Exploring Contextual Engagement for Trauma Recovery

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

Abstract

A wide range of research has used face data to estimate a person's engagement, in applications from advertising to student learning. An interesting and important question not addressed in prior work is if face-based models of engagement are generalizable and context-free, or do engagement models depend on context and task. This research shows that context-sensitive face-based engagement models are more accurate, at least in the space of web-based tools for trauma recovery. Estimating engagement is important as various psychological studies indicate that engagement is a key component to measure the effectiveness of treatment and can be predictive of behavioral outcomes in many applications. In this paper, we analyze user engagement in a trauma-recovery regime during two separate modules/tasks: relaxation and triggers. The dataset comprises of 8M+ frames from multiple videos collected from 110 subjects, with engagement data coming from 800+ subject self-reports. We build an engagement prediction model as sequence learning from facial Action Units (AUs) using Long Short Term Memory (LSTMs). Our experiments demonstrate that engagement prediction is contextual and depends significantly on the allocated task. Models trained to predict engagement on one task are only weak predictors for another and are much less accurate than context-specific models. Further, we show the interplay of subject mood and engagement using a very short version of Profile of Mood States (POMS) to extend our LSTM model.

1. Introduction

Engagement is a critical component of student learning, web-based interventions, commercial applications for marketing, etc. and face-based analysis is the most successful non-invasive approach for engagement estimation [36, 35, 53, 41, 59]. Techniques that involve face-based estimation of the six basic emotions are considered to apply to any situation, i.e. nearly universal. It is natural to ask if we can build a universal engagement predictor as well. This paper examines the role of "context" in engagement, in particular, engagement in trauma treatment. As shown in Fig 1, depending on the task, the same facial expression can be interpreted as engaged or disengaged.

Before getting deeper into the issues of context and engagement we layout our application space: mental trauma treatment. Each year, over 3 million people in the United States are affected by post-traumatic stress, a chronic mental disorder. Moreover, mental trauma following disasters, military service, accidents, domestic violence and other traumatic events is often associated with adversarial symptoms like avoidance of treatment, mood disorders, and cognitive impairments. Lack of treatment for serious mental health illnesses annually cost $193.2 billion in lost earnings [28]. Providing proactive, scalable and cost-effective web-based treatments to traumatized people is, therefore, a problem with significant societal impact [4].

Amongst the currently available e-health interventions, evidence to support the clinical effectiveness of most interventions exists, however, patient engagement with these interventions is still a major concern [36, 35, 53]. Such interventions measure user engagement from infrequent questionnaire's. Self-reported user engagement has been found, in many psychology studies, to be highly correlated with outcomes [12, 18, 16]. Research suggests that personalization and automated adaption in self-help websites can positively aid people with mental health issues and advance mental health computing [7]. In this work, we show that computer vision and deep-learning-based techniques can be used to predict user engagement from webcam feeds with content. Once we have tools for reliable engagement measurement during an intervention, the website and task can adapt to enhance or maintain engagement and recovery. We further prove that context-specific modeling is more accurate than a generic model of user engagement.

Emotions are well studied with multiple products/systems to estimate emotional response from face data in a context-free manner. While named emotions are important, the basic emotion categories fail to capture the complete richness of facial behavior. Furthermore, most studies have elicited emotional responses via a stimulus; prototypical expressions of basic emotions occur relatively

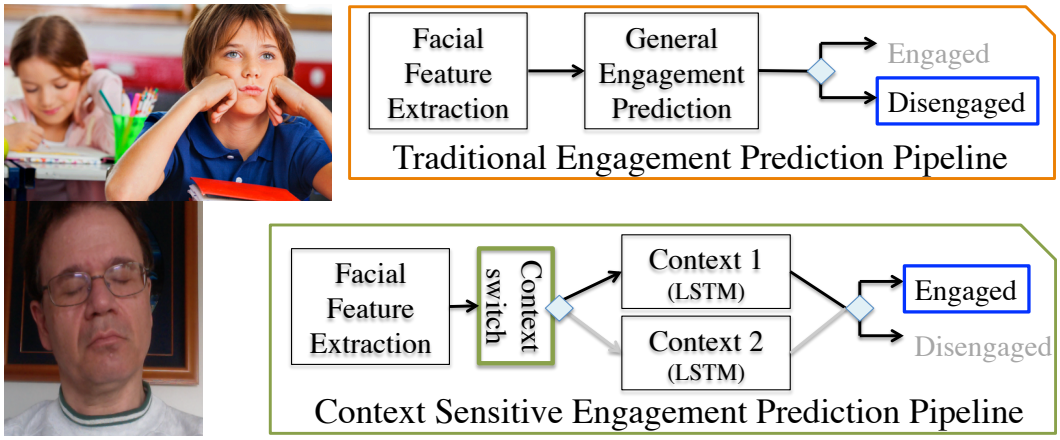

Figure 1. Consider the images on the left. Which subjects are engaged and which are disengaged? Would you change your answer if you knew one had a task of doing a relaxation exercise? What if it was reading web content, watching a video or taking a test? **We contend that face-based engagement assessment is context sensitive.** Traditional engagement prediction pipelines based on facial feature extraction and machine learning techniques learn a generic engagement model, and would consider the face in lower left disengaged. In trauma recovery, individuals are often advised to do particular exercises, e.g. self-regulation exercises where the task involves the subject to "close your eyes, relax and breathe". The image on the lower left is a highly engaged subject. Hence, there is a need to re-visit existing facial-expression-based engagement prediction techniques and augment them with the context of the task at hand. As shown in bottom right, this work develops context-sensitive engagement prediction methods based on facial expressions and temporal deep learning.

infrequently in natural interaction and even less so while people operate on a web-based intervention. Facial analysis gives strong clues about the internal mental state, e.g., a slight lowering of the brows or tightening of the mouth may indicate annoyance. De la Torre and Cohn [11] did seminal work in showing that face and voice are effective predictors of important psychological states, in particular for analysis of depression. Thus, to create engagement prediction model for trauma recovery, we use subtle facial movements or action-units rather than emotional responses as an intermediate representation.

The ground truth for the development of engagement prediction systems generally comes in one of two forms: sparse labels obtained from self-reports [15] or observational estimations from external human observers [59]. Generally, self-reports are collected from questionnaire's in which subjects report their level of engagement, attention, distraction, excitement, or boredom; in observational estimates, human observers use facial and behavioral cues to rate the subjects apparent engagement. This work uses self-reported engagement levels, which does not presume that face data necessarily reflects engagement. We evaluate our machine learning model on a dataset, described in detail in Sec 3, which is collected from trauma subjects while they work on a recovery website at `http://ease.vast.uccs.edu/`. The dataset comprises of hundreds of videos collected from 110 subjects with multiple video-synchronized self-reports of levels of engagement.

The contributions of this paper are as follows:

1. The first exploration of engagement in two contextually different tasks within the recovery regime: "Relaxation" and "Triggers". The associated dataset of AUs and engagement data will be publicly released.
2. Developing automated engagement prediction methods based on automatically computed AUs using LSTMs. We train/test this on both subtasks within trauma recovery.
3. We build context-specific, cross-context and mixed prediction models and show the importance of context in predicting engagement from facial expressions.
4. Exploring the relationship of a subject's mood as an initialization parameter for engagement estimation. User mood state is retrieved from a very short form of the Profile of Mood States (POMS) questionnaire's asked before the beginning of task [49]. Our experiments demonstrate that contextual engagement models show improvement in engagement prediction performance by combining AUs with POMS data.

While this paper addresses engagement in the context of treatment of mental trauma, we conjecture that our observations of contextual engagement hold for not just trauma recovery but also for other domains. For example, in student learning, a student looking up could be engaged if the current task is listening to lecture or watching something, or could be disengaged if the task is to be reading or taking a test.

2. Related work

The work presented in this paper is related to research from multiple communities such as context-models in vi-

sion, web-based intervention methods, trauma recovery, facial action units and expression analysis, automatic engagement recognition for student learning, deep-network-based sequence learning methods and others.

Context in vision research: Context-sensitive models have a wide range of uses in computer vision. Context, in general, can be grouped into feature level context, semantic relationship level context, and prior/priming information level context [58]. The feature and semantic levels address context within the scene/image, i.e., "context" implies spatial or visual context, and often a goal is to estimate that local/global from the data and to use that to help in primary vision task, e.g. [44, 63, 68, 55, 2, 33, 42]. There are also non-visual semantic-relationship level contexts, such as camera/photo metadata, geographical and temporal information, event labels, and social relationships pulled from social media tags [33].

Contextual priors are different as they are inputs to the system, e.g., using label/attribute priors [20, 46], 3D geometry priors [62], and scene/event priors [61]. In almost all of these works, using context improved vision-based estimation, and we hypothesize context will improve for engagement evaluation as well. The role of context for user engagement has been explored in affective computing literature [34] for designing intelligent user interfaces for office scenarios. In human-robot interaction [45, 9, 8] have shown additional knowledge of user context can better predict an affective state of the user and demonstrating that engagement prediction is contextual and task dependent.

In this work "context" is not a feature or semantic data in the image; contextual engagement is about the expected behaviors the system should be observing in engaged subjects. Because the system determines what is displayed and expects a particular user activity, the "context" is known *a priori*, i.e. at a high-level, we have a contextual prior with probability 0 or 1. Thus, our goal is not to estimate context for engagement or incorporate prior probabilities into a model, but rather to learn context-specific models and use them to predict user engagement. Inferring "context" for engagement may still be possible for the video, but it is beyond the scope of this paper and likely unnecessary in web-based settings.

Web-based trauma recovery: Researchers such as Bunge *et al*. [6] and Schueller *et al*. [48] have explored Behavioral Intervention Technologies (BITs) to augment traditional methods of delivering psychological interventions, of face-to-face (F2F) in one-to-one psychotherapy sessions, in order to expand delivery models and/or increase the outcomes of therapy. Such work has led to the development of various web-based intervention platforms with an aim to provide cost-effective, large-scale services and quality healthcare. Notable among these are works of Macea *et al*. [35], Mehta *et al*. [40] and Strecher *et al*. [54]. In the domain of web-based interventions for trauma recovery, the works of Benight *et al*. [5] and Shoji *et al*. [50] explored the role of engagement using voice analysis. The psychology study by Yeager *et al*. [64] explored in detail the role of engagement in Web-based trauma recovery. Influenced by these works, we postulate the need for the development of computer vision and machine learning-based methods for automated engagement prediction in the domain of web-based trauma recovery.

Student learning: The majority of research to predict automated engagement has been limited to the field of education where learning algorithms are built to determine student engagement from behavioral cues like facial expressions, gaze, head-pose, etc. [43, 41, 59, 22]. These works primarily rely on extracting facial features and developing machine-learning-based approaches to identify engagement activity of students in classroom performing various tasks,e.g., reading/writing, etc. The subjects are often assumed to be co-operative with control over their emotions and monitored by an external actor e.g. the teacher. One of the notable differences in these works and data collected from trauma subjects is that subject co-operativeness varies significantly, depending on the severity of mental illness and the task (self-regulation exercises) that they are assigned, leading to multiple challenges in applying methods from student learning directly [47].

Emotion recognition: The popularity of interactive social networking, and the exhaustive use of webcams has facilitated the understanding of human affect or emotions by analyzing responses to internet videos [56, 39, 19] and video-based learning[37, 67]. However, these methods are primarily geared towards developing methods to identify spontaneous emotions (e.g."happy', "sad" etc.). Due to their relevance to digital advertising and the availability of advanced learning techniques that are capable of exploring temporal dependencies, facial expression analysis to recognize emotional states is emerging as a new area of exploration [38]. Our research is also a step towards understanding the interpretation and use of "engagement" for real-world applications.

Facial features: Extracting facial features for expression analysis, facial action unit coding, face recognition, and face tracking has a rich history with some notable works [14, 29, 13]. Automatic detection of facial action units has proved to be an important building block for multiple non-verbal and emotion analysis systems [31]. For this work, we rely on leading facial landmark and action unit detection work, OpenFace, proposed by Baltrusaitis *et al*. [1]. OpenFace is the first open source tool capable of facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation. Some other representations, such as raw pixels, Gabor filters or LBP features could also have been considered for this work [52]. However, the work

of [52, 59] in affective computing suggests AUs are a better choice for an intermediate representation of facial data.

Deep Learning for Affect Detection: In recent years, significant advances in deep learning have lead to the development of various affect detection methods based on deep learning. More specifically, researchers have applied deep learning techniques to problem such as continuous emotion detection [52], facial expression analysis [57], facial action unit detection [10] and others. Deep learning methods have used video data, sensor data or multi-modal data [60]. As noted earlier, engagement is a rational response unlike the spontaneity of emotions and, hence, we model context and temporal variations in the input feature representation with AUs. Such problems are often modeled as sequence learning problems.

Sequence learning has a rich history in signal processing, machine learning and video classification with a number of notable techniques such as Hidden Markov Models (HMM), Dynamic Time Warping (DTW) and Conditional Random Field (CRF)-based approaches being common place [30]. Recurrent neural networks initially gained prominence in computer vision and deep learning areas since they take into account both the observation at the current time step and the representation in the previous one. More recently, in order to address the gradient vanishing problem of vanilla-RNNs when dealing with longer sequences, LSTMs [27] and Gated Recurrent Units (GRU) were proposed [26].

Most existing face-based engagement prediction methods make little use of the temporal information of the video sequence. Unlike more ephemeral emotion, which can be elicited from short stimuli, engagement while doing web-based treatment/education is a long-term process, requiring long-term integration of information. Our proposed prediction model includes context as well as a temporal sequence of facial Action Units (AUs). Using AUs for engagement prediction is a relatively nascent area of research. In this paper, we advocate the use of deep-learning techniques adapted for sequential data [65, 10, 24]. In particular we explore using Long Short-Term Memory (LSTM), a specialized form of recurrent neural networks, that are well suited for sequence learning problems and have emerged as a leading method for number of similar tasks such as video classification [65], video-affect recognition [60], as well as non-visual tasks such as speech, handwriting and music analysis [26]. Relative insensitivity to gap length gives an advantage to LSTMs over alternatives such as RNNs, HMMs and other sequence learning methods.

3. EASE dataset

In this section, we present details of the data used for our analysis of engagement prediction and the collection procedure. The dataset we use is called EASE (Engagement Arousal Self-Efficacy) [1].

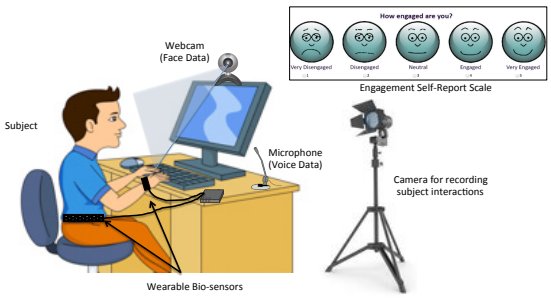

Figure 2. **Experimental setup for data collection:** Subjects interacted with the website while performing self-regulation exercises. Face data was captured using an external webcam; voice data was captured using a microphone. Additional data such as skin-conductance, respiration, and ECG signals were also recorded using wearable sensors. All the interactions were recorded using an external camera. Finally, while the subjects were viewing the trauma-recovery website, the system asks them about their engagement level, with self-report on a scale of 1-5 (top right corner) where 1 is "Very Disengaged" and 5 "Very Engaged"

Data collection procedure: The web-intervention used to collect the data was based on the findings of Social Cognitive Theory [3] and consisted of subjects undergoing six tasks (modules) namely: social-support, self-talk, relaxation, unhelpful coping, professional help and triggers. The broader study was divided into three sessions/visits in the form of a Randomized Control Trial (RCT). Each participant was assigned 2 out of the six modules in each visit. The first two visits were restricted to "Relaxation" and "Triggers" modules only, and in the third visit, the participants were free to choose from the remaining four modules. Each visit lasted for approx. 30 minutes - 1.5 hours. In the first visit, subjects were randomly allocated Relaxation or Triggers as the first module and a reverse order during the second visit and second module. At the beginning of each visit, the subjects listened to a neutral introductory video. During these sessions, a Logitech webcam with a resolution of 640x480 at 30 fps was placed on top of the monitor recording the participants face video along with audio. Physiological data was also recorded for the entire session. The participants could freely interact with the trauma recovery website, and their interactions were recorded in the form of Picture in Picture video using a Camtasia recorder (with screen and webcam recording simultaneously). During the module, participants provided self-reports about their engagement level. The recorded videos are being annotated by psychologists currently for behavioral arousal and engagement.

Although the EASE data is significantly rich in terms of its multi-modal nature, for this work we focus our attention

[1] We have IRB approval to release only unidentifiable data including AUs extracted from facial videos. We cannot display identifiable data, include video frames from the dataset.

primarily on facial video data captured by webcam placed on monitor, engagement self-reports and POMS responses provided by the participants.

Participants and demographics: Subjects were recruited from three health service providers: the TESSA domestic violence center, Peak Vista Community Health Partners, that has 21 centers for seniors, the homeless, a women's health center and six family health centers, student online research system (SONA), and the Veterans Health and Trauma Clinic, a clinic that provides services to the veterans, active duty service members, and their families who are struggling with combat trauma. Subject inclusion criteria were determined by the work of Benight *et al*. [5].

Dataset details: Fig 3 shows the details about the data. As mentioned earlier, each participant came in for three sessions/visits. The first two (controlled) visits are used for experiments in this paper. In each visit, the subjects undergo relaxation and trigger tasks. 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. Since few subjects dropped out during the study, for the first session, we have data from 95 subjects and from the second session we use data from 80 subjects. Some of the collected data was unusable due to either system issues, data corruption or lack of engagement self-reports. The participants provided self-reports about their engagement level with the task on a scale of 1 to 5, where 1 is "Very Disengaged" and 5 "Very Engaged" (see Fig 2). Fig 3 shows the total number of video frames available for each session and each module. Each video of the subject consisted of three self-reports (at the start, in middle and at the end of the segment).

Profile of Mood States (POMS): The Profiles of Mood States-Short Form (POMS-SF) [49, 17]) are used to measure reactive changes in the mood of a person. It is a list of 37 questions related to depression, vigor, tension, anger, fatigue, and confusion. Participants rate items for how they feel right now on a 5-point scale, ranging from 1 (not at all) to 5 (extremely). In the EASE dataset, POMS data was collected from a reduced load using only the first 24 questions. This self-report is given to the subjects at baseline and immediately before and after each module.

Contextual engagement data: The total number of engagement self-reports available for each session and task are shown in Fig 3. Due to a large number of video frames and sparse labels, we consider 30-second segments before each self-report for learning. For RX, we have 372 engagement self-reports leading to 372 segments of 900 samples (30 seconds@ 30 fps), totaling to 334800 frames of data [2]. With 20 AUs per frame and 900 frames per engagement

[2]Some segments shown in data in Fig 3 are not used due to synchronization issues.

	SESSION 1							
	MODULE 1				MODULE 2			
	Task	Number of Videos	Number of Frames	Number of Self Reports	Task	Number of Videos	Number of Frames	Number of Self-Reports
Trigger Task followed by Relaxation Task	Trigger	52	806855	166	Relaxation	52	1579927	122
Relaxation Task followed by Trigger Task	Relaxation	43	1391803	91	Trigger	43	590953	98

	SESSION 2							
	MODULE 1				MODULE 2			
	Task	Number of Videos	Number of Frames	Number of Self Reports	Task	Number of Videos	Number of Frames	Number of Self Reports
Trigger Task followed by Relaxation Task	Trigger	33	454510	94	Relaxation	33	1553409	63
Relaxation Task followed by Trigger Task	Relaxation	47	1139996	105	Trigger	47	544298	149

Figure 3. This figure displays information about participants and the distribution of modules taken by them in each session considered for engagement analysis in this work. 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.

sample, we have an 18000-dimensional feature vector. We use 334 segments for training and 38 segments for testing. Similarly, for TR we have 485 segments of 900 samples (30 seconds@30fps) totaling to 436500 frames of data. We use 437 segments for training and 48 segments for testing. The training/testing labels are a discrete set of engagement levels 1 (very disengaged) through 5 (very engaged). By doing this, each engagement level is treated as a separate and mutually exclusive output.

4. LSTM for engagement prediction

Due to the advantages of Long Short-Term Memory (LSTM)s to preserve information over time and the ability to handle longer sequences, for this work we use LSTMs to model long-term dependencies of AUs for engagement prediction. We model the problem of engagement prediction as a sequence learning problem where input consists of sequence x_i of AUs computed from facial video data of a particular length. Each sequence is associated with a label y_i which relates to the engagement self-reports provided by trauma subjects. Our implementation is based on TensorFlow which is turn is based on [23, 66], and we follow their notation.

We let subscripts denote timesteps and superscripts denote layers. All our states are n-dimensional equal to the number of AUs tracked, currently 20. Let $h_t^l \in \mathbb{R}^n$ be a hidden state in layer l at time-step t. Let $T_{n,m} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be an affine transform from n to m dimensions, i.e. $T_{n,m}x = Wx + b$ for some W and b). Let $\odot$ be element-wise multiplication and let h_t^0 be an input data vector at time-step t. We use the activations h_t^L to predict y_t, since L is the number of layers in our deep LSTM.

The LSTM has complicated dynamics that allow it to easily "memorize" information for an extended number of time-steps using *memory cells* $c_t^l \in \mathbb{R}^n$. Although many

LSTM architectures that differ in their connectivity structure and activation functions, all LSTM architectures have explicit memory cells for storing information for long periods of time, with weights for how to updated the memory cell, retrieve it, or keep it for the next time step. The LSTM architecture used in our experiments is given by the following equations [25], as implemented in TensorFlow basic cell LSTM:

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{2n,4n} \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^{l} \end{pmatrix}$$
$$c_t^l = f \odot c_{t-1}^l + i \odot g$$
$$h_t^l = o \odot \tanh(c_t^l)$$

where sigm is the sigmoid function, sigm and $\tanh$ are applied element-wise, i, f, o, c, h are the input gate, forget gate, output gate, cell activation vector and hidden vectors, respectively. In this work, we assume the length of the sequence is known apriori and hence use LSTMs with static RNN cells. In particular, as described in section 3 above, we use 900 temporal samples of 20 AUs, so $n = 20$, and we used Tensorflow's basicLSTMCell that comprises of 900 units in sequence $t = 1..900$. We optimize the LSTMs to predict the discrete set of engagement levels by computing the cross-entropy of the result after applying softmax function. Each engagement level is treated as a separate and mutually exclusive output.

5. Experiments

We now describe the AU computation procedure to extract intermediate feature representation, followed by the methodology for sequence learning using LSTMs. Finally, we discuss in detail the results obtained for engagement prediction across a variety of tasks and its task specificity.

Facial action units extraction: As noted in earlier section 3, the collected dataset consisted of a large number of facial video and engagement self-reports. While there are number of software available for extracting facial landmark points and facial action units (e.g. [14, 29]) we use the recent work on OpenFace [1] proposed by Baltrušaitis *et al.* It is an open-source tool which has shown state-of-the-art performance on multiple tasks such as head-pose, eye-gaze, and AU detection. For our work, we primarily focus on facial action units. The AUs extracted consisted of both intensity-based and presence-based AUs. Presence based AUs had a value of 1 if AU is visible in the face and 0 otherwise, intensity based AUs ranged from 0 through 5 (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 (intensity), Upper Lid Raiser, Cheek Raiser, Nose Wrinkler, Upper Lip Raiser, Lip Corner Puller (intensity), Dimpler, Lip Corner Depressor (intensity), Chin Raiser, Lip Stretched, Lips Part, Jaw Drop, Brow Lowerer (presence), Lip Corner Puller (presence), Lip Corner Depressor (presence), Lip Tightner, Lip Suck, Blink.

	RX	TR
Chance	32.8%	38.3%
SVC	31.4 ± 7%	35.2 ± 9%
LSTM	39.1 ± 8.8%	50.7 ± 11%

Table 1. The table above shows the enhanced engagement prediction accuracy by using LSTMs over a Support Vector Classifier, highlighting the importance of a deep-learning model for this task.

SVC model: As a baseline for engagement prediction, we Support Vector Classifier (SVC). Linear SVCs were trained using the 18000-dimensional feature vector (20 AUs for each of 900-time samples) for the hundreds of training segments for RX and TR. For training/testing, we used 10-fold cross-validation, and Table 1 reports average engagement prediction accuracy on 10-folds of test data. The table also shows "chance," but since the five engagement levels do not have uniform probability, we consider random chance the algorithm that guesses based on per context per engagement level priors. Hence chance is different for RX and TR. The SVC performance is not significantly different from chance, probably because the number of training samples (334/437) is much smaller than the dimensionality (18000).

LSTM model and tuning: We train the LSTM using the same data, by minimizing the cross-entropy loss of the predicted and actual class after applying softmax function. During the training process, we use Adam optimizer with a learning rate of 0.1. Since Adam optimizer uses larger effective step size, and our data has relatively sparse labels, we found that Adam optimizer performs better compared to gradient descent optimizer. We fix training at 15 epochs, which is after validation error stabilized. For the first test, we use the same 10-fold modeling. As seen in Table 1 the improvement in accuracy of LSTM over SVM supports the need of LSTMs for engagement prediction. Henceforth, for all further experiments and analysis in this paper, we use LSTM models only.

Contextual engagement prediction: We use AU data from subjects performing "Trigger" task to create Trigger (TR) model, and similarly, AU data from subjects performing "Relaxation" task is used to create Relaxation (RX) model. Once the model is trained, to study the effect of context, we test these models using context-specific (testing with the same task) and cross-context technique (testing with the opposite task).

We train context specific LSTMs i.e. we train AU data from subjects performing "Trigger" task to create Trigger (TR) model and similarly AU data from subjects performing "Relaxation" task is used to create Relaxation (RX) model. Table 2 shows the contextual engagement predic-

	RX - Test	TR - Test
RX - Train	39.1 $\pm$ 8.8 %	38.1 $\pm$ 4.4%
TR - Train	36.7 $\pm$ 3.3%	50.7 $\pm$ 11%
(RX + TR) - Train	39.5 $\pm$ 6.1%	49.1 $\pm$ 7.6%

Table 2. The first two rows above show contextual and cross-contextual engagement prediction results on EASE data obtained using LSTM in terms of prediction accuracy. If engagement was not contextual, the models to perform equally well in both cross-contexts, e.g. when the model is trained on Triggers (TR) and tested with Relaxation (RX) data. In the case of TR, context-specific and cross-context models show significant performance differences with the best accuracy being when training and testing are TR. Thus we can reject the hypothesis that context does not matter; it is formally rejected using a two-tailed paired t-test at the .01 level. Furthermore, combined-model (train RX+TR) had the highest amount of data for training and was still outperformed by the contextual models, showing again the importance of contextual modeling.

tion results. We note interesting trends in prediction results. Training on RX and testing on RX yields 39.1% prediction accuracy, and training on RX and testing on TR yields similar accuracy at 38.1%; these are not statistically different (p=.44). When trained on TR data and tested on TR data, results obtained 50.7% prediction accuracy, however, when the same model was tested on RX data, the accuracy dropped to 36.7%. This difference is statically significant (using two-sided heterostatic test p=.006) allowing us to reject the hypothesis that context does not matter. We also find that the two different contexts RX and TR have slightly significant differences (p=.02) in performance. This suggests that engagement models benefit when the context is known and modeled and that tasks differ in difficulty.

Since our experiments demonstrate that engagement prediction models are contextual, we take this a step further and ask the question: *"Does using current mood as context improve engagement prediction for a given task?"*. In order to answer this question, we use POMS data (see sec 3) that was collected before and after the session from each subject. Our POMS questionnaire has 24 questions, which are clustered into five negative sentiments (tension: 5 questions, depression: 6 questions, anger: 5 questions, fatigue: 2 questions, confusion: 2 questions) and one posiive sentiments vigor: 4 questions. The final POMStmd (total mood disturbance) level is computed as difference of sum of negative n(x) and positive p(x) sentiments:

$$\text{POMStmd} = \frac{1}{21.1} \sum_{x \in \text{neg. sentiments}} n(x) - \sum_{x \in \text{pos. sentiment}} p(x)$$

here we scaled the POMStmd scores by the observed value, so that the range is between [0,1].

The POMStmd score is then used to condition each AU input to obtain POMS-aware engagement prediction results. We precondition the basic engagement multi-class LSTM with POMStmd values obtained using self-reports by adding the normalized to the AU representation. Since the engagement scores are ordinal, not categorical, for testing of POMS-aware modeling we use the more traditional Leave-One-Subject-Out (LOSO) methodology, reporting root-mean-squared-error. The performance of POMS-aware engagement predictions are summarized in Table 3. The engagement LSTM baseline was re-calibrated to include subjects with valid POMS self-reports. Even though the LSTM model was optimized for categorical correctness, we notice a significant improvement in performance by augmenting AUs with POMS data. POMS-aware engagement model for TR (POMS-TR) showed a significant reduction in error (p=.0007) at 95% confidence intervals. Due to contextual nature of engagement prediction (as shown earlier), we do not create POMS-aware models for RX+TR (mixed) models.

	RX	TR
LSTM Baseline	0.9989	0.7653
POMS-aware LSTM	0.9493	0.6786

Table 3. The effect of POMS on contextual engagement prediction using Leave-One-Subject-Out (LOSO) validation. Augmenting AUs with POMS data shows clear reduction in RMSE; the difference for TR is statistically significant with a two-sided paired t-test at p=.01.

6. Conclusion and Future Work

In this work, we used data from subjects diagnosed with DSM-V level PTSD, performing self-regulation exercises during the trauma-recovery process. We presented engagement prediction as a sequence learning problem using AUs and LSTM-based deep-learning methods. We showed that user engagement is highly contextual and LSTM-based models can be used to study task specific facial behaviors and predict engagement. Further, we showed that augmenting AUs with data about subject's mood (POMS data) demonstrated a clear improvement in engagement prediction performance.

The primary goal of this work was to explore contextual engagement, rather than create best deep-learning-based classifiers. The work presented in this paper is the first step towards building contextual models for engagement prediction; we expect others will be able to further improve on the models herein. It is impractical to expect uniform sampling across engagement levels from PTSD subjects, so an issue that will need to be addressed in future efforts is to build machine-learning models that are aware of the data imbalances. The most natural way of building better classifiers is training with an even larger dataset and performing parameter optimization of LSTMs. Sophisticated methods like feature (AUs) pooling over space and time, jointly using additional tracking data such as head-pose, gaze, expressions

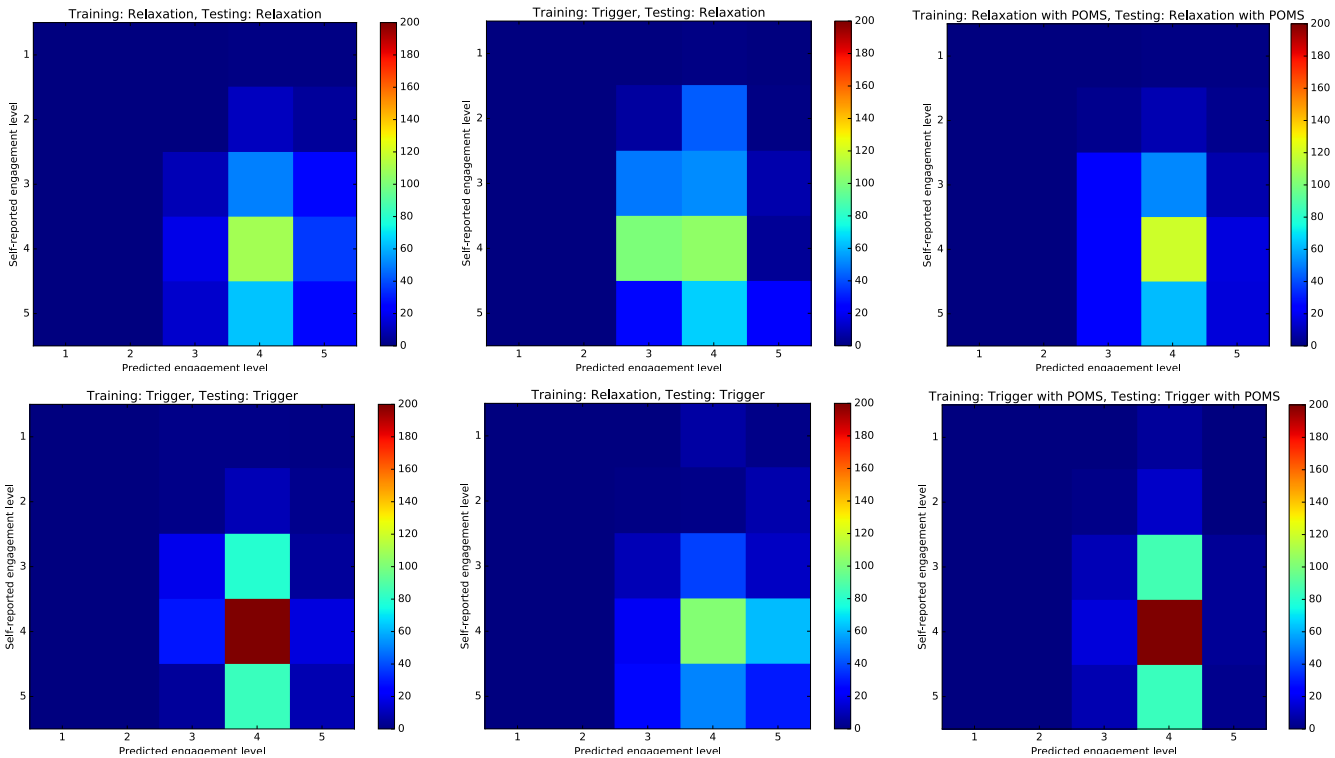

Figure 4. The figure above shows confusion matrices for results presented in Table 2. The top row represents confusion matrices for RX testing and bottom row represents TR testing. The first column is within-context, the second colums is cross-context and right column is within contex with POMS. Notice the cross-context results of Train-TR : Test-RX, with respect to the contextual model of Train-TR : Test-TR shows visible confusion for the "Engaged" class level-4 with the "Neutral" class level-3.

and other emotional states would also likely improve accuracy. The work can be further extended by exploring sophisticated deep learning models, multi-stream LSTM and exploring multiple modalities by taking advantage of recent advances in the field [51, 21].

In our dataset, we collected 15-50 minutes of data from each subject in a given module/session with 3-4 self-reports. In our experiments, for simplicity, we utilized 30-second segments of data before each engagement response. An important direction of research to pursue would be to study the effect of segment length on contextual engagement. Further, if continuous annotations are available, there is huge potential to learn LSTMs directly from annotated behavioral data instead of AUs for engagement prediction.

Finally, we presented first step towards exploring the relationship of subject's mood and his/her engagement level for a given task. This nascent area of research requires further exploration, e.g., can specific elements (anger, fatigue, rigor, etc.) from POMS data be more useful. In EASE dataset, the subjects were presented RX task followed by TR (or vice-versa). We did not consider the importance of presentation order. In actuality, subjects do not change their mood as an on/off switch; module order and interaction should be explored for web-based trauma recovery.

One of the potential applications of engagement prediction is in e-health through interned-based weight loss programs [32], web-based smoking cessation programs [54], PTSD recovery etc. Although these applications are focused on increasing user engagement, engagement prediction techniques using non-verbal cues have never been applied to them. Owing to the simplicity of Web-based interventions and availability of facial expressions coding software, we have stepped outside the realm of student learning for engagement prediction by considering a web-based intervention technique targeted at trauma recovery to understand contextual engagement and towards task-specific engagement.

7. Acknowledgement

This work supported in part by NSF Research Grant SCH-INT 1418520 "Learning and Sensory-based Modeling for Adaptive Web-Empowerment Trauma Treatment". We would like to thank Dr. Charles C. Benight and the UCCS Trauma Health and Hazards Center for the trauma-recovery website and his team for collecting the EASE dataset, Dr. Kotaro Shoji and Ms. Carolyn Yeager for organizing and sharing the self-reported POMS data.

References

[1] 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. 3, 6

[2] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015. 3

[3] C. Benight and A. Bandura. Social cognitive theory of posttraumatic recovery: the role of perceived self-efficacy. *Behaviour Research and Therapy,Elsevier*, 2004. 4

[4] C. Benight, J. Ruzek, and E. Waldrep. Internet interventions for traumatic stress: A review and theoretically based example. *J. of Trauma Stress*, 21(6):513–520, 2008. 1

[5] C. Benight, K. Shoji, CarolynYeager, 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. 3, 5

[6] E. Bunge, B. Dickter, M. Jones, G. Alie, A. Spear, and R. Perales. Behavioral intervention technologies and psychotherapy with youth: A review of the literature. *Current Psychiatry Reviews*, 12(1):14–28, 2016. 3

[7] R. A. Calvo, K. Dinakar, R. Picard, and P. Maes. Computing in mental health. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3438–3445. ACM, 2016. 1

[8] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan. Detecting engagement in hri: An exploration of social and task-based context. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 421–428. IEEE, 2012. 3

[9] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 119–126. ACM, 2009. 3

[10] W. Chu, F. D. la Torre, and J. Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. *Automatic Face and Gesture Conference*, 2017. 4

[11] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE, 2009. 2

[12] M. Couper, G. Alexander, N. Zhang, R. Little, N. Maddy, M. Nowak, J. McClure, J. Calvi, S. Rolnick, and M. Stopponi. Engagement and retention: measuring breadth and depth of participant use of an online intervention. *Journal of Medical Internet Research*, 12(4), 2010. 1

[13] M. Cox, J. Nuevo-Chiquero, J. Saragih, and S. Lucey. Csiro face analysis sdk. *Brisbane, Australia*, 2013. 3

[14] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn. Intraface. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015. 3, 6

[15] S. D'Mello, S. Craig, J. Sullins, and A. Graesser. Predicting affective states expressed through an emote-aloud procedure from autotutors mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education*, 16(1):2–28, 2006. 2

[16] L. Donkin and N. Glozier. Motivators and motivations to persist with online psychological interventions: A qualitative study of treatment completers. *Journal of Medical Internet Research*, 14(3), 2012. 1

[17] R. R. Edwards and J. Haythornthwaite. Mood swings: variability in the use of the profile of mood states. *Journal of pain and symptom management*, 28(6):534, 2004. 5

[18] G. Eysenbach. The law of attrition. *Journal of Medical Internet Research*, 7(1), 2005. 1

[19] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[20] A. C. Gallagher and T. Chen. Estimating age, gender, and identity using first name priors. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 3

[21] S. Ghosh, O. Vinyals, B. Strope, S. Roy, T. Dean, and L. Heck. Contextual lstm (clstm) models for large scale nlp tasks. *KDD Workshop on Deep Learning*, 2016. 8

[22] M. N. Giannakos, L. Jaccheri, and J. Krogstie. How video usage styles affect student engagement? implications for video-based learning environments. In *State-of-the-Art and Future Directions of Smart Learning*, pages 157–163. Springer, 2016. 3

[23] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 5

[24] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. *ICASSP*, 2013. 4

[25] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013. 6

[26] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 2016. 4

[27] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[28] T. R. Insel. Assessing the economic costs of serious mental illness, 2008. 1

[29] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015. 3, 6

[30] L. A. Jeni, A. Lőrincz, Z. Szabó, J. F. Cohn, and T. Kanade. Spatio-temporal event classification using time-series kernel based structured sparsity. In *European Conference on Computer Vision*, pages 135–150. Springer, 2014. 4

[31] F. D. la Torre and J. Cohn. Facial expression analysis. *Visual Analysis of Humans - Springer*, pages 377–409, 2011. 3

[32] R. C. Lefebvre, Y. Tada, S. W. Hilfiker, and C. Baur. The assessment of user engagement with ehealth content: The ehealth engagement scale. *Journal of Computer-Mediated Communication*, 15(4):666–681, 2010. 8

[33] H. Li, J. Brandt, Z. Lin, X. Shen, and G. Hua. A multi-level contextual model for person recognition in photo albums. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1305, 2016. 3

[34] L. Maat and M. Pantic. Gaze-x: Adaptive, affective, multimodal interface for single-user office scenarios. In *Artifical Intelligence for Human Computing*, pages 251–271. Springer, 2007. 3

[35] D. Macea, K. Gajos, Y. D. Calil, and F. Fregni. The efficacy of web-based cognitive behavioral interventions for chronic pain: a systematic review and meta-analysis. *J. of Pain*, 11(10):917–929, 2010. 1, 3

[36] S. U. Marks and R. Gersten. Engagement and disengagement between special and general educators: An application of miles and huberman's cross-case analysis. *Learning Disability Quarterly*, 21(1):32–56, 1998. 1

[37] D. McDuff. New methods for measuring advertising efficacy. *Digital Advertising: Theory and Research*, 2017. 3

[38] D. McDuff, R. El Kaliouby, J. F. Cohn, and R. W. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6(3):223–235, 2015. 3

[39] D. McDuff, R. El Kaliouby, and R. W. Picard. Crowdsourcing facial responses to online videos. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 512–518. IEEE, 2015. 3

[40] V. S. Mehta, M. Parakh, and D. Ghosh. Web based interventions in psychiatry: An overview. *International Journal of Mental Health & Psychiatry*, 2015, 2016. 3

[41] H. Monkaresi, P. 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. 1, 3

[42] W. Mou, H. Gunes, and I. Patras. Automatic recognition of emotions and membership in group videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016. 3

[43] H. O'Brien and E. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2010. 3

[44] O. Ramana Murthy and R. Goecke. Ordered trajectories for large scale human action recognition. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2013. 3

[45] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access*, 5:705–721, 2017. 3

[46] W. J. Scheirer, N. Kumar, K. Ricanek, P. N. Belhumeur, and T. E. Boult. Fusing with context: a bayesian approach to combining descriptive attributes. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8. IEEE, 2011. 3

[47] S. Scherer, G. Lucas, J. Gratch, A. Rizzo, and L. P. Morency. Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interview. *IEEE Trans. on Affective Computing*, 2016. 3

[48] S. M. Schueller, R. F. Muñoz, and D. C. Mohr. Realizing the potential of behavioral intervention technologies. *Current Directions in Psychological Science*, 22(6):478–483, 2013. 3

[49] S. Shacham. A shortened version of the profile of mood states. *Journal of Personality Assesment*, 47(3):305–306, 1983. 2, 5

[50] K. Shoji, C. Benight, A. Mullings, CarolynYeager, S. Dhamija, and T. Boult. Measuring engagement into the web-intervention by the quality of voice. In *International Society for Research on Internet Interventions*. ISRII, 2016. 3

[51] B. Singh, T. Marks, M. Jones, O. Tuzel, and M. Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. *CVPR*, 2016. 8

[52] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic. Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, 2016. 3, 4

[53] S. Steinmetz, C. Benight, S. Bishop, and L. James. My disaster recovery: a pilot randomized controlled trial of an internet intervention. *Anxiety Stress Coping*, 25(5):593–600, 2012. 1

[54] V. Strecher, J. McClure, G. Alexander, B. Chakraborty, V. Nair, J. Konkel, S. Greene, M. Couper, C. Carlier, C. Wiese, et al. The role of engagement in a tailored web-based smoking cessation program: randomized controlled trial. *Journal of medical Internet research*, 10(5):e36, 2008. 3, 8

[55] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev. Improving image classification with location context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1008–1016, 2015. 3

[56] T. Teixeira, M. Wedel, and R. Pieters. Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research*, 49(2):144–159, 2012. 3

[57] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, and M. Pantic. Deep structured learning for facial expression intensity estimation. *CVPR*, 2017. 4

[58] X. Wang and Q. Ji. A hierarchical context model for event recognition in surveillance video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3

[59] 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. 1, 2, 3, 4

[60] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, 2013. 4

[61] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal. Harnessing object and scene semantics for large-scale video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3112–3121, 2016. 3

[62] H. Yang and H. Zhang. Efficient 3d room shape recovery from a single panorama. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5422–5430, 2016. 3

[63] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3294–3301, 2014. 3

[64] C. Yeager. Understanding engagement with a trauma recovery web intervention using the health action process approach framework. *PhD Thesis*, 2016. 3

[65] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 4

[66] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014. 5

[67] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[68] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015. 3