A NOVEL METHOD FOR HUMAN BIAS CORRECTION OF CONTINUOUS-TIME ANNOTATIONS

Brandon M. Booth *Karel Mundnich* *Shrikanth S. Narayanan*

Signal Analysis and Interpretation Laboratory, University of Southern California
3740 McClintock Ave, EEB 400, Los Angeles, CA 90089

ABSTRACT

Human annotations are of integral value in human behavior studies and in particular for the generation of ground truth for behavior prediction using various machine learning methods. These often subjective human annotations are especially required for studies involving measuring and predicting hidden mental states (e.g. emotions) that cannot effectively be measured or assessed by other means. Human annotations are noisy and prone to the influence of several factors including personal bias, task ambiguity, environmental distractions, and health state. We propose a novel method for fusion of continuous real-time human annotations to generate accurate ground truth estimates. We introduce a signal warping method that uses additional comparative rank-based information about specific subsets of the annotations to correct for specific types of human annotation artifacts. This approach is validated using a mechanically simple but perceptually demanding psychophysical annotation experiment where objective truth labels are known. Our method yields ground truth estimates that are in better agreement with the objective truth than state-of-the-art approaches.

Index Terms— Annotation fusion, total variation denoising, ground truth

1. INTRODUCTION

Estimation of human mental states and traits that are not readily observable, such as emotional state, engagement, productivity, and attention is notoriously difficult. For these types of problems, self or observer-based annotations are often used to provide ratings for the target behavioral or experiential construct and establish ground truth labels for machine learning. However, the human annotation process is noisy and produces several types of artifacts in the labels due to factors such as perception bias, interpretation ambiguity, and distractions, to name a few [1]. The impact of these biases and cognitive effects on the annotations are magnified when the annotation task is complex or when it demands careful attention and vigilance over long sessions.

The usual strategy for combating these error sources involves gathering multiple annotations from different persons and fusing them to obtain a single ground truth thus mitigating the effect of individual biases and annotation artifacts, but there is yet no consensus on a best-practice fusion approach. Prior work on continuous annotation fusion has focused on ground truth estimation by modeling and removing different specific sources of annotation lag, noise, and/or artifacts. One evaluator-dependent approach [2] finds an optimal time shift for individual 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 [3] is a well-known alignment solution that maximizes the agreement between annotators by adjusting for the variance in each annotators' lag time, but it only corrects for temporal misalignments during fusion. Canonical correlation analysis (CCA) [4] and correlated spaces regression (CSR) [5] 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. Many other feature-based methods have been proposed [6, 7, 8, 9, 10] and perform reasonably well on various data sets, but require an informative set of features to be extracted from the stimuli for alignment and fusion. In cases where annotator fatigue or external distractions occur, annotation artifacts may not correlate with these features and affect the quality of the resulting ground truth estimation.

Several studies have shown that people are better at comparative ranking than absolute rating [1, 11, 12] suggesting that continuous annotations may not exhibit coherence and self-consistency. In this paper we present a correction method, applied after continuous annotation fusion, that uses relative rank information about subsets of the fused annotation to generate a ground truth closer to the objective truth. This rank information is collected in a separate annotation session. Several well-studied data sets involving continuous human annotation already exist [13, 14, 15], but the target construct in each has no objective truth for comparison. So, we present and examine the results from our own simple continuous annotation experiments and show that our proposed method can be used to improve the quality of the ground truth.

2. EXPERIMENT

In this study we use a simple but perceptually challenging annotation task where the objective truth is known. Ten annotators were asked to separately rate the intensity (luminance) of the color green in two videos in real-time and on a continuous scale by adjusting a standard user-interface slider widget.

Our thanks to the NSF and NIH for funding this research.

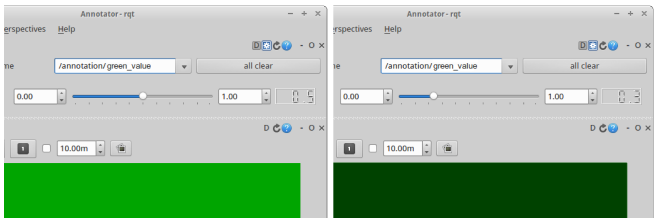


Fig. 1: A closeup snapshot of the user interface at different times during the green intensity annotation task. Annotators only adjusted the slider in sync with changes in the green video.

The same physical computer, monitor, and lighting conditions were used for all ten annotators. The videos were less than five minutes in length, 864x480 resolution, and comprised entirely of solid color frames of green at varying green channel intensities in RGB color space. In Task A's video, the green intensity changed at different speeds and times while avoiding discontinuous jumps and was designed to test annotator rating accuracy. Task B's video featured a perturbed slow oscillation of the green intensity and was chosen to test consistency in annotation over time. The annotation process was devised to be mechanically undemanding with a simple responsive interface to help ensure the main annotation challenge lay in the translation of perceived green intensity to annotation rating. A picture of the annotation interface is shown in Fig 1.

Fig 2 shows a plot of all ten annotations alongside the objective truth for these two annotation tasks. Intra-class correlation measures were computed to estimate annotator agreement per the guidelines in [16] and achieved approximately 0.97 at a 95% confidence interval for each task earning an *excellent* agreement rating according to [17]. The ICC values were calculated using the *psych* package version 1.6.9 in R using a mean-rating (k=3), consistency, two-way mixed effects model.

Although the annotator agreement measure is very high and Fig 2 shows that annotators were generally quite good at capturing large-scale changes and trends, they still had difficulties in other areas. First, annotators tended to over-shoot the target value when annotating increases or decreases in value over a period of time such as in Fig 2a between 200 and 250 seconds. This indicates they were perhaps fixated on annotating the rate of change rather than the actual rating. Secondly, we note that approximately half of the annotators struggled to capture the lack of change in green intensity especially during the 100 to 150-second time interval in Fig 2a. One possible explanation is that the longer duration of this constant segment gave annotators time to realize their current intensity ratings did not match their perception and then adjust the value to match in spite of what was (not) occurring in the video. Lastly, we note that similar green intensities were annotated inconsistently over time. In particular, there was a significant difference in average annotation value per and within annotators at different time intervals where the green intensity was actually at a constant 0.5 value (see Fig 2a). This last observation implies that even for this relatively simple annotation task, annotators struggled to accurately capture the trends while preserving self-consistency over time.

3. FUSED ANNOTATION CORRECTION

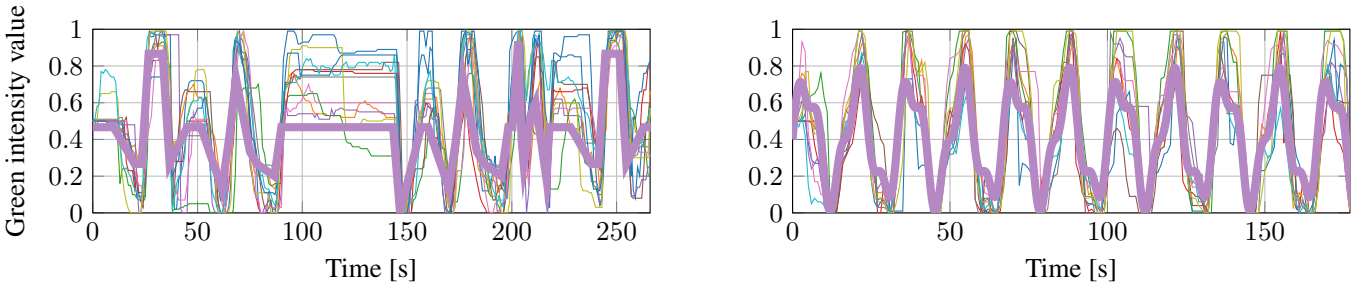
We propose a method for warping fused annotations to establish a ground truth signal that has been corrected for various global inconsistencies, artifacts, and errors introduced during the real-time continuous human annotation process. The method leverages a recurring observation that annotators more successfully capture trends and less accurately represent exact ratings [1, 11, 12]. In our approach, additional information in the form of similarity comparisons between unique time segments of the video must be collected from annotators after the continuous annotation task (these supplementary annotations are simulated in this work). We then leverage the structure of the fused annotations to identify peaks, valleys, and spans of time where the target construct does not appear to change and we only collect comparisons corresponding to these segments. Our method is summarized as a sequence of steps:

1. Annotation Fusion
2. Total variation (TV) denoising
3. Constant interval extraction
4. Triplet comparison collection
5. Ordinal embedding
6. Fused annotation warping

The first step fuses the raw annotations together to form a single time series, and, in principle, any existing annotation fusion method could be used at this stage. Total variation (TV) denoising is then used to approximate the fused signal as a piecewise-constant step function in order to facilitate the identification of segments in time where the target construct does not change noticeably (peaks, valley, and plateaus). Nearly constant intervals of the fused signal are extracted yielding these time segments and then additional rank information is procured from annotators to re-evaluate the proper sorting of these constant 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 fused 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 and Fig 3 shows the results after applying this technique.

3.1. Annotation Fusion

The first step involves fusing the annotations into a single representative signal. Many methods have been proposed for this [2, 3, 6, 7, 8, 18, 19] and in principle any choice works for this step. We use a simple per-annotator time shift (*EvalDep*) proposed by Mariooryad et al. [2]. This method requires some feature sequences to be extracted from the video for alignment, so we provide the green intensity and its forward difference per frame. These particular features allow for a nearly optimal time alignment, but in our tests the proposed method improves ground truth estimation even without time alignment. After shifting each annotation by its own lag estimate (approximately 1.6 seconds each), we truncate the trailing frames so all annotations are equal length and then average them in time.



(a) Task A: Objective truth and real-time annotations

(b) Task B: Objective truth and real-time annotations

Fig. 2: Plots of the objective truths (bold) and annotations of green channel intensity from ten annotators in two separate tasks.

3.2. Total Variation Denoising

Total variation (TV) denoising has been successfully used to remove salt and pepper noise from images while simultaneously preserving signal edges [20]. In our context, we want to identify the set of peaks and valleys where the annotation rating may be inaccurate, and we also want to find the set of nearly constant regions of the fused annotation signal corresponding to a lack of noticeable change in the target construct.

We use the TFOCS MATLAB library [21] to find a new vector y that approximates the fused annotation x by minimizing:

$$y = \min_{y} \Big[\sum_{t} \|x_t - y_t\|_{\ell_2}^2 + \lambda \sum_{t} \|y_{t+1} - y_t\|_{\ell_1} \Big]$$

The parameter λ controls the influence of the temporal variation term and degree to which y is approximately piecewise-constant. For this study, we hand-tune λ by increasing it by multiples of ten from a very tiny value (e.g. 10^{-8}) until it first starts appearing piecewise constant and we settle on the value 0.05. In theory, this parameter can be automatically selected based on other criteria and heuristics, but we leave this endeavor for future work.

3.3. Constant Interval Extraction

A simple heuristic is used to extract nearly constant intervals from the TV-denoised signal. In this step, we scan the TV-denoised signal to find the smallest set of (largest) intervals where each interval satisfies two criteria: (1) the total height does not exceed threshold h, and (2) the frame length of the interval is at least T frames.

We use a height threshold $h = 0.003$ and a minimum frame count threshold $T = 17$ (recall the videos are 30Hz). We choose h to be quite small relative to the annotation scale and find that for well-TV-denoised signals the method is not very sensitive to this parameter. The T parameter is selected to be the smallest value greater than 10 (average human reaction time) that produces a manageable number of intervals. The number of triplets for n intervals grows $O(n^3)$ so we minimize the interval count.

3.4. Triplet Comparisons

In this step during an actual experiment, annotators are asked to compare three extracted video segments corresponding to each unique triplet of constant intervals. One video segment serves as a reference and the other two as test candidates and the annotator is instructed to select which of the two candidate video segments is most similar to the reference. We simulate these comparison results in this study using the objective truth as an oracle.

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. For our application, we are interested in the case where the ordinal comparisons are given in triplet form. Given a set of inputs $\mathcal{Y} = \{y_1, ..., y_n\}$ with each $y \in \mathbb{R}^m$ and a set of similarity relations on 3-tuples from $\mathcal{Y}$ of the form $s(y_i, y_j) < s(y_i, y_k)$ where $\{i, j, k\}$ is a 3-subset of $\{1, 2, ..., n\}$, the goal is to find a set $\mathcal{Z} = \{z_1, ..., z_n\}$ with each $z \in \mathbb{R}^d$ such that:

$$\|z_i - z_j\| < \|z_i - z_k\| \iff s(y_i, y_j) < s(y_i, y_k)$$

for some norm on $\mathcal{Z}$. These triplet comparisons express a similarity relationship where sample i is more similar to sample j than k. Collecting comparisons from humans over triplets has been studied and proven useful in other works [22].

In our method, ordinal embedding is used to reorder the constant intervals to try to sort them properly with respect to the objective truth. Many ordinal embedding solvers over triplets have been proposed [22, 23]. We employ the t-stochastic triplet embedding (t-STE) approach [22] because, as the authors highlight, it aggregates similar points and repels dissimilar ones leading to simpler solutions. Fig 3 shows the embedding results for the extracted constant intervals on Task A that have been rescaled to the proper $[0, 1]$ range and computed using all possible triplet comparisons from the oracle. Note that the embedding only preserves the relative similarity relationships, so it is expected to be off by a (unknown) monotonic transformation of the objective truth.

3.6. Spatial Warping

In the final step, the fused annotation is spatially warped to rectify inconsistencies using the ordinal embedding results.

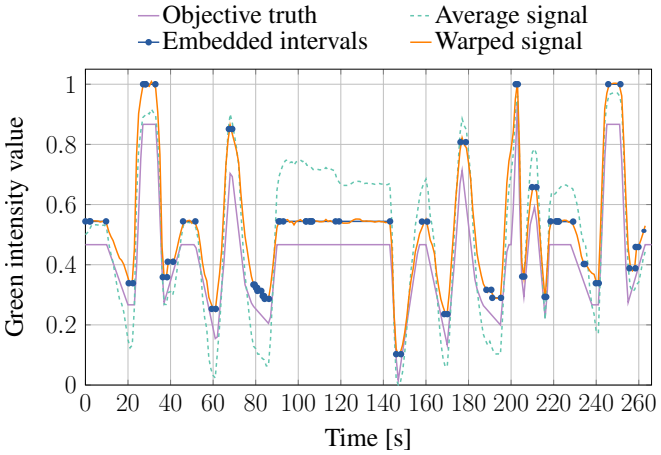


Fig. 3: Plot of the objective truth signal, time-shifted average annotation signal (EvalDep), warped signal (proposed), and the 1-D embedding for extracted constant intervals for Task A. The spatially warped signal better approximates the structure of the objective truth and also achieves greater self-consistency over the entire annotation duration.

Within the time frame of each interval, the fused annotation is shifted so its average over the interval is equal to its corresponding embedding value. The fused annotation between each constant interval is offset and linearly scaled to align with its neighboring repositioned constant intervals. We select a linear inter-interval warping function because it minimally distorts the signal. A formal definition is given in Fig 4.

$$\mathcal{I}_i = \begin{cases} \{t : \min(C_i) \le t \le \max(C_i)\} & i \in \{1, 2, ..., |C|\} \\ \{0\} & i = 0 \\ \{T\} & i = |C| + 1 \end{cases}$$

$$\mathcal{S}_i = \begin{cases} \mathcal{E}_i - \frac{1}{|\mathcal{I}_i|} \sum_{t \in \mathcal{I}_i} y_t & i \in \{1, 2, ..., |C|\} \\ 0 & \text{else} \end{cases}$$

$$y'_t = \begin{cases} y_t + \mathcal{S}_i & \exists \mathcal{I}_i : t \in \mathcal{I}_i \\ y_t + \left(\frac{y_t - y_a}{y_b - y_a}\right)\mathcal{S}_{i+1} + \left(\frac{y_b - y_t}{y_b - y_a}\right)\mathcal{S}_i & \exists i : a \le t \le b \end{cases}$$

$$\text{where} \quad a = \max(\mathcal{I}_i), \quad b = \min(\mathcal{I}_{i+1})$$

Fig. 4: Equations for our proposed spatial warping method. 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 let C be the ordered sequence of non-overlapping time intervals corresponding to the extracted constant intervals. We define $\mathcal{E}$ as the sequence of embedding values in $\mathbb{R}^d$ corresponding to the time interval sequence C. The sequence $\mathcal{I}$ is used instead of C to handle edge cases. For notational simplicity, we also introduce a new sequence S whose i^{th} element is the difference between interval i's average value and the corresponding embedding value.

Task	Signal Type	Pearson	Spearman	Kendall's Tau	NMI
A	EvalDep Average	0.906	0.946	0.830	0.484
	Warped EvalDep	0.967	0.939	0.835	0.562
B	EvalDep Average	0.969	0.969	0.855	0.774
	Warped EvalDep	0.988	0.987	0.906	0.862

Table 1: Agreement measures for baseline and proposed warped fused annotation approaches. All warped results use a complete set of ordinal comparisons from the oracle. NMI = normalized mutual information.

4. RESULTS

Table 1 shows various objective truth agreement measures for the proposed method and the *EvalDep* method from [2] used as a baseline. The proposed method improves the accuracy of the ground truth estimate in both tasks and Fig 3 clearly shows that it produces a more self-consistent signal over large periods of time.

Although one would expect the proposed rank-based signal warping procedure to improve rank-based correlation metrics, the Spearman correlation decreases slightly primarily due to frame-level rank disagreements over the warped constant intervals rather than disagreements at a large scale due to the ordinal embedding. The same decrease in rank-based correlation can occur when any non-injective function is piecewise linearly warped and thus is not a particular artifact of this method.

5. FUTURE WORK

There are several compelling research directions for expanding on this work which we aim to address in the future. The total variation denoising and constant interval extraction procedures require selection of tunable constants to achieve desirable results, which we would like to eliminate. Further analysis of this method's ability to produce accurate ground truth estimates for more complex continuous annotation tasks, like 2-D dimensional core affect, is another exciting avenue. Large reductions in the number of required triplet comparisons may also be possible by using adaptive sampling techniques, by automatically inferring comparisons via stochastic transitivity, and by exploiting information redundancy in the triplets. Further investigation of the interplay between annotator uncertainty and trend annotations may reveal additional ways to improve the ground truth estimate.

6. CONCLUSION

In this paper we propose a novel method for improving the accuracy and consistency of fused continuous annotations for use as ground truth. We test our approach in experiments where objective truths are known and show that our approach yields a ground truth in better agreement with the objective truth in spite of the presence of several annotation artifacts.

7. REFERENCES

[1] Angeliki Metallinou and Shrikanth 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.

[2] Soroosh Mariooryad and Carlos 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.

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

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

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

[6] Galen Andrew, Raman Arora, Jeff A Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1247–1255.

[7] Mihalis A Nicolaou, Vladimir Pavlovic, and Maja Pantic, "Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1299–1311, 2014.

[8] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015.

[9] Tingting Zhu, Nic Dunkley, Joachim Behar, David A Clifton, and Gari D Clifford, "Fusing continuous-valued medical labels using a bayesian model," *Annals of biomedical engineering*, vol. 43, no. 12, pp. 2892–2902, 2015.

[10] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, "Modeling multiple time series annotations based on ground truth inference and distortion," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2016.

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

[12] Georgios N Yannakakis and Héctor P Martínez, "Ratings are overrated!," *Frontiers in ICT*, vol. 2, pp. 13, 2015.

[13] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[14] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc 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, 2012.

[15] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[16] Terry K Koo and Mae Y Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[17] Domenic V Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology.," *Psychological assessment*, vol. 6, no. 4, pp. 284, 1994.

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

[19] George Trigeorgis, Mihalis A Nicolaou, Stefanos Zafeiriou, and Bjorn W Schuller, "Deep canonical time warping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5110–5118.

[20] Leonid I Rudin, Stanley Osher, and Emad Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[21] Stephen R Becker, Emmanuel J Candès, and Michael C Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Mathematical programming computation*, vol. 3, no. 3, pp. 165, 2011.

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

[23] Omer Tamuz, Ce Liu, Ohad Shamir, Adam Kalai, and Serge J. Belongie, "Adaptively learning the crowd kernel," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 673–680, ACM.