# Automated Action Units Vs. Expert Raters: Face off

Svati Dhamija, Terrance E. Boult
University of Colorado Colorado Springs
{sdhamija,tboult}@vast.uccs.edu

## Abstract

*User engagement is an essential component of any application design. Finding reliable methods to forecast continuos engagement can aid in creating adaptive applications like web-based interventions, intelligent student tutoring, the creation of socially intelligent human-robots, etc. In this paper, we compare observational estimates from expert raters to vision-based learning, for estimating user engagement. The vision-based approach uses automated computation of Action Units combined with an RNN.*

*Several data collection techniques have been explored in the past that capture different modalities for engagement from obtaining self-reports and gathering external observations via crowd-sourcing or even trained expert raters. Traditional machine learning approaches discard annotations from inconsistent raters, use rater averages or apply rater-specific weighting schemes. Such approaches often end up throwing away expensive annotations.*

*We introduce a novel approach that exploits the inherent confusion and disagreement in raters annotations to build a scalable engagement estimation model that learns to appropriately weigh subjective behavioral cues. We show that actively modeling the uncertainty, either explicitly from expert raters or from automated estimation with AU, significantly improves prediction over prediction from just the average engagement ratings. Our approach performs significantly better or on par with experts in predicting engagement for a trauma-recovery application.*

## 1. Introduction

> The face is most considered and regarded, as is natural from its being the chief seat of expression and the source of the voice.
>
> *Charles Darwin [10]*

Exploration of new research avenues in the fields of computer vision and affective computing have leveraged tools and techniques from the field of crowdsourcing [24, 36]. The emergence of sources like MTurk, Crowdflower, etc. to obtain external annotations for large-scale data mining are now commonplace. Most affective computing applications rely on expert raters to obtain continuous labels of affective and cognitive states [8, 32, 7]. The current gold-standard for measuring user engagement is estimated from self-reports. Questionnaires are simple and inexpensive, but they distract the user from the task at hand or increase cognitive load. Obtaining self-reports is problematic, especially for people suffering from mental-disorders whose symptoms include but are not limited to reduced ability to concentrate, trouble understanding and frequent mood-swings. Automated continuous prediction of engagement is therefore of utmost importance to numerous applications.

In recent years, psychologist have turned to web/mobile based intervention as an effective way to provide treatment and therapy for a number of mental health challenges such as depression, mood management, anxiety disorders, trauma recovery and others [3]. A number of commercial solutions such as those offered by Pacifica, Ginger.io [2, 1] have been developed for stress/anxiety management and emotional well-being as an effective way for sustainable mental health. These solutions help patients monitor their progress towards recovery through interactive questionnaires about their well-being (mood, stress, sleep patterns, etc.) and providing remote psychologist and peer-community support. Such self-support tools cannot be too generic for trauma recovery and need to autonomously adapt to the patient‘s needs based on their mental and physical state [44]. Although ample evidence exists for the clinical effectiveness of web/mobile interventions across a wide base of interventions, in case of trauma subjects who often experience emotional detachment and disengagement from mundane tasks, effective engagement with these interventions remains a significant concern [25].

Advances in machine learning, computer vision, and facial expression analysis have led to development of wide range of applications from methods to analyze student engagement [43, 26] in classrooms, measuring heart-rate in realistic conditions [22], estimating valence and arousal [29] to higher level states such as monitoring moods and depressions [37]. With the ubiquity of cameras on smartphones,

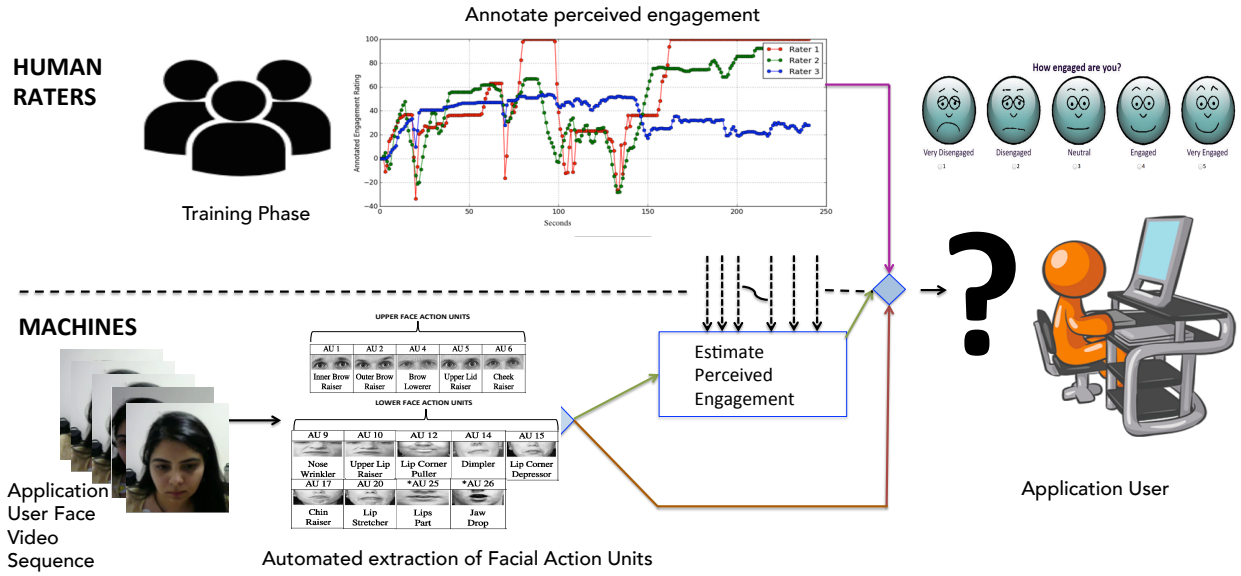



Figure 1. Designing effective applications involve understanding the engagement levels of the target population. Machine learning models that estimate user engagement have two choices: learning external annotations of perceived engagement provided by expert raters or learn vision based models from facial expression representations. The top pipeline in the above figure represents the expert raters annotation process. Raters undergo a training phase to formulate rules for the modality being measured, e.g., behaviors indicative of engagement/disengagement. Despite rigorous training efforts, the raters do not always agree upon engagement levels. The bottom pipeline is automated processes using facial feature extraction in the form of Action Units (AUs). We can directly estimate engagement from the AU or use the AUs to estimate the expert rater scores. This work compares the expert raters to automated AUs for estimating user engagement and also proposes a new scalable approach to combine the best of both worlds.

facial analysis based emotional state monitoring methods provide a scalable approach for deploying solutions [38]. Therefore, it is natural to ask: How can we develop facial analysis tools to continuously monitor subject engagement in trauma recovery applications? In this paper our visual features are Action Units (AUs), computed with OpenFace [4], which are then used as inputs to train different RNN-based models.

As noted by in prior work [11], the ground truth for engagement prediction consists of either sparse labels obtained from self-reports or continuous annotations (observational estimations) from expert external human observers. Self-reports, questionnaires completed by subject to reporting their level of engagement, provide accurate feedback about subject's well-being. However, repeated questionnaires impose a cognitive burden on trauma subjects. An alternative approach is to obtain continuous annotations from trained psychology experts who use facial and behavioral cues to make judgments about subjects engagement [16]. These judgments are standardized across observers through use of measurement tools [15]. Training psychologist to make such judgments is an iterative, cumbersome and labor-intensive process. Trauma subjects, depending on the condition, can either be very expressive or can tend to suppress emotions and thereby make video annotation process tiring, difficult and costly for raters. Furthermore, estimating inter-observer reliability is not always straight-forward and can lead to significantly varying annotations (see Figure 1). The contributions of this paper are addressing these fundamental questions:

- How robust are annotations obtained from expert human observers?
- How do human observers fare compared to machine learned features for engagement prediction?
- How can we develop vision systems that can learn and use inter-observer variability in prediction?

## 2. Related Work

This work lies at the intersection of multiple emerging areas which we briefly review and compare in this section.

**Visual Engagement Prediction:** Predicting subject engagement from facial data has been an active area of research in computer vision and affective computing literature in recent years. Monkaresi *et al*. [26] developed methods for detecting engagement when subjects were performing a structured writing activity. Hernandez *et al*. [19] developed facial features and head gesture-based method to measure the engagement of viewers while watching television. Mcduff *et al*. [24] explored the role of emotions and engagement for media applications to understand the effectiveness

of video advertising. Whitehill *et al*. [43] developed various student engagement methodologies by analyzing facial expressions and machine learning techniques. They developed binary engagement detectors to estimate high/low engagement levels in video segments (monitoring students) and found the performance of the system comparable to humans. In recent past, Kamath *et al*. [20] proposed a crowdsourced discriminative learning approach/system for e-learning and estimated student engagement. Almost all of the aforementioned techniques have been confined to student-learning or advertising. Most works also create a machine learning approach with a chosen set of features (either vision based or from external annotations), and no comparison is done for perceived engagement with automated vision-based techniques. Unlike student learning, in trauma recovery applications, subjects often do not have control over their emotions and suffer from varying degrees of disengagement raising unique challenges.

**Behavioral Intervention Technologies (BITs):** BITs are the application of behavioral and psychological intervention strategies through the use of technology features to address behavioral, cognitive and affective targets that support physical, behavioral and mental health [25]. Such technologies are employed to implement change strategies that include self-monitoring, goal-setting, skill building and others. These techniques have potential to reach large populations, who otherwise would struggle to receive such care. In recent years, such techniques have led to the development of various web/mobile based intervention techniques in academic [31] and commercial domains [2, 1]. In the domain of web-based interventions for trauma recovery, the work of Benight *et al*. is seminal and have explored in detail solutions and challenges faced when designing such solutions. More recent work by Yeager *et al*. explored in detail the role of engagement in Web-based trauma recovery [44].

**Multi-Rater Annotations:** Supervised learning methods typically consist of a label from single annotator per training sample. When working with perceived emotions, it is common practice to get feedback from multiple annotators to avoid bias. However, this process often leads to noisy labels with varying degrees of inter-rater agreements. Raykar *et al*. [30] presented a probabilistic approach for learning with data from multiple annotators and proposed algorithm that evaluates the labels obtained from different annotators and also gives an estimate of the hidden labels. Wellinder *et al*. [42] presented an approach to model multiple non-expert annotators as a multidimensional entity with variables representing competence, expertise, and bias. Our work consisted of psychology research assistants who underwent rigorous training for this task (experts). Hence, such model would not be applicable. In Bioinformatics, Valizadegan *et al*. [40] presented a consensus approach for learning with multiple annotators, wherein a separate model was trained from each annotator, and the results were fused to obtain a consensus model. In such approaches, it is hard to estimate the source of variance [28]. In the domain of affective computing, continuous measurement of perceived emotions (annotations) was popularized by Gottman *et al*. [18] with the introduction of "affect rating dial". Raters were instructed to turn the knob clockwise or counter-clockwise to report their affective experience More recently number of annotations tools have been developed to collect continuous ratings as proposed by Brugman *et al*. [6] (ELAN), Kipp *et al*. [21] (ANVIL), Nagel *et al*. [27] (EmuJoy) and others. The dataset we used for this work consisted of audio, video and physiological signals (only video signals considered in this work). For the EASE dataset that we use for this work (details in Section 4), an annotation tool developed by Jirard *et al*. called Continuous Affect Rating and Media Annotation (CARMA) was used. The tool is based on the works of Gottman *et al*. CARMA is a media annotation program that collects continuous ratings from observers with the flexibility of using data from multiple modalities, and hence, was suitable for data collection (see Figure 2, for details on annotation process).

## 3. Recurrent Neural Network (RNN)

Recurrent Neural Networks are well suited for problems involving sequential information and have found relevance in a wide range of tasks, in recent years, ranging from machine translation, video classification, natural language processing, image captioning, visual question answering and others. RNNs learn a representation from a sequence of input vectors [34, 17]. Engagement prediction from facial videos is inherently a sequence prediction task. There are series of facial expressions (characterized by facial action units) that the system sees and it has to make predictions about the level of engagement. Hence, we seek to develop algorithms and models for engagement prediction with RNNs.

A recurrent neural network (RNN) is a neural network that consists of a hidden state $h$ which operates on a variable-length sequence $\mathbf{x} = (x_1, \ldots, x_T)$. At each time step $t$, the hidden state $h_t$ of RNN is updated by:

$$h_t = \theta\phi(h_{t-1}) + \theta_x \mathbf{x}_t \tag{1a}$$

$$y_t = \theta_y \phi(h_t) \tag{1b}$$

where $\phi$ is the non-linear activation function and $y$ is the target output unit. The activation function $\phi$ may be as simple as an element-wise logistic sigmoid function and as complex as a long short-term memory (LSTM) unit [14]. In this work, the input sequence $\mathbf{x}$ is composed of either facial action units $\mathbf{x}_{AU}$ or expert ratings $\mathbf{x}_{HA}$. The target output $y$ for a given video segment consist of self-reports obtained from trauma recovery subjects.

During the back propagation through recurrent units, the derivative of each node is dependent of all the nodes which processed earlier. To compute $\frac{\partial h_t}{\partial h_k}$ a series of multiplication from $k = 1$ to $k = t - 1$ is required. Assume that $\dot{\phi}$ is bounded by $\alpha$ then $||\frac{\partial h_t}{\partial h_k}|| < \alpha^{t-k}$

$$\frac{\partial E}{\partial \theta} = \sum_{t=1}^{t=S} \frac{\partial E_t}{\partial \theta} \tag{2}$$

$$\frac{\partial E_t}{\partial \theta} = \sum_{k=1}^{k=t} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial \theta} \tag{3}$$

$$\frac{\partial h_t}{\partial h_k} = \prod_{i=k+1}^{t} \frac{\partial h_i}{\partial h_{i-1}} = \prod_{i=k+1}^{t} \theta^T diag[\dot{\phi}(h_{i-1})] \tag{4}$$

where $E$ is the loss of $t^{th}$ layer. A solution to this problem is to use gated structures. The gates can control back propagation flow between each node [14, 39, 17].

## 4. EASE Dataset & Annotations

We use Engagement, Arousal and Self-Efficacy (EASE) dataset [11] for this work. We briefly review the dataset followed by a description of the extensions used in this work. EASE dataset is a unique dataset consisting of videos, audio and physiological signals collected from subjects while they worked on a web-intervention for trauma recovery. The subjects were recruited from various clinical centers (Veterans Trauma Clinic, family health centers, and others), came from diverse backgrounds such as active duty service members and their families, veterans suffering from Post Traumatic Stress Disorders, victims of domestic violence, etc. [5]. Participants consisted of total 110 subjects with 88 Female, 17 Male, 5 did not specify in the age group of 18-79 years, with 80% being under the age of 46. The dataset consists of two modules: Relaxation and Triggers. The relaxation module presents the user with video demonstrations of various exercises like breathing, muscle relaxation, etc. The triggers module educates the user about trauma symptoms and prevention. The total size of the dataset consists of 8M+ video frames. For this work, we use a subset of the original EASE dataset for which expert rater annotations are available (the Triggers module) [1].

Expert annotations by psychologists is a multi-step process [16]. In the first phase, graduate students/post-doctoral fellows with a background in clinical psychology undergo training to understand the process of labeling. As the goal of the process is to annotate perceived engagement by external observers, annotators were asked to label for "How engaged does the subject appear?". Annotators also formulated the guidelines on how to perceive engagement. The guidelines included if the subject showed expressions like "brow furrow", "squinting focus", "fast reading", "head scanning" etc. were signs that the subject is engaged. Similarly, signs like "gaze avoidance", "fidgety movement", "dropping eyes", "closing eyes when tasks do not require" etc. were signs of the subject being disengaged. These guidelines were created by experienced clinical psychologists who work with trauma subjects. Following this familiarization process, annotators rated engagement levels between -100 (very disengaged) to 100 (very engaged). The annotations were obtained by post-analysis of video sequences (rather than live annotations) due to ease of collection. We ensured that not all annotators rate same set of videos to avoid annotator perceived engagement bias. The annotations were obtained every second. When labeling videos, the audio was turned off, and raters were instructed to label engagement based on appearance. As can be inferred, this is a cumbersome, costly and painstakingly slow process due to challenges associated with recruitment of expert annotators, training and labeling large corpus of data. In this work, we use 6K+ annotations from a total of 183K+ frames (see Table 1).

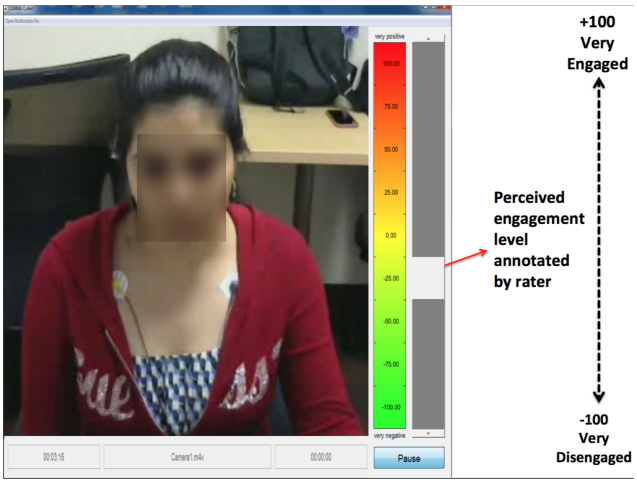



Figure 2. CARMA tool: The figure above illustrates the tool used for obtaining expert annotations from trained psychologists on EASE Dataset [11]. expert raters are presented with a video feed of trauma subject (shown above) undertaking a particular task. Annotators provide a rating between -100 (very disengaged) to +100 (very engaged) to record the perceived engagement levels of the trauma subject. The ratings are recorded as continuous annotations sampled on a per second basis. Note: the face of the subject is blurred due to IRB restrictions.

## 5. Machine Learning Models

We now describe the methodology adopted to develop models trained from data derived from expert raters and machines (facial action units). Training data from expert

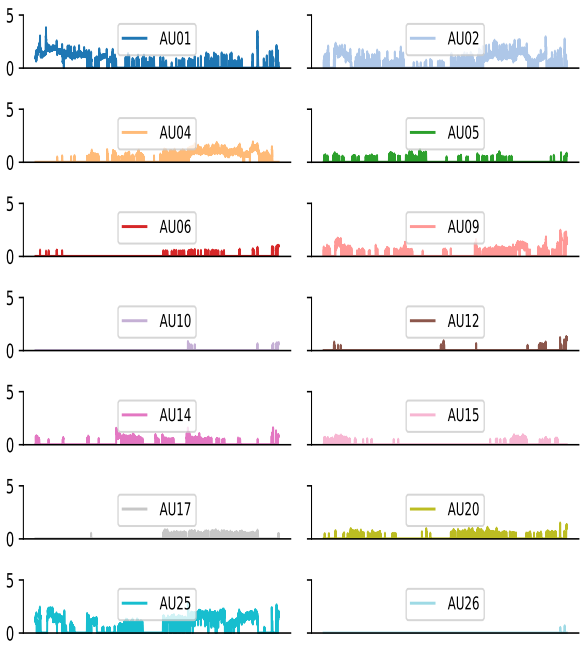
[1] Code and data for experiments can be found at `http://vast.uccs.edu/~sdhamija/`

| Dataset Details | No. of Instance |
|---|---|
| Subjects | 54 |
| Self-Reports | 204 |
| Videos | 65 |
| Frames | 183.6K |
| Annotations | 6120 |
| **Distribution of Self-Reports** | |
| Very Engaged - 5 | 51 |
| Engaged - 4 | 94 |
| Neutral - 3 | 50 |
| Disengaged - 2 | 5 |
| Very Disengaged - 1 | 4 |

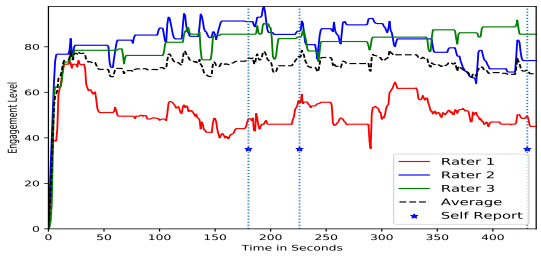
Table 1. Data distribution from EASE subset: User face videos and external annotations were available from 54 trauma subjects. There were a total of 64 videos from both Session1 and 2 combined. In all our experiments we use 30 seconds of video segment prior to engagement response making a total of 183.6K video frames at 30fps. expert raters annotated the videos on a per second basis resulting in 6120 annotations used in our work. Engagement responses were also obtained from 204 self-reports on a scale of 1-5 from 1 being "Very Disengaged" to 5 "Very Engaged". The bottom part of the table shows distribution of responses based on self-reported user engagement.

raters was obtained from engagement ratings, as described in the previous section. The data is obtained at 1 rating per second. The machine derived data for comparison was obtained by extracting facial action units from video frames of EASE dataset. Although there are a number of software available for extracting facial landmark points and facial action units we use the recent work on OpenFace [4] proposed by Baltrusaitis *et al*. In our prior work on contextual engagement [11], the AUs extracted from OpenFace consisted of both intensity-based and presence-based AUs, making a total of 20 feature dimensions. For all automated models in this work, we focus only on intensity-based AUs to reduce the effect of combining categorical and numeric features. This reduces our input feature dimensions from 20 to 14. Intensity-based AUs are generated from OpenFace on a 0 to 5 point scale (not present to present with maximum intensity). The list of AUs used in this paper are as follows: Inner Brow Raiser, Outer Brow Raiser, Brow Lowerer, Upper Lid Raiser, Cheek Raiser, Nose Wrinkler, Upper Lip Raiser, Lip Corner Puller, Dimpler, Lip Corner Depressor, Chin Raiser, Lip Stretcher, Lips Part, Jaw Drop. AUs are extracted at 30 fps.

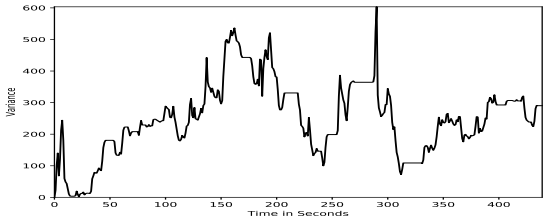
We represent both human annotations $\mathbf{x}_{HA}$ and action units $\mathbf{x}_{AU}$ obtained from facial data as a series of inputs and associated subject engagement self-reports $y$ as target outputs. Hence, we model it as a sequence learning problem and develop recurrent neural network based learning frame-




(a)




(b)




(c)

Figure 3. The figure above shows the automated AUs, rater annotations and subject self-reports for a single subject from the EASE dataset. The plot 3a shows 14 intensity based action units extracted from OpenFace. The y-axis is the intensity of the AU. The plot 3b shows perceived engagement from 3 expert raters and their average. Notice the vertical dotted line in the plot is user self-report. Since the self-reports are on a scale of 1-5, which is different from the expert rater annotation range of -100 to +100, we cannot plot them in the same manner. All three self-reports submitted by the subject in the plot presented above were "5" and the video segment is from a highly engaged subject. The variance of the three raters are also shown in the plot 3c. The x-axis in all plots is the time in seconds. These plots are obtained from an 7.3 minute video segment. Notice how the inter-rater variance changes radically.

work. We conduct a number of experiments (as described below) to gain deeper insight into the relationship between models learned from machine extracted features and expert human ratings. In order to have a fair comparison of all four models mentioned below, we use only 30-second segments prior to self-reported engagement for feature extraction.

**Raters Average:** As noted earlier, for each video three ratings were obtained. We compute average over the three ratings and train a sequential RNN model as a regressor over self-reported engagement levels to learn temporal dependencies. Following this process, we obtain model learned from expert annotations.

**Raters Average + Variance:** We extend the Raters Average model by augmenting it with variance computed over each instant, along with the average over annotations. Similarly, we learn a multi-input (average & variance) single output RNN model to predict self-reported engagements. We term this approach as Raters Average + Variance.

**Automated AUs:** We use facial action units extracted from video frames of trauma subjects as multi-dimensional inputs and train sequential RNN model as a regressor over self-reported engagement levels. The model is trained to integrate long-term temporal variations in facial expressions via facial action units. We term this approach as Automated AUs.

**Facial Average and Variance Estimations (FAVE):** We extend the Automated AUs based approach through a multi-step approach for engagement self-report prediction. In the first step, we use multi-dimensional action units to predict continuous ratings obtained from expert annotations, more specifically ratings average and variance. We train RNN models to learn temporal dependencies to jointly predict expert raters average and variance. In the second step, we use "machine estimated" human annotations (average and variance ratings obtained from the first step) and train a regressor over engagement self-reports for each subject.

## 6. Model Training/Tuning

Deep learning models like RNN have several hyper-parameters that can impact performance. We analyze three basic hyper-parameters namely batch-size, number of epochs and learning rate. Selection of batch size is especially important because of the EASE subset engagement self-reports data distribution mentioned in Table 1. A small batch-size selection may learn subject-specific bias during training, and a large batch size might cause overfitting on the entire dataset. Similarly, if the learning rate is small the learning algorithm may get stuck in local minima, and if too high then the algorithm may bypass the local minima and never converge. In order to find an effective batch size and learning rate, we first split the 204 self-reports randomly to create a training set comprising of 90% of the total samples

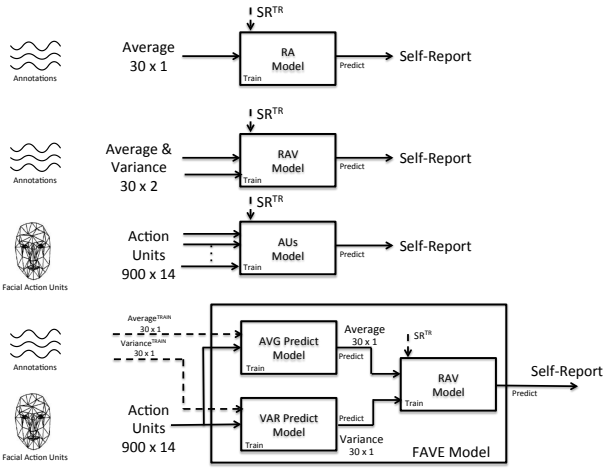



Figure 4. Machine Learning Models: The four machine learning models are illustrated in the figure above along with the input and output dimensionality for each. The input to all models are obtained from 30 second video sequences of 30fps for each self-report (204). The two models on the top, RA and RAV are trained with expert annotations using rater's average and rater's average & variance respectively. The expert annotations are on a per second basis. The AUs model was trained with automated intensity based AUs extracted from OpenFace on a frame level. The FAVE model was a two step model which learned to estimate expert annotators average and variance from automated AUs and used the estimated average and variance for final predictions. The final predictions for all four models are user self-reports.

(184) with the remaining samples in the validation set. We perform a grid search on all four learning models for batch sizes of 20,40,60 and 80 along with learning rates ranging from 1,0.1,0.01,0.001 and 0.0001. We created convergence plots with the training and validation L2 losses for 1500 epochs. For the three models, Raters Average Raters Average + Variance and Automated AUs mentioned in Section 5, we found that the algorithms converged for a batch size of 40 samples and a learning rate of 0.1. Also the training and validation losses after approx. 5-7 epochs. Since our hypothesis was to compare expert raters to machine-generated features, we keep the model parameters to be constant for all four models. We select 15 epochs for all models. At 15 epochs the L2 validation loss was minimum and stable. The constant model parameters enable us to compare the effects of feature inputs directly rather than study algorithm effects.

As mentioned in Section 5, the FAVE model was a two-step process. The first step in the aforementioned model had a total of 6120 samples, i.e., the number of expert rated annotations mentioned in Table 1. The convergence plot for the model with varied batch sizes of 1000,1500 and 2000 and a fixed learning rate of 0.001 are shown in Figure 5. Notice that even though the training and validation errors do not converge for batch sizes 1500 and 2000, the L2 loss converges for a batch size of 1000 around 15 epochs. As

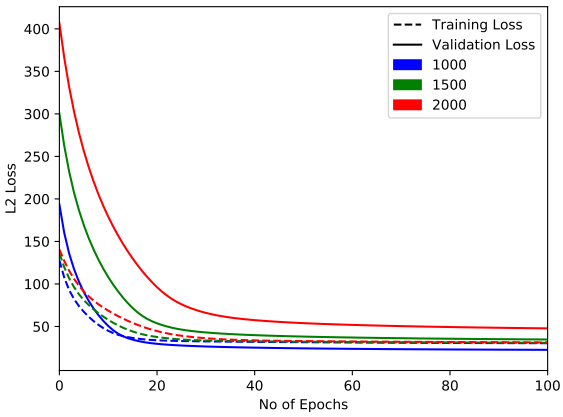



Figure 5. Model tuning for FAVE : The above plot shows the training and validation loss for 100 epochs and training batch sizes of 1000,1500 and 2000. The L2 loss stabilized at a learning rate of 0.001. Notice that the lowest training loss for all three batch sizes is approximately the same. The lowest validation loss is for a batch size of 1000. Therefore, for our FAVE model, we selected 100 epochs and reported results on a batch size of 1000.

an initial test, we tried to jointly estimate raw average and variance of expert raters using automated AU data extracted from 30 frames i.e. 1 second. However, we noticed that the model learned the bias of the ground-truth labels. This finding was not surprising as the engagement estimation for self-reports was an unbounded regression problem and the range of the ground-truth was high (-100 to +100). As a next step we normalized the raters average by the maximum value of expert rater annotations, such that the estimates lie in the range of -1 to +1 and variance from 0 to 1. Once all the model parameters are fine tuned and fixed, we use Leave One Subject Out methodology to evaluate the results from all four models.

## 7. Results

Because we have very limited training/testing data, we used Leave One Subject Out (LOSO) cross-validation methodology for computing results and computed mean squared error (MSE) along with standard deviation (SD) between the predicted engagement levels and ground-truth self-reported engagement levels. That is, for each of the four different machine learning models, we train 54 different regressors, using each to predict the results on missing sample. The results are summarized in Table 2.

The Raters Average model used average annotations obtained from 3 raters to predict self-reports. With an MSE of 1.1499, it was predicting, just not that well. Prediction using average expert rater data was the worst performing model in terms of MSE and SD.

Perceived observer annotations vary significantly from each other depending on how annotators perceive suggested guidelines for annotations and facial expressions. This variation was illustrated back in Figure 3, which shows raw annotation, average, variance and facial action units for a sample video from EASE dataset. In case of the Raters Average + Variance model, the system was trained with using average expert annotations and corresponding annotation variance to predict self-reported engagement levels. The model training using rater mean and variance performs significantly better than Raters Average probably because it can use the inter-rater variance and associated uncertainty in weighting the prediction for its sequential representation. It can naturally learn to place greater emphasis on samples where the raters agree.

For the Automated AUs model, 14-dimensional inputs per frame, for 30 seconds worth of data provide a high-dimensional input for which the RNNs learned to directly predict engagement self-reports. We note that prediction using the machine extracted features perform statistically significantly better (p=0.02) than predict engagement levels when compared to expert annotators. Thus the costly and labor-intensive annotation process is may be unnecessary. We also note the reduction in SD of results suggesting the stability of predictions.

Finally, noting these improvements in engagement prediction performance from machine extracted features, and also the improvement from using mean and variance we explore combining the two ideas expecting improved performance. We developed a two-step process: first to estimate mean and variance of annotations from action units followed by using these annotations predictions (average and variance) to as inputs to predict engagement self-reports with the FAVE model. We note the major reduction in MSE and SD over both Raters Average + Variance and Automated AUs models. Moreover, statistical evidence (p=0.025) is found using at 2-paired 1-tailed t-test that the combined model is better than the basic Automated AUs model. Why does this happen and what are the implications of these results? First, the total number of learned parameters is smaller for the two-stage model. Hence there is a lower risk of overfitting. Secondly, there is the possibility is that machines are more consistent at rating same set of facial expressions to a particular engagement level. This consistency propagates in the average & variance of estimated annotations, which in turn leads to more stable engagement predictions. With each associated engagement level, the model is able to predict uncertainty (variance) in the corresponding engagement levels. Hence, such two-step process as investigated for FAVE is ideal for scaling annotations for a large number of video segments without having to annotate each video by expert annotators. Furthermore, it lays a foundation to explore active learning techniques [41, 12] to aid in making the annotation process more

| Machine Learning Model | Average MSE | Standard Deviation |
|---|---|---|
| Raters Average | 1.1499 | 1.6610 |
| Raters Average + Variance | 0.8240 | 1.3583 |
| Automated AUs | 0.8131 | 1.2994 |
| **FAVE** | **0.7554** | **1.2976** |

Table 2. This table summarizes the results obtained from experiments described in Section 5 using Leave One Subject Out (LOSO) methodology. The first column represents the inputs to the temporal sequence learning model for user engagement estimation. The second and third columns show the average mean squared error and standard deviation respectively. Rows 1 and 2 above are models created from expert rater annotations and 3 and 4 from automated action units. Automated AUs shows significant reduction in error from Raters Average and Raters Average + Variance using a 2-paired tailed t-test (p=0.02).

consistent by developing annotator specific adaptive active learning [23].

## 8. Discussion

As noted earlier, collecting annotated data for trauma recovery applications is extremely expensive. Data collection process adds significant cognitive burden on subjects, leading to very cumbersome rater recruitment process. The annotation process involves finding a large population of subject expert annotators, training them and ensuring the quality of the data. Furthermore, due to the need of a specialized setup, costs and associated IRB processes such data collection can be done only at approved labs. Hence, there is a need to develop systems that will learn, scale and maximize the usability from limited annotated data. While we show them in the context of a trauma recovery application, the ideas should apply to a wide range of vision-based affective computing.

In this work, we used only AUs extracted from 30 seconds prior to self-reports and also expert ratings from 30 seconds for a fair comparison. We noticed performance saturation beyond 30-second video segments with RNN networks and hence there is a need to explore more sophisticated temporal deep learning methods that can learn over longer time durations. Hierarchical and multi-scale temporal representations like the ones proposed by Chung *et al*. [9] are promising next steps to develop more accurate, flexible and scalable engagement estimators. Beyond representations, there is a need to develop better annotation tools that reduce the cognitive burden of annotators by taking into account rater reaction time and rater variance correlated with action units [35].

We presented an engagement prediction system that can learn from expert-generated engagement annotations and machine-generated facial data. Our results show that automated systems with extracted facial action units perform better than expert annotated data. We captured the non-agreement or confusion of raters and used it to build a more accurate engagement estimator.

Further, we enhanced the estimator with associated prediction uncertainty by incorporating predicted variance for the given video segment. For fair comparison in training, all our models used the same temporal window for data. However, we have significantly more expert labeled data. The power of the FAVE model could be advanced even further by using more training data, greater than the 30-second segments selected in this work. The increase in training data has a strong potential to create a better engagement estimator.

Surprisingly, our results also revealed that combining expert raters annotations with the automated action units in a two-step process has the potential to estimate user engagement in a better way. A possible reasoning for the enhanced performance in FAVE could be attributed to the reduction in input feature dimensions. The Automated AUs model had an input feature dimension of 900 for each input sample. The FAVE model, on the other hand, had 30 features for a single sample which is a drastic reduction from Automated AUs Ḟusion techniques like FAVE of estimating expert annotations from automated AUs can help scale to different datasets where only self-reports are available, decreasing the cost of training raters and annotation time [13]. Scheirer *et al*. [33] proposed a system for perceptual annotations to leverage the abilities of human subjects to build better machine learning systems. This work could be extended, to instead incorporate rater confusion and non-agreement in loss functions of the learning system to build more accurate engagement estimators.

## References


[1] Pacifica labs. `https://www.thinkpacifica.com/LastAccessed12/1/17`.

[2] Ginger.io. "`http://ginger.io` Last Accessed 12/1/17".

[3] E. Anthes. Pocket psychiatry: mobile mental-health apps have exploded onto the market, but few have been thoroughly tested. *Nature*, 532(7597):20–24, 2016.

[4] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[5] C. Benight, K. Shoji, C. Yeager, A. Mullings, S. Dhamija, and T. Boult. Changes self-appraisal and mood utilizing a web-based recovery system on posttraumatic stress symptoms: A laboratory experiment. In *International Society for Traumatic Stress Studies*. ISTSS, 2016.

[6] H. Brugman, A. Russel, and X. Nijmegen. Annotating multimedia/multi-modal resources with elan. In *LREC*, 2004.

[7] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2017.

[8] E. Camilleri, G. N. Yannakakis, and A. Liapis. Towards general models of player affect. In *Affective Computing and Intelligent Interaction (ACII), 2017 International Conference on*, 2017.

[9] J. Chung, S. Ahn, and Y. Bengio. Hierarchical multiscale recurrent neural networks. *arXiv preprint arXiv:1609.01704*, 2016.

[10] C. Darwin and P. Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[11] S. Dhamija and T. Boult. Exploring contextual engagement for trauma recovery. *CVPR Workshop on Deep Affective Learning and Context Modelling*, 2017.

[12] R. Di Salvo, C. Spampinato, and D. Giordano. Generating reliable video annotations by exploiting the crowd. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[13] S. K. D'Mello. On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing*, 7(2):136–149, 2016.

[14] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. In *9th International Conference on Artificial Neural Networks: ICANN '99*. IET, 1999.

[15] J. Girard. Carma: Software for continuous affect rating and media annotation. *Journal of Open Research Software*, 2(1), 2014.

[16] J. M. Girard and J. F. Cohn. A primer on observational measurement. *Assessment*, 23(4):404–413, 2016.

[17] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[18] J. M. Gottman and R. W. Levenson. A valid procedure for obtaining self-report of affect in marital interaction. *Journal of consulting and clinical psychology*, 53(2):151, 1985.

[19] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang. Measuring the engagement level of tv viewers. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.

[20] A. Kamath, A. Biswas, and V. Balasubramanian. A crowdsourced approach to student engagement recognition in e-learning environments. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[21] M. Kipp. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[22] X. Li, J. Chen, G. Zhao, and M. Pietikainen. Remote heart rate measurement from face videos under realistic situations. *Computer Vision and Pattern Recognition*, 2014.

[23] X. Li and Y. Guo. Adaptive active learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866, 2013.

[24] D. J. McDuff. *Crowdsourcing affective responses for predicting media effectiveness*. PhD thesis, Massachusetts Institute of Technology, 2014.

[25] D. C. Mohr, M. N. Burns, S. M. Schueller, G. Clarke, and M. Klinkman. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. *General hospital psychiatry*, 35(4):332–338, 2013.

[26] H. Monkaresi, N. Bosch, R. Calvo, and S. D'Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Trans. on Affective Computing*, 2017.

[27] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmüller. Emujoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, 39(2):283–290, 2007.

[28] Q. Nguyen. *Efficient learning with soft label information and multiple annotators*. PhD thesis, University of Pittsburgh, 2014.

[29] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[30] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

[31] S. C. Reid, S. D. Kauer, S. Hearps, A. H. Crooke, A. S. Khor, L. A. Sanci, and G. C. Patton. A mobile phone application for the assessment and management of youth mental health problems in primary care: a randomised controlled trial. *BMC Fam Pract*, 12(1):131, 2011.

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

[33] W. J. Scheirer, S. E. Anthony, K. Nakayama, and D. D. Cox. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1679–1686, 2014.

[34] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[35] J. Schöning, P. Faion, G. Heidemann, and U. Krumnack. Providing video annotations in multimedia containers for visualization and research. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 650–659. IEEE, 2017.

[36] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Workshop on Crowdsourcing for Search Evaluation, SIGIR 2010.*, 2010.

[37] G. Stratou and L.-P. Morency. Multisense context aware nonverbal behavior analysis framework: A psychological distress use case. *IEEE Transactions on Affective Computing*, 8(2):190–203, 2017.

[38] M. Suk and B. Prabhakaran. Real-time facial expression recognition on smartphones. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 1054–1059. IEEE, 2015.

[39] S. Valipour, M. Siam, M. Jagersand, and N. Ray. Recurrent fully convolutional networks for video segmentation. In *Ap-*

*plications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 29–36. IEEE, 2017.

[40] H. Valizadegan, Q. Nguyen, and M. Hauskrecht. Learning classification models from multiple experts. *Journal of biomedical informatics*, 46(6):1125–1135, 2013.

[41] C. Vondrick and D. Ramanan. Video annotation and tracking with active learning. In *Advances in Neural Information Processing Systems*, pages 28–36, 2011.

[42] P. Welinder, S. Branson, P. Perona, and S. J. Belongie. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pages 2424–2432, 2010.

[43] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan. Faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Trans. on Affective Computing*, 5(3):86–98, 2014.

[44] C. Yeager. Understanding engagement with a trauma recovery web intervention using the health action process approach framework. *PhD Thesis*, 2016.