: \ d_{r-1}(z_i, z_j) > d_{r-1}(z_i, z_k) \notin \mathcal{T} \Rightarrow w_{r-1}^{>} \\
\text{Decision of } \mathcal{A}_r &: \ d_r(z_i, z_j) < d_r(z_i, z_k) \in \mathcal{T} \Rightarrow w_r^{<}
\end{aligned}$$

We use w_a ($a \in \{1, 2, ..., r\}$) to denote the weight of annotator a relative to other annotators, and we index the dissimilarity functions $d_a(\cdot, \cdot)$ as a reminder that each annotator perceives events differently. Each annotator's weight is assigned beforehand using one of many techniques described in the next subsection, but intuitively each weight is proportional to the trust that we assign to the corresponding annotator. Then, if

$$\sum_r w_a^< > \sum_r w_a^>, \tag{5}$$

we conclude that $(i, j, k) \in \mathcal{T}$.

Algorithm 1 shows the implementation for the triplet generation through weighted majority voting. The implementation takes a matrix $A \in \mathbb{R}^{n \times r}$, where each column represents an annotation time series from one of r annotators, and the rows represent time frames. The implementation takes a weight vector $w \in \mathbb{R}^r$, representing the *trust* in each annotator. The implementation is flexible enough that setting $w_a = 0$ will remove annotator a from the decision process (leave-one-annotator-out), while recovering a simple majority vote if w_a is constant for all annotators.

Algorithm 1: Generate set of triplets $\mathcal{T}$ using annotator weights and a majority vote strategy.

Data: $A \in \mathbb{R}^{n \times r}$: Annotations matrix
Input: weights $\in \mathbb{R}^r$: Annotator weights
Result: Set of triplets $\mathcal{T}$

```
triplets ← [];
for a ← 1 to r do
    /* Compute all pairwise distances between
       values of each column of A. Each D[a] is a
       distances matrix between all points in each
       A[:, a].                                   */
    D[a] = distances(A[:, a]);
end
for k ← 1 to n do
    for j ← 1 to k − 1 do
        for i ← 1 to n do
            // Iterate over unique triplets (i, j, k)
            if i ≠ j and i ≠ k then
                w^< = 0;
                w^> = 0;
                for a ← 1 to r do
                    // Compute each annotator decision
                    // for each unique triplet (i, j, k)
                    if D[a][i, j] < D[a][i, k] then
                        w^< += weights[a];
                    else if D[a][i, j] < D[a][i, k] then
                        w^> += weights[a];
                    end
                end
                if w^< > w^> then
                    triplets.append((i, j, k));
                else if w^< < w^> then
                    triplets.append((i, k, j));
                end
            end
        end
    end
end
```

(a) AVEC arousal dev_7 (b) AVEC valence train_2

Figure 1: Agreement between annotators for two example subjects from the RECOLA emotion dataset in two different annotation tasks: arousal and valence. Agreement is measured using CCC. The overall agreement in valence is higher than the overall agreement for arousal.

An interesting feature of this majority vote embedding approach to annotation fusion is that once $\boldsymbol{X}$ is computed, the number of triplets in $\mathcal{T}$ that violate the distances in $\boldsymbol{X}$ may be computed, leading to a measure of agreement in the construction of the embedding itself. We revisit this idea in our discussion in Section 6.

3.1 Annotator Weights

We test three approaches to assigning weights to annotators: *unweighted*, *weighted*, and *weighted leave-one-out*. In the *unweighted* scenario, each annotator is given an equal weight w_a so each triplet is given an equal vote. In the *weighted* scenario, we use the concordance correlation coefficient (CCC) to assess the similarity between two annotators, which is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{6}$$

We obtain a matrix of correlations $\boldsymbol{S}$ such as in Fig. 1 and then set $w_a = \sum_i S(a, i) - 1$. Thus annotations that agree with each other are given higher weights and annotations in disagreement with the majority (potentially adversarial) are given less weight. Lastly, we adopt a *weighted leave-one-out* strategy where the annotator with the lowest weight is left out entirely and the embedding depends only on the remaining annotations.

4 EXPERIMENTS

We apply our majority vote triplet embedding method in two scenarios. First, we empirically motivate our approach by using it to fuse annotations in a designed experiment where the construct of interest is known *a priori*. We use this test to show that this embedding algorithm produces intuitive gold-standard labels qualitatively similar to the average of annotations. Then we apply the method to continuous annotations from the RECOLA data set using

various weighting schemes and also using well-established time series warping methods to try to produce a gold standard fusion that best approximates both arousal and valence.

4.1 Color Intensity Annotations

In a human annotation experiment conducted by Booth et al. [1], human annotators were asked to watch a video containing frames of a single color at different intensities and then rate the intensity in real-time on a continuous scale. In this experiment, the true color intensity was known, so the annotations and any proposed fusion could be directly compared to the true signal. We apply our majority vote triplet embedding algorithm to the individual annotations in both annotation tasks from this data set to produce gold-standard labels which we evaluate in Section 5.1.

4.2 RECOLA Emotion Annotations

The RECOLA data set contains real-time human annotations of dimensional affect (valence and arousal) on a continuous scale. Each emotional dimension is separately annotated, so we employ our triplet embedding method separately for each.

Since the true valence and arousal signals are unknown, we use a different means for comparing our proposed gold-standard annotation fusions per the AVEC 2018 gold-standard emotion sub-challenge guidelines. The underlying assertion made to establish an evaluation criteria is that good gold-standard labels minimize the unexplained variance among the individual annotations and also are easier to learn from the features using simple models. A variety of unimodal and multi-modal linear regression models with different regularizers are trained on all available features in the RECOLA data set (physiologic, video-based, and audio-based) with the gold-standard serving as labels, then each model attempts to reconstruct the gold-standard as accurately as possible. The gold-standard is compared to its projection into each model's representation space using CCC which provides a measure of the "quality" of the gold-standard. The baseline paper for the challenge describes these assumptions and regression models in more detail [2].

Our proposed majority vote triplet embedding approach results in non-linear spatial warping of the annotations, but makes no adjustments in time. Therefore, we test it with two state-of-the-art temporal warping methods to achieve better time alignment of annotations and features: dynamic time warping (DTW) [9] and generalized time warping (GTW) [14].

Dynamic time warping adjusts the sequence of temporal indices of one signal given another reference signal to achieve an optimal correlation between the two. Since the true emotional signal is unknown, we elect to use a single feature as a reference signal and then apply DTW to each individual annotation to align it with the reference feature. We perform an exhaustive search of all corresponding video, audio, and physiologic features provided in the RECOLA data set to find the single feature with the highest average Pearson correlation with each corresponding annotation, which turns out to be the geometric video-based feature "geometric_feature_245_amean". Fig. 3 shows an example annotation of arousal before and after using DTW for temporal alignment to this reference feature. We use Python's *dtw* package to apply DTW.

Generalized time warping is an enhanced version of canonical time warping which attempts to learn a monotonic temporal warping function and feature projection function that together maximize the correlation between the projected features and the temporally warped signal. We use the Matlab code provided by [15] to implement GTW and perform a grid search for tuning d, the CCA energy threshold hyperparameter, by maximizing the CCC obtained when evaluating using the AVEC challenge gold-standard evaluation metric. We find the performance is not impacted significantly for $d \in \{0.6, 0.7, 0.8, 0.9, 0.95\}$ and that $d = 0.7$ is slightly better.

Since the number of triplets necessary to fully specify each annotation's vote scales $O(n^3)$ with the number of frames, we first downsample the annotations from 25Hz to 1Hz using a polyphase FIR filter. This leads to 13,365,300 unique triplets for a signal with 300 samples. Once the optimal embedding is obtained, we upsample to the original sampling rate. These resampling steps are performed in Julia using the DSP.jl package. Furthermore, both DTW and GTW require that the features and annotations are temporally aligned. Thus, when applying these two methods, we downsample the annotations by a factor of 10 to match the sampling rate of the features. After running the time warping method, we piecewise linearly interpolate the monotonic temporal map to provide time indices at the original resolution.

5 RESULTS

We present the results from our color intensity validation experiment and our application of our proposed fusion technique to the RECOLA data set using the 2018 AVEC gold-standard emotion sub-challenge evaluation measure to judge the quality of each proposed fused annotation.

5.1 Color Intensity Validation

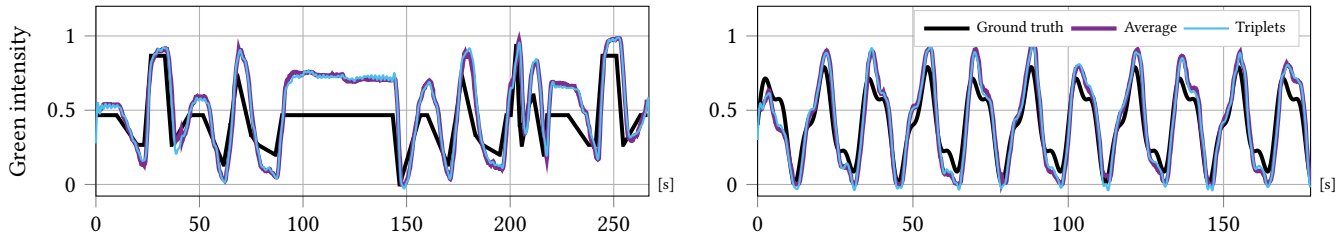
Fig. 2 shows the resulting proposed gold standard from our majority vote triplet embedding on the color intensity data set. Table 1 shows the correlations between our proposed approach and the average of the individual annotations without time alignment.

5.2 RECOLA Emotion Annotations

The fused annotations serving as proposed gold-standards for the RECOLA emotion annotations are evaluated using CCC as described in Section 4.2. Table 2 shows the correlations of our results compared to the AVEC challenge baseline algorithm (for details see [2]). Fig. 1 shows example annotator agreement matrices used to

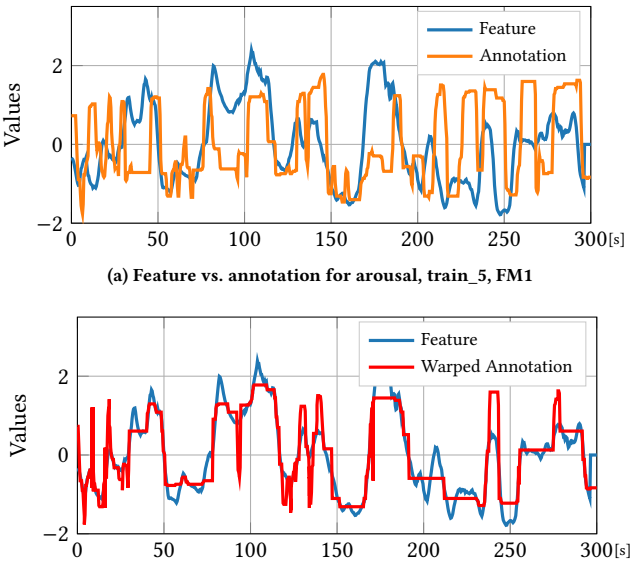
Table 1: Agreement measures between different fusion methods and the objective truth signal in two tasks from the color intensity data set [1].

	Method	RMSE	Pearson	CCC
TaskA	Unweighted Average	0.1916	0.7756	0.6392
	Triplet Embedding	0.1907	0.7762	0.6410
TaskB	Unweighted Average	0.1057	0.9523	0.9172
	Triplet Embedding	0.1005	0.9594	0.9248



(a) Task A (b) Task B

Figure 2: Plots for TaskA (left) and TaskB (right) from the color intensity annotation dataset. The true color intensity signal is shown (black) alongside the unweighted average of individual annotations (purple) and a gold-standard produced using an unweighted version of our triplet embedding algorithm (light blue). This shows the proposed method producing the gold-standard (triplets) is sensible and qualitatively similar to the average signal.



(a) Feature vs. annotation for arousal, train_5, FM1

(b) Feature vs. warped annotation for arousal, train_5, FM1

Figure 3: An example plot of an individual annotation (from FM1) of arousal from the RECOLA data set [11] for subject "train_5" and the corresponding feature with the highest average Pearson correlation with all annotations (geometric_feature_245_amean). The two signals are shown in (a) and the DTW-warped annotation is shown in (b).

compute the weights as described in Section 3.1. Fig. 4 plots the percentage of triplet violations remaining after running the majority vote triplet embedding algorithm to convergence. Since we initialize the triplet optimization subroutine with a signal very close to the average of individual annotations, the resulting embedding is extremely stable over multiple trials. Thus, we only plot the triplet violations for one trial run per annotation task.

6 DISCUSSION AND FUTURE WORK

It is clear from the results in our validation experiment in Table 1 that our method produces gold-standard labels comparable to simple unweighted averages. This empirically helps illustrate that even though our majority vote triplet embedding technique is more complicated, it produces sensible annotation fusions.

The results from Table 2 comparing different gold-standard proposals on the RECOLA emotion data set are somewhat surprising. None of the triplet embeddings are able to surpass the baseline's arousal score. As much research in the past has noted, machine learning on human-produced labels for arousal typically outperforms valence, which is reflected in our experiments by the CCC scores. But, neither the non-linear time nor space warping effects of the triplet embedding seem to improve the learnability of emotional arousal. The drop in CCC for arousal in most cases is quite small, however, and dominated by the improved valence correlations. Even when no time warping is performed, the triplet embedding improves over the baseline method's CCC score for valence substantially.

Additionally, when the most adversarial annotator is left out (there is usually only one in the RECOLA data), DTW performs quite well. This indicates that our embedding approach applied to DTW-warped annotations is very sensitive to annotation outliers, which may in part be due to our choice to use a single feature as a reference signal for the time warping. GTW can accommodate all features and works well in our tests here, so we consider this to be a more robust approach when time warping is used.

A couple of unexpected results come from our experiments on the RECOLA data set. First, the performance difference between using and not using GTW when applying majority vote triplet embeddings is negligible in spite of the noticeable difference in the temporal warping shown in Fig. 3). Secondly, the number of violated triplets at convergence shown in Fig. 4 is less in the *weighted* and *unweighted* DTW cases, but the CCC values associated with them is often worse than the other methods. So, the relationship between the quality of a gold-standard and the quality of an embedding cannot be easily inferred from the number of triplet violations. We save further investigation of these two observations for future work.

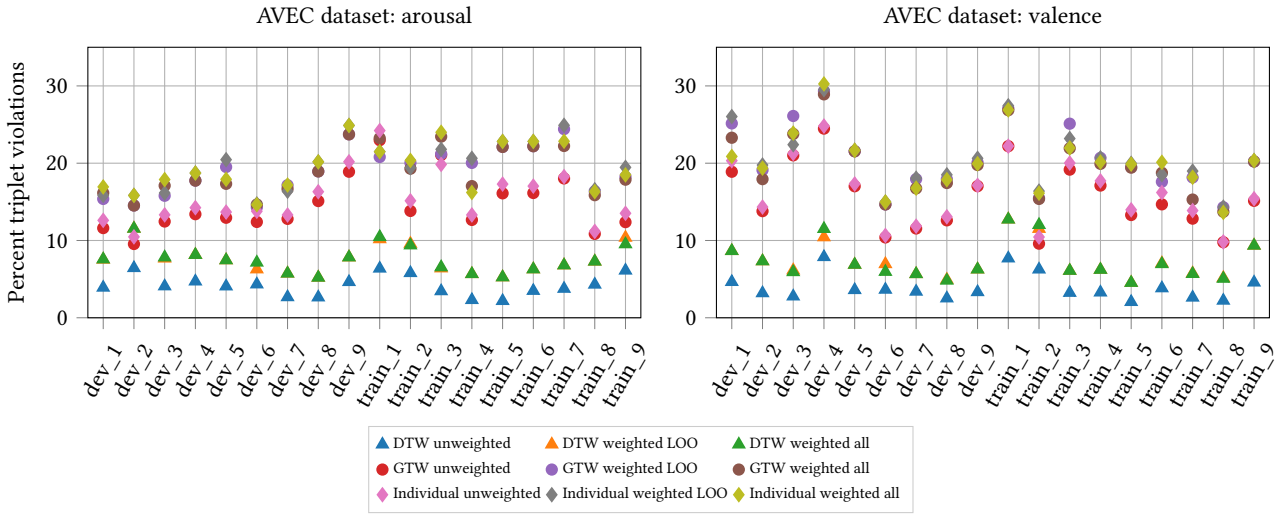


Figure 4: Plots showing the percentages of triplet violations after tSTE convergence

Table 2: CCC values for various proposed gold-standard annotation fusions using the 2018 AVEC emotion sub-challenge evaluation metric explained in Section 4.2

Annotation Scheme	Prediction Model	Baseline		Triplet Embedding					
		Arousal	Valence	Arousal	Valence	DTW arousal	DTW valence	GTW arousal	GTW valence
Unweighted	Unimodal	N/A	N/A	0.751	0.561	0.469	0.559	0.751	0.561
	Multimodal-Multi Rep	N/A	N/A	0.751	0.693	0.679	0.687	0.751	0.696
	Multi-Rep	N/A	N/A	0.751	0.676	0.669	0.674	0.751	0.676
	Multimodal Hierarchic	N/A	N/A	0.751	0.685	0.674	0.677	0.751	0.683
Weighted	Unimodal	0.760	0.506	0.753	0.561	0.576	0.561	0.751	0.561
	Multimodal-Multi Rep	0.760	0.506	0.753	0.691	0.687	0.693	0.751	0.692
	Multi-Rep	0.772	0.506	0.753	0.671	0.673	0.678	0.751	0.676
	Multimodal Hierarchic	0.775	0.570	0.753	0.685	0.684	0.677	0.751	0.686
Weighted Leave One Out	Unimodal	N/A	N/A	0.753	0.561	0.753	0.561	0.751	0.561
	Multimodal-Multi Rep	N/A	N/A	0.753	0.692	0.753	0.691	0.751	0.690
	Multi-Rep	N/A	N/A	0.753	0.676	0.753	0.676	0.751	0.676
	Multimodal Hierarchic	N/A	N/A	0.753	0.685	0.753	0.685	0.751	0.686

Though we have empirically shown the success of the majority vote triplet embedding method for annotation fusion, it is still difficult to reason why it works well for valence prediction. The method makes a bold assumption about the consistency in each annotator's ability to assign similar values to different moments when the construct of interest is comparable. One idea worth exploring is restricting the triplet inference to frames in close temporal proximity. This would effectively reduce the number of triplets in $\mathcal{T}$ to a potentially more reliable subset and hopefully lead to a more accurate gold standard especially for lengthy annotations where human fatigue is expected to impact individual annotation quality.

7 CONCLUSION

In this paper we present a novel majority vote triplet embedding scheme for annotation fusion. We test the viability of the method in a human annotation experiment where the true annotation time series is known *a priori* and show that it can be used to fuse annotations. We also apply the proposed method on the human annotations in the RECOLA emotion data set both with and without two state-of-the-art time warping methods to propose gold-standard labels of dimensional emotion. We evaluate these proposed gold-standards using a comparison algorithm from the 2018 AVEC gold-standard emotion sub-challenge and show that it achieves similar or better correlations, especially for emotional valence, which suggests it is more suitable as a gold-standard than the weighted average method serving as the baseline fusion technique.

REFERENCES

[1] Brandon M Booth, Karel Mundnich, and Shrikanth S Narayanan. 2018. A Novel Method for Human Bias Correction of Continuous-time Annotations. (2018). Accepted for publication in the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[2] Fabien Ringeval and Björn Schuller and Michel Valstar and Roddy Cowie and Heysem Kaya and Maximilian Schmitt and Shahin Amiriparian and Nicholas Cummins and Denis Lalanne and Adrien Michaud and Elvan Çiftçi and Hüseyin Güleç and Albert Ali Salah and Maja Pantic. 2018. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*, Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic (Eds.). ACM, Seoul, Korea.

[3] Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28, 3/4 (1936), 321–377.

[4] Lalit Jain, Kevin G Jamieson, and Rob Nowak. 2016. Finite Sample Prediction and Recovery Bounds for Ordinal Embedding. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2711–2719.

[5] Kevin G Jamieson, Lalit Jain, Chris Fernandez, Nicholas J Glattard, and Rob Nowak. 2015. Next: A system for real-world development, evaluation, and application of active learning. In *Advances in Neural Information Processing Systems*. 2656–2664.

[6] Matthäus Kleindessner and Ulrike von Luxburg. 2014. Uniqueness of Ordinal Embedding.. In *COLT*. 40–67.

[7] Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. 2017. RankTrace: Relative and unbounded affect annotation. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 158–163.

[8] 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.

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

[10] Mihalis A Nicolaou, Stefanos Zafeiriou, and Maja Pantic. 2013. Correlated-spaces regression for learning continuous emotion dimensions. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 773–776.

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

[12] George Trigeorgis, Mihalis A Nicolaou, Bjorn W Schuller, and Stefanos Zafeiriou. 2018. Deep canonical time warping for simultaneous alignment and representation learning of sequences. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 5 (2018), 1128–1138.

[13] Laurens Van Der Maaten and Kilian Weinberger. 2012. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*. IEEE, 1–6.

[14] Feng Zhou and Fernando De la Torre. 2012. Generalized time warping for multimodal alignment of human motion. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 1282–1289.

[15] 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.

[16] Feng Zhou and Fernando Torre. 2009. Canonical time warping for alignment of human behavior. In *Advances in neural information processing systems*. 2286–2294.