Joint Multi-Dimensional Model for Global and Time-Series Annotations

Anil Ramakrishna, *Student Member, IEEE,* Rahul Gupta, *Member, IEEE,* and Shrikanth Narayanan, *Fellow, IEEE*

Abstract—Crowdsourcing is a popular approach to collect annotations for unlabeled data instances. It involves collecting a large number of annotations from several, often naive untrained annotators for each data instance which are then combined to estimate the ground truth. Further, annotations for constructs such as affect are often multi-dimensional with annotators rating multiple dimensions, such as valence and arousal, for each instance. Most annotation fusion schemes however ignore this aspect and model each dimension separately. In this work we address this by proposing a generative model for multi-dimensional annotation fusion, which models the dimensions jointly leading to more accurate ground truth estimates. The model we propose is applicable to both global and time series annotation fusion problems and treats the ground truth as a latent variable distorted by the annotators. The model parameters are estimated using the Expectation-Maximization algorithm and we evaluate its performance using synthetic data and real emotion corpora as well as on an artificial task with human annotations.

Index Terms—Annotation fusion, Emotion annotations, Multi-dimensional annotations, Time series annotation modeling, Expectation Maximization, Factor Analysis.

1 Introduction

Crowdsourcing is a popular tool used in collecting human judgments on subjective constructs such as emotion. Typical examples include annotations of images and video clips with categorical emotions or with continuous affective dimensions such as *valence* or *arousal*. Online platforms such as Amazon Mechanical Turk[1] (MTurk) and Crowdflower[2] have risen in popularity owing to their inexpensive annotation costs and their ability to scale efficiently.

Crowdsourcing is also a popular approach in collecting labels for training supervised machine learning algorithms. Such labels are typically obtained from domain experts, which can be slow and expensive. For example, in the medical domain, it is often expensive to collect diagnosis information given laboratory tests since this requires judgments from trained professionals. On the other hand, unlabeled patient data may be easily available. Crowdsourcing has been particularly successful in such settings with easy availability of unlabeled data instances since we can collect a large number of annotations from untrained and inexpensive workers over the Internet, which when combined together may be comparable or even better than expert annotations [1].

A typical crowdsourcing setting involves collecting annotations from a large number of workers; hence there is a need to robustly combine them to estimate the ground truth. The most common approach for this is to take simple averages for continuous annotations or perform majority voting for categorical annotations. However, this assumes uniform competency across all the workers which is not always guaranteed or justified. Several alternative approaches have been proposed to address this challenge, each assuming a specific function modeling the annotators' behavior. In practice, it is common to collect annotations on multiple questions for each data instance in order to reduce costs, the annotators' mental load or even to improve annotation accuracy. For example, if we're annotating valence and arousal for a given data instance (such as a single image or video segment), collecting annotations on both these dimensions in one session per instance may be preferred over collecting valence annotations for all instances followed by arousal.

Such a joint annotation task may entail *task specific* or *annotator specific* dependencies between the annotated dimensions. In the aforementioned example, task specific dependencies may occur due to inherent correlations between the valence and arousal dimensions depending on the experimental setup. Annotator specific dependencies may occur due to a given annotator's (possibly incorrect or incomplete) understanding of the annotation dimensions. Hence it is of relevance to model the dimensions jointly. However, most state of the art models in annotation fusion combine the annotations by treating the different dimensions independently.

Joint modeling of the annotation dimensions may result in more accurate estimates of the ground truth as well as in giving a better picture of the annotators' behavior. In this work, we address this goal by proposing a multi-dimensional model which makes use of any potential relationships between the annotation dimensions while combining them. The model we propose is applicable to both the global annotation setting (such as while collecting emotion annotations on a picture, judgment about the overall tone of a conversation, etc.) as well as time series annotations (for example, time continuous annotations of audio/video clips on dimensions such as engagement or affect). Our model treats the hidden ground truth as latent variables and estimates them jointly with the annotator parameters

1. www.mturk.com
2. www.crowdflower.com

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3006418, IEEE Transactions on Affective Computing

using the Expectation Maximization (EM) algorithm [2]. We evaluate the model in both settings with both synthetic and real emotion corpora. We also create an artificial annotation task with controlled ground truth which is used in the model evaluation for both settings.

The main contributions of this work are as follows:

1) We propose a unified model to capture relationships between annotation dimensions. For ease of exposition we focus on the linear case in this paper.
2) The linear model we propose results in an annotator specific matrix which captures this annotator level relationship between the annotation dimensions.
3) We create a novel multi-dimensional annotation task with controlled ground truth and use it to evaluate both the global and time series annotation settings of the model.

The rest of the paper is organized as follows. In Section 2, we review related work and motivate the problem in Section 3. In Section 4, we describe the proposed model and provide equations for parameter estimation using EM algorithm (derivations are deferred to the appendix). We evaluate the model in Section 5 and provide conclusions in Section 6.

2 Related work

Several authors, most notably [1], assert the benefits of aggregating opinions from many people which is often believed to be better than those from a small number of experts, under certain conditions. Often referred to as the *wisdom of crowds*, this approach has been remarkably popular in recent times, specially in fields such as psychology and behavioral sciences where a ground truth may not be easily accessible or may not exist. This popularity can be largely attributed to online crowdsourcing platforms such as Mturk that connect researchers with low cost workers from around the globe. Along with cost, scalability is another major appeal with such tools leading to their frequent use in machine learning, leveraging large scale annotation of data instances such as images [3], audio/video clips [4] and text snippets [5].

Figure 1 shows a common setting in the crowdsourcing paradigm. For each data instance m, annotator k provides a noisy annotation $a_k^{m,d}$ which depends on the ground truth $a_*^{m,d}$ where d is the dimension being annotated. Since we collect several annotations for each m, we need to aggregate them to estimate the unknown ground truth. The most common technique used in this aggregation is to take the average value in case of numeric annotations or perform majority voting in the case of categorical annotations as shown in Equation 1.

$$a_*^{m,d} = \underset{j}{\operatorname{argmax}} \sum_k \mathbb{1}\{a_k^{m,d} == j\} \tag{1}$$

where, $\mathbb{1}\{\}$ is the indicator function.

While simple and easy to implement, this approach assumes consistent reliability among the different annotators which seems unreasonable, especially in online platforms such as Mturk. To address this, several approaches have been suggested that account for annotator reliability in estimating the ground truth.

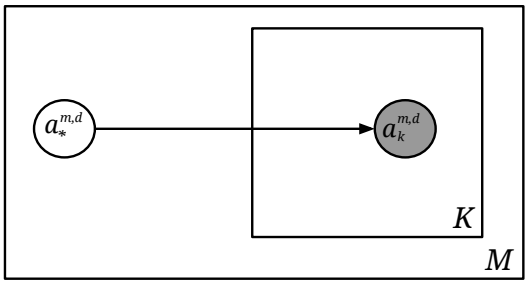


Fig. 1: Plate notation for a basic annotation model. $a_*^{m,d}$ is the latent ground truth for the given data instance (for the d^{th} question) and $a_k^{m,d}$ is the rating provided by the k^{th} annotator.

Early efforts to capture reliability in annotation modeling [6], [7] assumed specific structure to the functions modeled by each annotator. Given a set of annotations $a_k^{m,d}$ along with the corresponding function parameters, the ground truth is estimated using the Maximum A Posteriori (MAP) estimator.

$$a_*^{m,d} = \underset{j}{\operatorname{argmax}} \sum_k \log p(a_k^{m,d}|a_*^{m,d} = j) + \log p(a_*^{m,d} = j) \tag{2}$$

where $p(a_*^{m,d})$ is the prior probability of ground truth.

In [6], the categorical ground truth label $a_*^{m,d} = i$ is modified probabilistically by annotator k using a stochastic matrix Π_k as shown in Equation 3 in which each row is a multinomial conditional distribution given the ground truth.

$$P(a_k^{m,d} = j|a_*^{m,d} = i) = \pi_{ij}^k \tag{3}$$

Given annotations from K different annotators, their parameters Π_k and prior distribution of labels $p_j = P(a_*^{m,d} = j)$, the ground truth is estimated using MAP estimation as before.

$$a_*^{m,d} = \underset{j}{\operatorname{argmax}} \sum_k \log \pi_{j(a_k^{m,d})} + \log p_j \tag{4}$$

The above expression makes a conditional independence assumption for annotations given the ground truth label. Since we do not typically have the annotator parameters Π^k, these are estimated using the EM algorithm.

Figure 2 shows an extension of the model in Figure 1 in which we learn a predictor (classifier/regression model) for the ground truth jointly with annotator parameters. Such a predictor may be used to estimate the ground truth for new unlabeled data instances. This strategy of jointly modeling the annotator functions as well as the ground truth predictor has been shown to have better performance when compared to predictors trained independently using the estimated ground truth [8]. The ground truth estimate in this model is given by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3006418, IEEE Transactions on Affective Computing

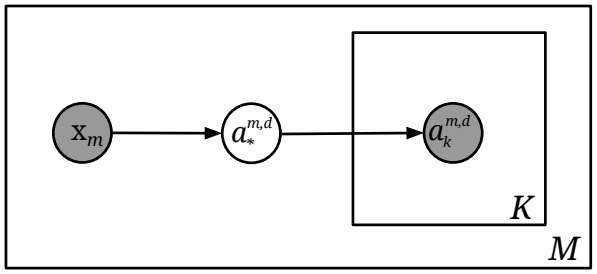


Fig. 2: Annotation model proposed by [8] with a jointly learned predictor. $\mathbf{x}_m$ is the set of features for the m^{th} data instance; $a_*^{m,d}$ is the d^{th} dimension of the latent ground truth which is modeled as a function of $\mathbf{x}_m$; $a_k^{m,d}$ is the rating provided by the k^{th} annotator.

$$a_*^{m,d} = \underset{a_*^{m,d}}{\operatorname{argmax}} \sum_k \log p(a_k^{m,d}|a_*^{m,d}) + \log p(a_*^{m,d}|\mathbf{x}_m) \quad (5)$$

Recently, several additional extensions have been proposed to the model in Figure 2; For example, in [9], the authors assume varying regions of annotator expertise in the data feature space and account for this using different probabilities for label confusion for each region. The authors show that this leads to a better estimation of annotator reliability and ground truth.

The models described so far have been designed for annotation tasks in which the task is to rate some global property of the data. For example, in image based emotion annotation, the task may be to provide annotations on affective dimensions such as valence and arousal conveyed by each image. However, human interactions often involve variations of these dimensions over time [10] which are captured using time series annotations from audio/video clips. Various tools have been developed to collect such annotations, including Anvil [11], Feeltrace [12], EMuJoy [13], Gtrace [14] and DARMA [15] (for a review of available tools and their properties, see [16] and [15]). In fusing such time series annotations, the previously mentioned models are applicable only if annotations from each frame are treated independently. However, this entails several unrealistic assumptions such as independence between frames, zero lag in the annotators and synchronized response in the annotators to the underlying stimulus.

Several works have been proposed to capture the underlying reaction lag in the annotators. [17] proposed a generalization of Probabilistic Canonical Correlation Analysis (PCCA) [18] named Dynamic PCCA which captures temporal dependencies of the shared ground truth space in a generative setting, and incorporated a latent time warping process to implicitly handle the reaction lags in annotators. They have further proposed a supervised extension of their model which jointly learns a predictor function for the latent ground truth signal similar to [8]. [19] address the reaction lag by explicitly finding the time shift that maximizes the mutual information between expressive behaviors and their annotations. [20] generalize the work of [19] by using a linear time invariant (LTI) filter which can also handle any bias or scaling the annotators may introduce.

More recent works in annotation fusion include [21] in which the authors propose a variant of the model in Figure 1 with various annotator functions to capture four specific types of annotator behavior. [22] describes a mechanism named approval voting that allows annotators to provide multiple answers instead of one for instances where they are not confident. [23] uses repeated sampling for opinions from annotators over the same data instances to increase reliability in annotations.

Most of the models described above focus on combining annotations on each dimension separately. However, the annotation dimensions are often related. For example, many studies in emotion literature have reported interrelationships between discrete emotion categories [24], [25]. The circumplex model [26], which attempts to capture these relationships by modeling the emotions as points on a two dimensional space, has also been noted to exhibit *v-shaped* patterns in the joint distribution of valence and arousal [27]. In addition, in most practical applications, the annotation tasks themselves are multi-dimensional. For example, while collecting ratings on affective dimensions it is routine to collect annotations on valence, arousal and dominance together. Further, there may be dependencies between the internal definitions the annotators hold for the annotation dimensions; for example, while annotating emotional dimensions, a particular annotator may associate certain valence values with only a certain range of arousal. Hence it may be beneficial to model the different dimensions jointly while performing annotation fusion. However, research in this direction has been limited. [28] proposed a model which assumes joint Gaussian noise between the annotation dimensions, but their model fails to capture structural dependencies described above between the annotation and ground truth dimensions. The model proposed in [17] can indeed be generalized to combine the different annotation dimensions together but they do not evaluate with joint annotated dimensions from a real dataset as that is not the focus of their work. [29] jointly model continuous annotations on valence and arousal using *personalized basis spline* functions, on which functional PCA is applied to identify the dominant spline functions. Using this model, they estimate the ground truth for each data instance using a heuristic algorithm, but their model does not include a jointly trained ground truth predictor. It is therefore of relevance to model multi-dimensional annotation fusion as part of the unified annotator function and predictor modeling paradigm.

In this work, we propose a joint multi-dimensional model to address many of the gaps mentioned above. Our model captures annotator specific linear relationships between different annotation dimensions, and is an extension of the Factor Analysis model [30]. It incorporates an annotator specific transformation matrix parameter F_k, which explicitly captures the relationship between the annotation dimensions and enables clear interpretations of the estimated relationships; the matrix F_k is jointly estimated with a predictor for the ground truth signal. We further provide generalizations of our model to both global and time series annotation settings. We begin with a motivation followed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3006418, IEEE Transactions on Affective Computing

by a detailed description of the model and its parameter estimation in the next sections.

3 Motivation

To examine the relationships between the annotation dimensions, we created a plot of absolute values of Pearson's correlation between annotation dimensions from four commonly studied emotional corpora in Figure 3: IEMOCAP [31], SEMAINE [32], RECOLA [33] and the movie emotion corpus from [27]. Each of these corpora include annotations over affective dimensions such as valence, arousal, dominance and power. For the IEMOCAP corpus, we used global annotations while the others include time series annotations of the affective dimensions from videos. In each case, the correlations were computed from concatenated annotation values between all the dimensions.

As is evident, in almost all cases, the annotation dimensions exhibit non-zero correlations. We attribute the inconsistent correlations between the dimensions across corpora to varying underlying affective narratives as well as differences in perceptions and biases introduced by individual annotators themselves (see Section A.1). The non-zero correlations highlight the benefit of modeling the annotation dimensions jointly. The model we propose is aimed at addressing this. We explain the model in detail in the next section.

4 Joint Multi-dimensional Annotation Model

4.1 Setup

The proposed model is shown in Figure 4. Each data instance m has a feature vector $\mathbf{x}_m$ and an associated multidimensional ground truth $\mathbf{a}_*^m$, which is defined as follows,

$$\mathbf{a}_*^m = f(\mathbf{x}_m; \Theta) + \boldsymbol{\epsilon}_m \tag{6}$$

We assume that from a pool of K annotators, a subset operates on each data instance and provides their annotation $\mathbf{a}_k^m$.

$$\mathbf{a}_k^m = g(\mathbf{a}_*^m; F_k) + \boldsymbol{\eta}_k \tag{7}$$

where index k corresponds to the k^{th} annotator; F_k is an annotator specific matrix that defines his/her linear weights for each output dimension; $\boldsymbol{\epsilon_m}$ and $\boldsymbol{\eta_k}$ are noise terms defined individually in the next sections along with the functions f and g. In the global annotation setting, both $\mathbf{a}_*^m$ and $\mathbf{a}_k^m \in \mathbb{R}^{\mathrm{D}}$ where D is the number of items being annotated; for the time series setting $\mathbf{a}_*^m$ and $\mathbf{a}_k^m \in \mathbb{R}^{\mathrm{T}\times\mathrm{D}}$, where T is the total duration of the data instance (audio/video signal). In all subsequent definitions, we use uppercase letters M, K, T, D to denote various counts and lowercase letters m, k, t, d to denote the corresponding index variables.

We make the following assumptions in our model.

- A1 Annotations are independent for different data instances.
- A2 The annotations for a given data instance are independent of each other given the ground truth.
- A3 The model ground truths for different annotation dimensions are assumed to be conditionally independent of each other given the features $\boldsymbol{x}_m$.

4.2 Global annotation model

In this setting, the ground truth and annotations are d dimensional vectors for each data instance. We define the ground truth $\mathbf{a}_*^m$ and annotations $\mathbf{a}_k^m$ as follows.

$$\mathbf{a}_*^m = \Theta^T \mathbf{x}_m + \boldsymbol{\epsilon}_m \tag{8}$$

$$\mathbf{a}_k^m = F_k \mathbf{a}_*^m + \boldsymbol{\eta}_k \tag{9}$$

where, $\mathbf{x}_m \in \mathbb{R}^{\mathrm{P}}$; $\Theta \in \mathbb{R}^{\mathrm{P}\times\mathrm{D}}$; $\boldsymbol{\epsilon}_m \sim N(\mathbf{0}, \sigma^2 I)$; $\sigma^2 \in \mathbb{R}$. The annotator noise $\boldsymbol{\eta}_k$ is defined as $\boldsymbol{\eta}_k \sim N(\mathbf{0}, \tau_k^2 I)$; $\tau_k^2 \in \mathbb{R}$. $F_k \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}$ is the annotator specific weight matrix. Each annotation dimension value $a_k^{m,d}$ for annotator k is defined as a weighted average of the ground truth vector $\mathbf{a}_*^m$ with weights given by the vector $F_k(d, :)$.

4.2.1 Parameter Estimation

The model parameters $\Phi = \{F_k, \Theta, \sigma^2, \tau_k^2\}$ are estimated using Maximum Likelihood Estimation (MLE) in which they are chosen to be the values that maximize the likelihood function $\mathcal{L}$.

$$\begin{aligned}\log \mathcal{L} &= \sum_{m=1}^{M} \log p(\mathbf{a}_1^m \dots \mathbf{a}_K^m; \Phi) \\ &= \sum_{m=1}^{M} \log \int_{\mathbf{a}_*^m} p(\mathbf{a}_1^m \dots \mathbf{a}_K^m | \mathbf{a}_*^m; F_k, \tau_k^2) p(\mathbf{a}_*^m; \Theta, \sigma^2)\, d\mathbf{a}_*^m\end{aligned} \tag{10}$$

Optimizing Equation 10 directly is intractable because of the presence of the integral within the log term, hence we use the EM algorithm. Note that the model we propose assumes that only some random subset of all available annotators provide annotations on a given data instance, as shown in Figure 4. However, for ease of exposition, we overload the variable K and use it here to indicate the number of annotators that attempt to judge the given data instance m.

4.2.2 EM algorithm

The Expectation Maximization (EM) algorithm to estimate the model parameters is shown below. It is an iterative algorithm in which the E and M-steps are executed repeatedly until an exit condition is encountered. Complete derivation of the model can be found in Appendix B.

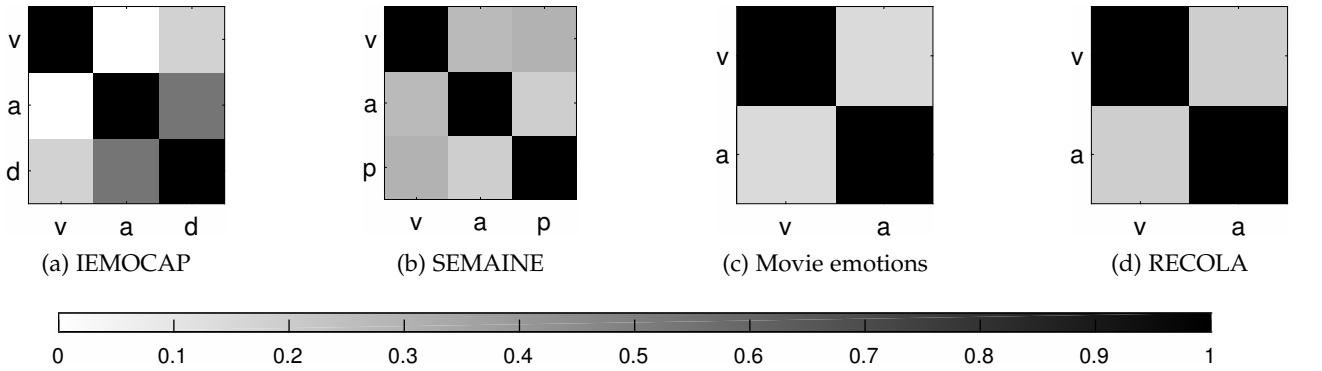
Initialization We initialize by assigning the expected values and covariance matrices for the m ground truth vectors $\mathbf{a}_*^m$ to their sample estimates (i.e. sample mean and sample covariance) from the corresponding annotations. We then estimate the parameters as described in the maximization step using these estimates.

E-step In this step we take expectation of the log likelihood function with respect to $p(\mathbf{a}_*^m | \mathbf{a}_1^m \dots \mathbf{a}_K^m)$ and the resulting objective is maximized with respect to the model parameters in the M-step. Equations to compute the expected value and covariance matrices for the latent variable $\mathbf{a}_*^m$ in the E-step are listed below.

$$\mathbb{E}_{\mathbf{a}_*^m | \mathbf{a}_1^m \dots \mathbf{a}_K^m}[\mathbf{a}_*^m] = \Theta^T \mathbf{x}_m + \Sigma_{\mathbf{a}_*^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} \Sigma^{-1}_{\mathbf{a}_1^m \dots \mathbf{a}_K^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} (\mathbf{a}^m - \boldsymbol{\mu}^m)$$

$$\Sigma_{\mathbf{a}_*^m | \mathbf{a}_1^m \dots \mathbf{a}_K^m}[\mathbf{a}_*^m] = \Sigma_{\mathbf{a}_*^m, \mathbf{a}_*^m} - \Sigma_{\mathbf{a}_*^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} \Sigma^{-1}_{\mathbf{a}_1^m \dots \mathbf{a}_K^m, \mathbf{a}_1^m \dots \mathbf{a}_K^m} \Sigma_{\mathbf{a}_1^m \dots \mathbf{a}_K^m, \mathbf{a}_*^m}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3006418, IEEE Transactions on Affective Computing



(a) IEMOCAP (b) SEMAINE (c) Movie emotions (d) RECOLA

Fig. 3: Correlation heatmaps for annotations from a representative sample of emotion annotated datasets; v - valence, a - arousal, d - dominance, p - power

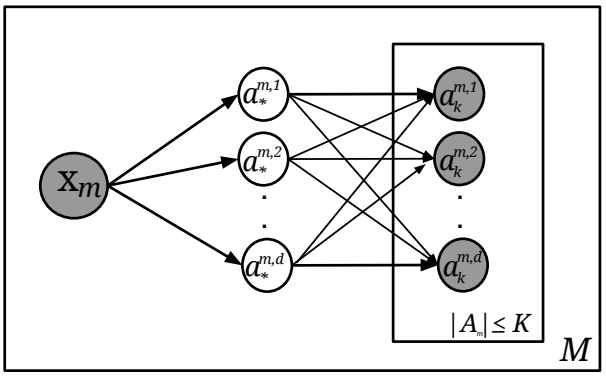


Fig. 4: Proposed model. $\mathbf{x}_m$ is the set of features for the m^{th} data instance, $a_*^{m,d}$ is the latent ground truth for the d^{th} dimension and $a_k^{m,d}$ is the rating provided by the k^{th} annotator. Vectors $\mathbf{x}_m$ and $\mathbf{a}_k^m$ (shaded) are observed variables, while $\mathbf{a}_*^m$ is latent. $\mathbf{A}_m$ is the set of annotator ratings for the m^{th} instance.

The Σ terms are covariance matrices between the subscripted random variables. $\mathbf{a}^m$ and $\boldsymbol{\mu}^m$ are DK dimensional vectors obtained by concatenating the K annotation vectors $\mathbf{a}_1^m, \ldots \mathbf{a}_K^m$ and their corresponding expected values.

M-step In this step, we compute current estimates for the parameters as follows. The expectations shown below are over the conditional distribution $\mathbf{a}_*^m | \mathbf{a}_1^m \ldots \mathbf{a}_K^m$.

$$\Theta = (\mathrm{X}^T\mathrm{X})^{-1}(\mathrm{X}^T\,\mathbb{E}[\mathrm{a}_*^m])$$

$$F_k = \left(\sum_{m=1}^{M_k} \mathbf{a}_K^m\,\mathbb{E}[(\mathbf{a}_*^m)^T]\right)\left(\sum_{m=1}^{M_k} \mathbb{E}[\mathbf{a}_*^m\,(\mathbf{a}_*^m)^T]\right)^{-1}$$

$$\sigma^2 = \frac{1}{md}\sum_{m=1}^{M}\Big(\mathbb{E}[(\mathbf{a}_*^m)^T\mathbf{a}_*^m] - 2tr\left(\Theta'^T\mathbf{x}_m\,\mathbb{E}[(\mathbf{a}_*^m)^T]\right) + tr(\mathbf{x}_m^T\Theta'\Theta'^T\mathbf{x}_m)\Big)$$

$$\tau_k^2 = \frac{1}{m_k d}\sum_{m=1}^{M_k}\Big((\mathbf{a}_K^m)^T\mathbf{a}_K^m - 2tr\big(F_k'^T\mathbf{a}_K^m\,\mathbb{E}[(\mathbf{a}_*^m)^T]\big) + tr\big(F_k'^T F_k'\,\mathbb{E}[\mathbf{a}_*^m(\mathbf{a}_*^m)^T]\big)\Big)$$

Note the similarity of the update equation for Θ with the familiar normal equations. We are using the soft estimate of $\mathbf{a}_*^m$ to find the expression for Θ in each iteration. Here, X is the feature matrix for all data instances; it includes individual feature vectors x_m in its rows. Θ' and F_k' are parameters from the previous iteration.

Termination We run the algorithm until convergence, and stop model training when the change in log-likelihood falls below a threshold of 0.001%.

4.3 Time series annotation model

In this setting, the ground truth and the annotations are matrices with T rows (time) and D columns (annotation dimensions). The ground truth matrix $\mathbf{a}_*^m$ is defined as follows.

$$\mathrm{vec}(\mathbf{a}_*^m) = \mathrm{vec}(\mathrm{X}_m\Theta) + \boldsymbol{\epsilon}_m \tag{11}$$

where $\mathbf{a}_*^m \in \mathbb{R}^{\mathrm{T}\times\mathrm{D}}$, $\mathrm{X}_m \in \mathbb{R}^{\mathrm{T}\times\mathrm{P}}$ and $\Theta \in \mathbb{R}^{\mathrm{P}\times\mathrm{D}}$; T represents the time dimension and is the length of the time series. X_m is the feature matrix where each row corresponds to features extracted from the data instance for one particular time stamp. vec(.) is the vectorization operation which flattens the input matrix in column first order to a vector. $\boldsymbol{\epsilon}_m \sim \mathcal{N}(\mathbf{0}, \sigma^2 I) \in \mathbb{R}^{\mathrm{TD}}$ is the additive noise vector with $\sigma \in \mathbb{R}$.

In [20], the authors propose a linear model where the annotation function $g(\mathbf{a}_*^m; F_k)$ is a causal linear time invariant (LTI) filter of fixed width. The advantage of using an LTI filter is that it can capture scaling and time-delay biases introduced by the annotators.

The filter width W is chosen such that $W \ll T$, where T is the number of time stamps for which we have the annotations. The annotation function for dimension d' can be viewed as the left multiplication of a filter matrix $B_k^{d'} \in \mathbb{R}^{\mathrm{T}\times\mathrm{T}}$ as shown in Equation 12.

$$B_k^{d'} = \begin{bmatrix} b_1^{d'} & 0 & 0 & 0 & 0 & \dots & 0 \\ b_2^{d'} & b_1^{d'} & 0 & 0 & 0 & \dots & 0 \\ b_3^{d'} & b_2^{d'} & b_1^{d'} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & b_W^{d'} & \dots & b_1^{d'} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & 0 & b_W^{d'} & \dots & b_1^{d'} \end{bmatrix} \quad (12)$$

We extend this model in our work to combine information from all of the annotation dimensions. Specifically, the ground truth is left multiplied by D horizontally concatenated filter matrices, each $\in \mathbb{R}^{\mathrm{T}\times\mathrm{T}}$ corresponding to a different dimension as shown below.

$$\mathbf{a}_k^{m,d} = F_k^d \mathrm{vec}(\mathbf{a}_*^m) + \boldsymbol{\eta}_k \quad (13)$$

where,

$$F_k^d = [B_k^{d,1}, B_k^{d,2}, \dots, B_k^{d,D}] \quad (14)$$

$F_k^d \in \mathbb{R}^{\mathrm{T}\times\mathrm{TD}}$ with WD unique parameters. $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \tau_k^2 I) \in \mathbb{R}^{\mathrm{T}}$ with $\tau_k^2 \in \mathbb{R}$.

4.3.1 Parameter Estimation

Estimating the model parameters similar to the global model requires computing the expectations over a vector of size TD. Since T is the number of time stamps in the task and can be arbitrarily long, this may not be feasible in all tasks. For example, in the movie emotions corpus [27], annotations are computed at a rate of 25 frames per second with each file of duration ∼30 minutes or of ∼45k annotation frames. To avoid this we use a variant of EM named *Hard EM* in which instead of taking expectations over the entire conditional distribution of $\mathbf{a}_*^m$ we find its mode. This variant has been shown to be comparable in performance to the classic EM (*Soft EM*) despite being significantly faster and simple [34]. This approach is similar to the parameter estimation strategy devised by [20] in their time series annotation model.

The likelihood function is similar to the global model in Equation 10 as shown below.

$$\log \mathcal{L} = \sum_{m=1}^{M} \log \int_{\mathbf{a}_*^m} p(\mathbf{a}_1^m \dots \mathbf{a}_K^m | \mathbf{a}_*^m; F_k, \tau_k^2) p(\mathbf{a}_*^m; \Theta, \sigma^2)\, d\mathbf{a}_*^m$$

However the integral here is with respect to the flattened vector $\mathrm{vec}(\mathbf{a}_*^m)$.

4.3.2 EM algorithm

The EM algorithm for the time series annotation model is listed below. Complete derivations can be found in Appendix C.

Initialization Unlike the global annotation model, we initialize $\mathbf{a}_*^m$ randomly since we observed better performance when compared to initializing it with the annotation means. Given this $\mathbf{a}_*^m$, the model parameters are estimated as described in the maximization step below.

E-step In this step we assign $\mathbf{a}_*^m$ to the mode of the conditional distribution $q(\mathbf{a}_*^m) = p(\mathbf{a}_*^m | \mathbf{a}_1^m, \dots, \mathbf{a}_K^m)$. Since this distribution is normal (see appendix B) finding the mode is equivalent to minimizing the following expression.

$$\mathbf{a}_*^m = \underset{\mathbf{a}_*^m}{\mathrm{argmin}} \sum_k \sum_d ||\mathbf{a}_k^{m,d} - F_k^d \mathrm{vec}(\mathbf{a}_*^m)||_2^2 + ||\mathrm{vec}(\mathbf{a}_*^m) - \mathrm{vec}(\mathrm{X}_m\Theta)||_2^2$$

M-step Given the estimate for $\mathbf{a}_*^m$ from the E-step, we substitute it in the likelihood function and maximize with respect to the parameters in the M-step. The estimates for the different parameters are shown below.

$$\Theta = \Big(\sum_{m=1}^{M} \mathrm{X}_m^T \mathrm{X}_m \Big)^{-1} \Big(\sum_{m=1}^{M} \mathrm{X}_m^T \mathbf{a}_*^m \Big)$$

$$f_k^d = \Big(\sum_{m=1}^{M_k} A^T A \Big)^{-1} \Big(\sum_{m=1}^{M_k} A^T \mathbf{a}_k^{m,d} \Big)$$

$$\sigma^2 = \frac{1}{MTD} \sum_{m=1}^{M} ||\mathrm{vec}(\mathbf{a}_K^m) - \mathrm{vec}(\mathrm{X}_m\Theta)||_2^2$$

$$\tau_k^2 = \frac{1}{M_k TD} \sum_{m=1}^{M_k} \sum_d ||\mathbf{a}_k^{m,d} - F_k^d \mathrm{vec}(\mathbf{a}_*^m)||_2^2$$

M_k is the number of files annotated by user k; A is a matrix obtained by reshaping $\mathrm{vec}(\mathbf{a}_*^m)$ as described in subsection C.1.2.

Termination We run the algorithm until convergence, and stop model training when the change in log-likelihood falls below a threshold of 0.5%.

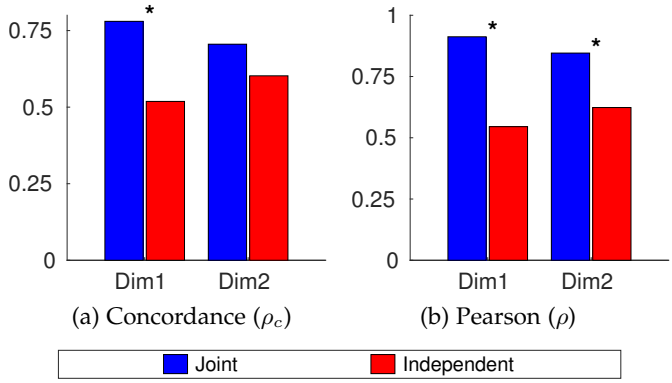
5 Experiments & Results

We evaluate the models described above on three different types of data: synthetic data, an artificial task with human annotations, and finally with real data. We describe these below. We compare our joint models with their *independent* counterparts as baselines, in which each annotation dimension is modeled separately. This allows us to highlight the benefits of moving to a multi-dimensional annotation fusion scheme with everything else kept constant. Update equations for the independent model can be obtained by running the models described above for each dimension separately with $D = 1$. Note that the independent model is similar in the global setting to the regression model proposed in [8] (with ground truth scaled by the singleton f_k^d). In the time series setting it is identical to the model proposed by [20].

The models are evaluated by comparing the estimated $\mathbf{a}_*^m$ with the actual ground truth. We report model performance using two metrics: the Concordance correlation coefficient (ρ_c) [35] and the Pearson's correlation coefficient (ρ). ρ_c measures any departures from the *concordance line* (line passing through the origin at 45° angle). Hence it is sensitive to rotations or rescaling in the predicted ground truth. Given two samples x and y, the sample concordance coefficient $\hat{\rho}_c$ is defined as shown below.

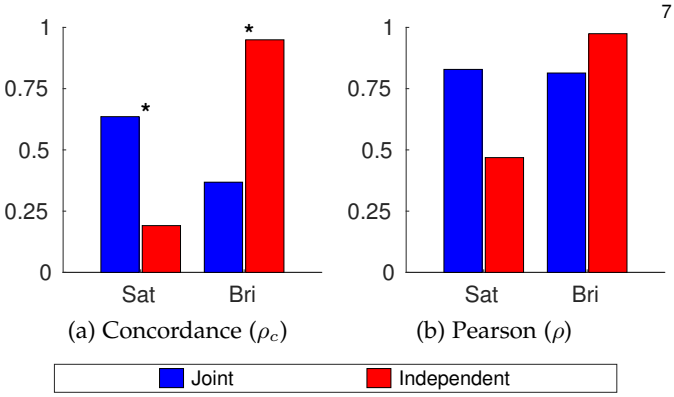
$$\hat{\rho}_c = \frac{2 s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3006418, IEEE Transactions on Affective Computing



(a) Concordance (ρ_c) (b) Pearson (ρ)

Fig. 5: Performance of global annotation model on synthetic dataset; **-statistically significant*



(a) Concordance (ρ_c) (b) Pearson (ρ)

Fig. 6: Performance of global annotation model on artificial dataset; *Sat-Saturation, Bri-Brightness; *-statistically significant*

We also report results in Pearson's correlation to highlight the accuracy of the models in the presence of rotations.

As noted before, the models proposed in this paper are closely related to the Factor Analysis model, which is vulnerable to issues of unidentifiability [36], due to the matrix factorization. Different types of unidentifiability have been studied in literature, such as factor rotation, scaling and label switching. In our experiments, we handle label switching through manual judgment (by reassigning the estimated ground truth between dimensions if necessary) as is common in psychology [37], but defer the task of choosing an appropriate prior on the rotation matrix F_k to address other unidentifiabilities for future work.

We report aggregate test set results using C-fold cross validation. To address overfitting, within each fold, we evaluate the parameters obtained after each iteration of the EM algorithm by estimating the ground truth on a disjoint validation set, and pick those with the highest performance in concordance correlation ρ_c as the parameter estimates of the model. We then estimate the performance of this parameter set in predicting the ground truth from a separate held out test set for that fold. Finally, we also report statistically significant differences between the joint and independent models at 5% false-positive rate ($\alpha = 0.05$) in all our experiments.

5.1 Global annotation model

The global annotation model uses the EM algorithm described in Section 4.2.2 to estimate the ground truth for discrete annotations. We evaluate the model in three different settings described below. Statistical significance tests were run by computing bootstrap confidence intervals [38] on the differences in model performances across the C-folds. To establish the statistical significance, we ran the joint and independent models to obtain C test set model predictions from C folds. Given these, we ran 1000 bootstrap iterations in which the test set predictions were sampled with replacement, from which ρ and ρ_c were estimated for each dimension. We conclude significance if the evaluation metric being examined was higher in at least 95% of the bootstrap runs.

5.1.1 Synthetic data

We created synthetic data according to the model described in Section 4.2 with random features $\mathrm{X} \in \mathbb{R}^{500}$ for 100 data instances each with 2 dimensions of annotations (i.e. D=2). 10 artificial annotators, each with unique random F_k matrices were used to produce annotations for all the data instances. Elements of the feature matrices were sampled from the standard normal distribution, while the elements of F_k matrices were sampled from $\mathcal{U}(0, 1)$. Elements of ground truth $\mathbf{a}_*^m$ were sampled from $\mathcal{U}(-1, 1)$ and θ was estimated from $\mathbf{a}_*^m$ and X. Since its off diagonal elements are non-zero, our choice of F_k represents tasks in which the annotation dimensions are related to each other.

Figure 5 shows the performance of joint and independent models in predicting the ground truth $\mathbf{a}_*^m$. For both dimensions, the proposed joint model predicts the $\mathbf{a}_*^m$ with considerably higher accuracy as shown by the higher correlations, highlighting the advantages of modeling the annotation dimensions jointly when they are expected to be related to each other.

5.1.2 Artificial data

Since crowdsourcing experiments typically involve collecting subjective annotations, they seldom have well defined ground truth. As a result, most annotation models are evaluated on expert annotations collected by specially trained users. For example, while collecting annotations on medical data, labels estimated by fusing annotations from naive users may be evaluated against those provided by experts such as doctors. However, this poses a circular problem since the expert annotations themselves may be subjective and combining them to estimate the ground truth is not straightforward. To address this, we created an artificial task with controlled ground truth on which we collect annotations from multiple annotators and evaluate the fused annotation values with the known ground truth values, similar to [39]. In our task, the annotators were asked to provide their best estimates on perceived saturation and brightness values for monochromatic images. The relationship between perceived saturation and brightness is well known as the Helmholtz—Kohlrausch effect [40], according to which, increasing the saturation of an image leads to an increase in the perceived brightness, even if the actual brightness was constant.

In our experiments, we collected annotations on images from two regimes: one with fixed saturation and varying brightness, and vice versa. This approach was chosen since

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3006418, IEEE Transactions on Affective Computing

it would allow us to evaluate the impact of change in either brightness or saturation while the other was held constant. The color of the images were chosen randomly (and independent of the image's saturation and brightness) between green and blue. Annotations were collected on Mturk and the annotators were asked to familiarize themselves with saturation and brightness using an online interactive tool before providing their ratings. In both experiments, a reference image with fixed brightness and saturation was inserted after every ten annotation images to prevent any bias in the annotators. The reference images were hidden from the annotators and appeared as regular annotation images. For parameter estimation, RGB values were chosen as the features for each image.

We used the joint model to estimate the ground truth for the two regimes separately since we expect the relationship between saturation and brightness to be dissimilar in the two cases. From each experiment, predicted values of the underlying dimension being varied was compared with the actual $\mathbf{a}_*^m$ values. For example, in the experiment with varying saturation and fixed brightness, the joint model was run on full annotations, but only the estimated values of saturation were compared with ground truth saturation. For the independent model, we use annotation values of the underlying dimension being varied from each regime, and compare the estimated values with ground truth.

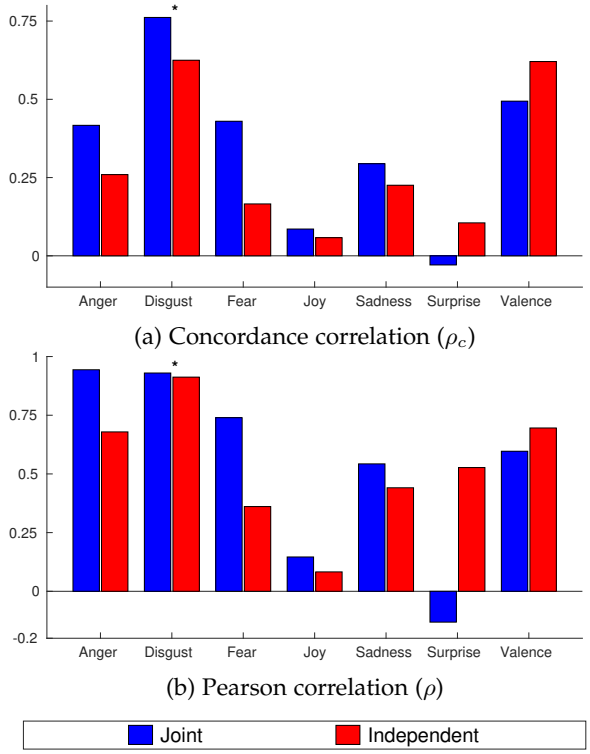
Figure 6 shows the performance of the joint and independent models for this experiment. The joint model leads to better estimates of saturation when compared to the independent model by making use of the annotations on brightness. This agrees with the Helmholtz—Kohlrausch phenomenon described above, since the annotators can perceive the changing saturation as a change in brightness, leading to correlated annotations for the two dimensions. On the other hand, the independent model leads to better estimates of brightness, which seems to have no effect on perceived saturation annotations. This experiment highlights the benefits of jointly modeling annotations in cases where the annotation dimensions may be correlated or dependent on one another.

5.1.3 Real data

Our final experiment for the global model was on the task of annotating news headlines in which the annotators provide numeric ratings for various emotions. This dataset was first described in the 2007 SemEval task on affective text [41]. Numeric ratings from the original task were labeled by trained annotators and we treat these as expert annotations. We use Mturk annotations from [5] as the actual input to our model. Sentence level annotations are provided on seven emotions (D=7): *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *valence* (positive/negative polarity). We use sentence level embeddings computed using the pre-trained sentence embedding model sent2vec[3] [42] as feature vectors x for the model.

Figure 7 shows the performance of the joint and independent models on this task. The joint model shows better performance in predicting the reference emotion labels for

3. https://github.com/epfml/sent2vec



(a) Concordance correlation (ρ_c)

(b) Pearson correlation (ρ)

Fig. 7: Performance of global annotation model on the text emotions dataset; *-*statistically significant*

anger, *disgust*, *fear*, *joy* and *sadness*, but performs worse than the independent model in predicting *surprise* and *valence*.

5.2 Time series annotation model

In this setting, the annotations are collected on data with a temporal dimension, such as time series data, video or audio signals. Similar to the global model, we evaluate this model in 3 settings: synthetic, artificial and on real data. The evaluation metrics ρ_c and ρ are computed over estimated and actual ground truth vectors $\mathbf{a}_*^m$ by concatenating the data instances into a single vector. The time series models have the window size W as an additional hyperparameter, which is selected using a validation set. In each fold of the dataset, we train model parameters for different window sizes from the set $\{5, 10, 20, 50\}$, and pick W and related parameters with the highest concordance correlation ρ_c on the validation set. These are then evaluated on a disjoint test set, and we repeat the process for each fold. In each experiment, the parameters were initialized randomly, and the process was repeated 20 times at different random initializations, selecting the best starting point using the validation set. To identify significant differences, we compute the test set performance of the two models for each fold, and run the paired t-test between the C sized samples of ρ and ρ_c corresponding to the joint and independent models. We do not bootstrap confidence intervals due to smaller test set sizes.

5.2.1 Synthetic data

The synthetic dataset was created using the model described in Section 4.3. Elements of the feature matrix were sampled from the standard normal distribution while elements of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3006418, IEEE Transactions on Affective Computing

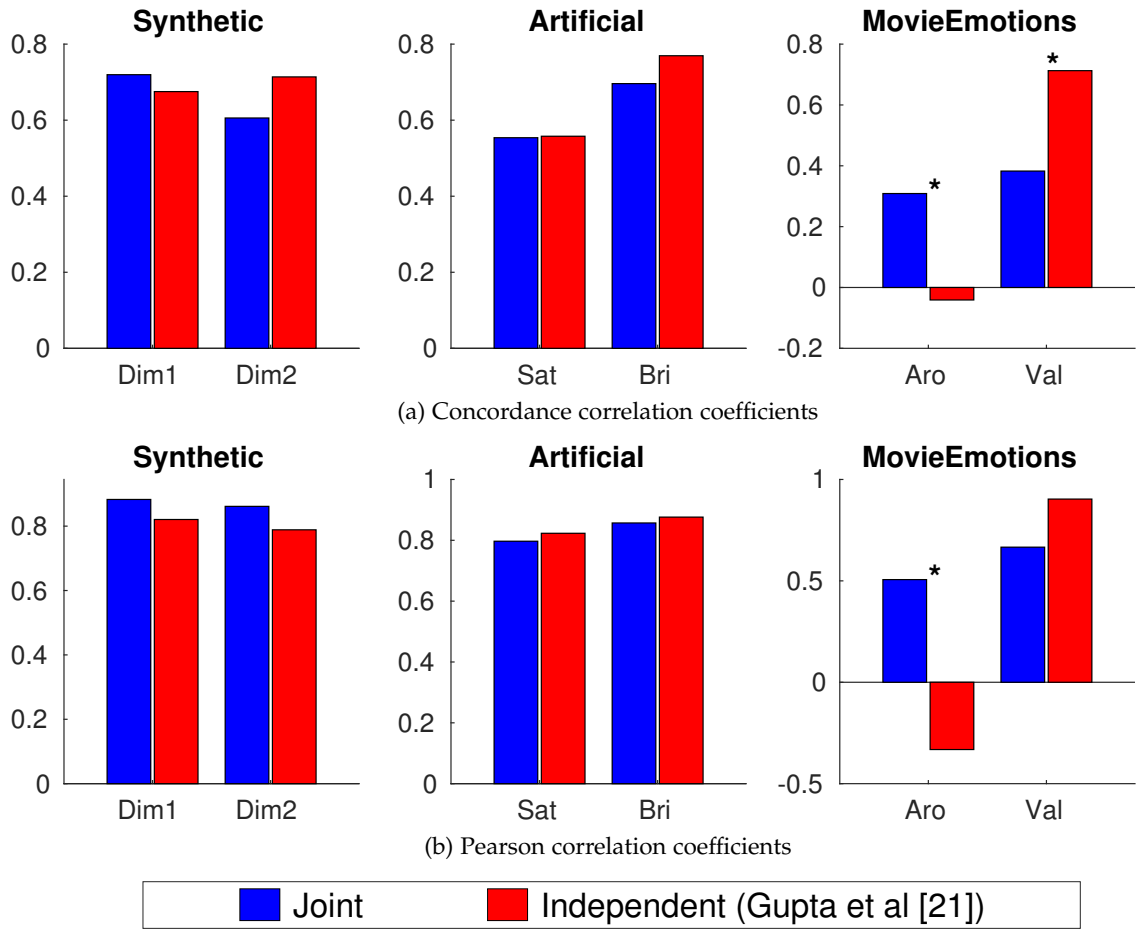


Fig. 8: Concordance and Pearson correlation coefficients between ground truth/reference and model predictions for the time series annotation model; *-*statistically significant*

F_k and ground truth were sampled from $\mathcal{U}(0,1)$. In this setting each data instance includes T feature vectors, one for each time stamp. The time dependent feature matrices were created using a random walk model without drift but with *lag* to mimic a real world task. In other words, while creating the P dimensional time series, the features vectors were held fixed for a time period arbitrarily chosen to be between 2 to 4 time stamps. This was done because in most tasks the underlying dimension (such as emotion) is expected to remain constant at least for a few seconds. In addition, the transition between changes in the feature vectors were linear and not abrupt. In our experiments, we chose $P = 500$, $T = 350$, $D = 2$, $M = 18$ and the number of annotators $K = 6$.

Figure 8 shows the aggregate results across C-folds ($C = 5$) for the joint and independent models in the 3 settings. In the synthetic dataset, the joint model achieves higher values for Pearson's correlation ρ for both the dimensions and higher value for ρ_c for dimension 1. For dimension 2 however, the independent model achieves better ρ_c.

5.2.2 Artificial data

We collected annotations on videos with the artificial task of identifying saturation and brightness, described in the previous section. The videos consisted of monochromatic images with the underlying saturation and brightness varied independent of each other. The dimensions were created using a random walk model with lag as described in Section 5.2.1. The annotations were collected in house using an annotation system developed using the Robot Operating System [43]. 10 graduate students gave their ratings on the two dimensions. Each dimension was annotated independently using a mouse controlled slider. For parameter estimation, the feature vectors for each time stamp were RGB values.

As seen in Figure 8, both models achieve similar performance in predicting the ground truth for saturation and brightness in terms of ρ, as well as in predicting saturation in terms of ρ_c. The independent model achieves slightly better performance in predicting brightness in terms of concordance correlation (though not statistically significant); however, their performance in terms of ρ suggests that the joint model output differs only in terms of a linear scaling. The joint model appears to be at par with the independent model for the most part, suggesting that the transformation matrix F_k connecting the two dimensions for each annotator, is unable to accurately capture the dependencies between the dimensions, likely due to the fact that, unlike the global annotation model, the underlying brightness and saturation were varied simultaneously and independent of each other

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3006418, IEEE Transactions on Affective Computing

(leading to non-linear dependencies between them), and that we limit F_k to only capture linear relationships.

5.2.3 Real data

We finally evaluate our model on a real world task with time series annotations. We chose the task of predicting the affective dimensions of valence and arousal from movie clips, first described in [27]. The associated corpus includes time series annotations of valence and arousal on contiguous 30 minute video segments from 12 Academy Award winning movies. This task was chosen because the data set includes both expert annotations as well as annotations from naive users. We treat the expert annotations as reference and evaluate the estimated dimensions against them; however, we note that the expert labels were provided by just one annotator, which may itself be noisy.

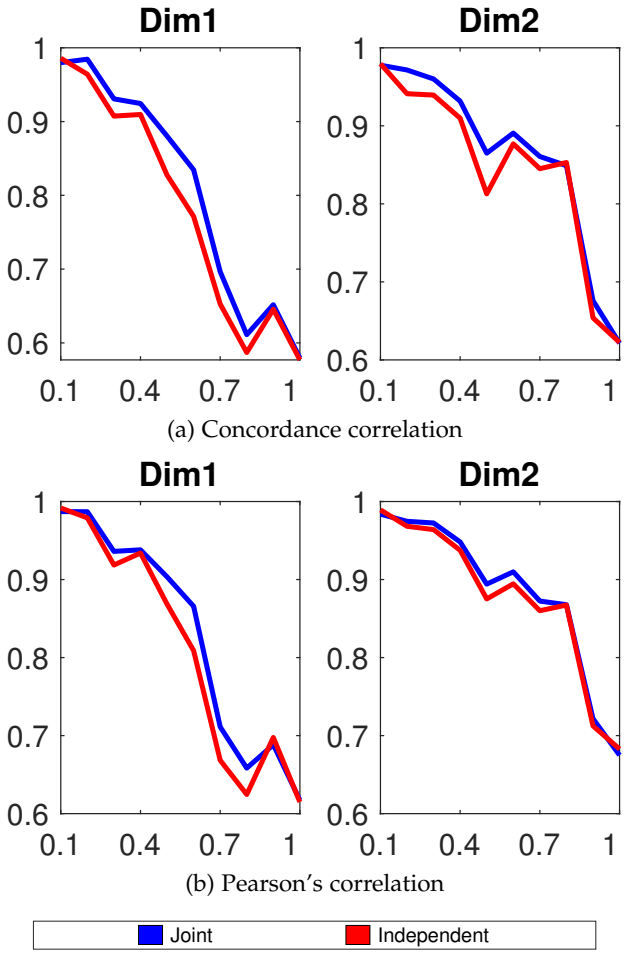
For each movie clip, 6 annotators provide annotations on their *perceived* valence and arousal using the Feeltrace [12] annotation tool. The features used in our parameter estimation include combined audio and video features extracted separately. The audio features were estimated using the emotion recognition baseline features from Opensmile [44] at 25 fps (same frame rate as the video clips) and aggregated at a window size of 5 seconds using the following statistical functionals: mean, max, min, std, range, kurtosis, skewness and inter-quartile range. The video features were extracted using OpenCV [45] and included frame level luminance, intensity, Hue-Saturation-Value (HSV) color histograms and optical flow [46], which were also aggregated to 5 seconds using simple averaging. The combined features were of size $P = 1225$ for each frame.

Figure 8 shows the performance of the two models in estimating the affective dimensions for the dataset. The joint model seems to considerably outperform the independent model while estimating arousal while the independent models seem to produce better estimates of valence from the annotations. The independent model seems to perform poorly in arousal prediction, but the joint model shows a balanced performance, with the joint modeling constraint likely acting as a regularizer.

5.3 Effect of dependency among dimensions

To evaluate the impact of the magnitude of dependency between the annotation dimensions on the performance of the models, we created a set of synthetic annotations for the global model similar to Section 5.1.1. We created 10 synthetic datasets, each with constant F_k matrices across all annotators. The principal diagonal elements were fixed to 1 while the off diagonal elements were increased between 0.1 to 1 with a step size of 0.1. Similar to the previous setting, we created 100 annotators, each operating on 10 files. Note that despite the annotators having identical F_k matrices, their annotations on a given file were different because of the noise term η_k in Equation 7.

Figure 9 shows the 5-fold cross validated performance of the joint and independent models on this task. As seen in the figure, the joint model consistently outperforms the independent model in both metrics. Both the models start with similar performance when the off diagonal elements are close to zero since this implies no dependency between the annotation dimensions, and the performance of both models continues to degrade as the off diagonal elements increase. However, the joint model is able to make better predictions of the ground truth by making use of the dependency between the dimensions, highlighting the benefits of modeling the annotation dimensions jointly. Visualizations for averaged estimates of the F_k matrices from this experiment can be found in Section A.2



(a) Concordance correlation

(b) Pearson's correlation

Fig. 9: Effect of varying dependency between annotation dimensions for the synthetic model

6 Conclusion

We presented a model to combine multi-dimensional annotations from crowdsourcing platforms such as Mturk. The model assumes the ground truth to be latent and distorted by the annotators. The latent ground truth and the model parameters are estimated using the EM algorithm. EM updates are derived for both global and time series annotation settings. We evaluate the model on synthetic and real data. We also propose an artificial task with controlled ground truth and evaluate the model.

Weaknesses of the model include vulnerability to unidentifiability issues like most variants of factor analysis [36]. Typical strategies to address this issue involve adapting a suitable prior constraint on the factor matrix. For example, in PCA, the factors are ordered such that they are orthogonal to each other and arranged in decreasing order of variance.

In our experiments, the model was found to be vulnerable to unidenfiability due to label switching, which was addressed through manual judgements. We defer the task of choosing an appropriate prior constraint on F_k for future work.

Future work includes generalizing the model with Bayesian extensions, in which case the parameters can be estimated using variational inference, in addition to adding model constraints to ensure identifiability of all model parameters. Though we limit our analysis here to linear relationships between the transformation matrix F_k and the ground truth vector $\mathbf{a}_*^m$, we note that extending the model to capture non-linear relationships is straightforward. For example, the vector $\mathbf{a}_*^m$ in Equation 7 can be replaced by one that includes a non-linear dependence on $\mathbf{a}_*^m$. Providing theoretical bounds to the model performance, specially with respect to the sample complexity may also be possible since we have assumed normal distributions throughout the model.

7 Acknowledgements

The authors would like to thank Zisis Skordilis for all the helpful discussions and feedback.

References

[1] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

[4] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, vol. 101, no. 1, pp. 184–204, 2013.

[5] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the conference on empirical methods in natural language processing*, pp. 254–263. Association for Computational Linguistics, 2008.

[6] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied statistics*, pp. 20–28, 1979.

[7] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Advances in neural information processing systems*, pp. 1085–1092, 1995.

[8] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *The Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[9] K. Audhkhasi and S. Narayanan, "A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 769–783, 2013.

[10] A. Metallinou and S. S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, April 2013.

[11] M. Kipp, "Anvil-a generic annotation tool for multimodal dialogue," in *Seventh European Conference on Speech Communication and Technology*, pp. 1367–1370. ISCA, 2001.

[12] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELtrace: An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[13] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmüller, "EMuJoy: Software for continuous measurement of perceived emotions in music," *Behavior Research Methods*, vol. 39, no. 2, pp. 283–290, 2007.

[14] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with emotionml," in *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pp. 709–710, 2013.

[15] J. M. Girard and A. G. Wright, "DARMA: Software for dual axis rating and media annotation," *Behavior research methods*, vol. 50, no. 3, pp. 902–909, 2018.

[16] D. Dupre, D. Akpan, E. Elias, J.-M. Adam, B. Meillon, N. Bonnefond, M. Dubois, and A. Tcherkassof, "Oudjat: A configurable and usable annotation tool for the study of facial expressions of emotion," *International Journal of Human-Computer Studies*, vol. 83, pp. 51–61, 2015.

[17] M. A. Nicolaou, V. Pavlovic, and M. 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.

[18] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," University of California, Berkeley, Tech. Rep., 2005.

[19] 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.

[20] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. S. Narayanan, "Modeling multiple time series annotations as noisy distortions of the ground truth: An expectation-maximization approach," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 76–89, Jan. 2018.

[21] Y. E. Kara, G. Genc, O. Aran, and L. Akarun, "Modeling annotator behaviors for crowd labeling," *Neurocomputing*, vol. 160, pp. 141–156, 2015.

[22] N. B. Shah, D. Zhou, and Y. Peres, "Approval voting and incentives in crowdsourcing," *arXiv preprint arXiv:1502.05696*, 2015.

[23] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 614–622. ACM, 2008.

[24] D. Watson and L. A. Clark, "On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model," *Journal of personality*, vol. 60, no. 2, pp. 441–476, 1992.

[25] J. A. Russell and J. M. Carroll, "On the bipolarity of positive and negative affect." *Psychological bulletin*, vol. 125, no. 1, p. 3, 1999.

[26] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[27] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2376–2379. IEEE, 2011.

[28] A. Ramakrishna, R. Gupta, R. B. Grossman, and S. S. Narayanan, "An expectation maximization approach to joint modeling of multidimensional ratings derived from multiple annotators," in *Proceedings of Interspeech*, pp. 1555–1559, 2016.

[29] K. Sharma, M. Wagner, C. Castellini, E. L. van den Broek, F. Stulp, and F. Schwenker, "A functional data analysis approach for continuous 2-d emotion annotations," in *Web Intelligence*, vol. 17, pp. 41–52, no. 1. IOS Press, 2019.

[30] H. H. Harman, *Modern factor analysis*. University of Chicago press, 1976.

[31] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[32] 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, 2012.

[33] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8. IEEE, 2013.

[34] V. I. Spitkovsky, H. Alshawi, D. Jurafsky, and C. D. Manning, "Viterbi training improves unsupervised dependency parsing," in *Proceedings of the Fourteenth Conference on Computational Natural*

Language Learning, pp. 9–17. Association for Computational Linguistics, 2010.
[35] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.
[36] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research." *Psychological methods*, vol. 4, no. 3, p. 272, 1999.
[37] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, vol. 23, no. 3, pp. 187–200, 1958.
[38] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.
[39] 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)*, pp. 3091–3095. IEEE, 2018.
[40] D. Corney, J.-D. Haynes, G. Rees, and R. B. Lotto, "The brightness of colour," *PloS one*, vol. 4, no. 3, p. e5091, 2009.
[41] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, ser. SemEval '07, pp. 70–74. Association for Computational Linguistics, 2007.
[42] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," *CoRR*, vol. abs/1703.02507, 2017.
[43] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, Kobe, Japan, May 2009.
[44] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10, pp. 1459–1462. ACM, 2010.
[45] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
[46] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 366–371. IEEE, 1998.
[47] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

Shrikanth (Shri) Narayanan (StM'88-M'95-SM'02-F'09) is the Niki & C. L. Max Nikias Chair in Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical and Computer Engineering, Computer Science, Linguistics, Psychology, Neuroscience, Otolaryngology and Pediatrics, Research Director of the Information Science Institute, and director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC, he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biosignal processing and machine intelligence, and their applications with direct societal relevance. [http://sail.usc.edu]

Prof. Narayanan is a Fellow of the National Academy of Inventors, the Acoustical Society of America, IEEE, the International Speech Communication Association (ISCA), the Association for Psychological Science, the American Institute for Medical and Biological Engineering (AIMBE), and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is VP-Education for IEEE Signal Processing Society, an Editor for the Computer Speech and Language Journal and an Associate Editor for the APSIPA Transactions on Signal and Information Processing. He was also previously Editor in Chief for IEEE Journal of Selected Topics in Signal Processing and an Associate Editor of the IEEE Transactions on Speech and Audio Processing (20002004), IEEE Signal Processing Magazine (20052008), IEEE Transactions on Multimedia (2008-2011), IEEE Transactions on Signal and Information Processing over Networks (2014-2015), IEEE Transactions on Affective Computing (2010-2016), and the Journal of the Acoustical Society of America (2009-2017). He is a recipient of several honors including the 2015 Engineers Councils Distinguished Educator Award, a Mellon award for mentoring excellence, the 2005 and 2009 Best Journal Paper awards from the IEEE Signal Processing Society and serving as its Distinguished Lecturer for 2010-11, a 2018 ISCA Best Journal Paper award, and serving as an ISCA Distinguished Lecturer for 2015-16 and the Willard R. Zemlin Memorial Lecturer for ASHA in 2017. Papers co-authored with his students have won awards including the 2014 Ten-year Technical Impact Award from ACM ICMI and at several conferences. He has published over 800 papers and has been granted seventeen U.S. patents.

Anil Ramakrishna received his B.E. degree from the Visvesvaraya Technological University in 2010, M.S in Computer Science in 2014, M.S in Electrical Engineering in 2019, Ph.D. in Computer Science in 2019, all from the University of Southern California (USC). His dissertation focused on developing models for multidimensional annotations of affect. His research interests include sentiment analysis, natural language processing and machine learning. He is a member of the IEEE.

Rahul Gupta received a B.Tech. degree in Electrical Engineering from Indian Institute of Technology, Kharagpur in 2010 and a Ph.D. degree in Electrical Engineering from University of Southern California (USC), Los Angeles in 2016. His research concerns development of machine learning algorithms with application to human behavioral data. His dissertation work is on the development of computational methods for modeling non-verbal communication in human interaction. He is the recipient of Info-USA exchange scholarship (2009), Provost fellowship (2010-2014) and the Phi Beta Kappa alumni in Southern California scholarship (2015). He was part of the team that won the INTERSPEECH-2013 and INTERSPEECH-2015 Computational Paralinguistics Challenges. He is a member of the IEEE.