ABSTRACT

Post-traumatic stress disorder (PTSD) is an anxiety disorder that affects a large population and that is currently diagnosed mostly through subject interviews and manual analysis of self-reported symptoms and of subject behavior. However, most PTSD cases are believed to go underdiagnosed and undertreated. We present a multi-modal system for computer-aided diagnosis of PTSD and stress that requires no clinician interview and relies principally in the elicitation of multi-modal neurophysiological responses to audio-visual stimuli. We conduct a thorough evaluation of the discriminative power of the modalities involved (electro encephalography, galvanic skin-response, electrocardiography, head motion and speech), type of stimuli presented (audio, images, audio-and-images and video), and emotions evoked (positive, negative, and trauma-specific) between PTSD subjects and high and low-stress control groups. Our analysis indicates that the multi-modal prediction from the elicitation of trauma-specific emotions from images and audio is a promising approach to computer-aided diagnosis.

Index Terms— computer aided diagnosis, multi-modal fusion, EEG

1. INTRODUCTION

Post-traumatic stress disorder (PTSD) is a complex anxiety disorder [1] caused by traumatic experience. Approximately 50% of the population are estimated to experience serious traumatic exposure during their lifetimes. The estimated risk of developing PTSD after such exposure is around 14% in the general population, 24% in the young urban population [2], and 10-30% in the combat veteran population (e.g. 20% for veterans returning from Iraq and Afghanistan) [3]. It is estimated that approximately 8% of the US population suffers from PTSD symptoms at some point during their lifetime [4] and PTSD prevalence is highest in combat veteran populations, ranging from 10% to 30% depending on study and conflict [5, 6].

The clinical diagnosis of PTSD is based on DSM-IV diagnostic criteria [1] that score multiple behavior dimensions. In particular, symptoms of interest last at least a couple of weeks and include: trauma re-experiencing sequences triggered by trauma reminders, avoidance of trauma related thoughts and feelings and hyperarousal. Since the behaviors over long time intervals cannot be directly observed by an expert, the diagnosis is based on self-reporting information provided by the subject. Therefore, the diagnosis depends on subject motivation, an opportunity to have an interview with a trained professional, and on the accuracy of the self-assessment, factors that can be affected by the diagnosis-related stigma [7].

In this paper we present three main contributions:

• We present a novel kiosk and protocol for the multi-modal computer-aided diagnosis of PTSD and high-scoring Holmes and Rahe stress [8] - i.e. individuals who experienced Major Life Stress (MLS). Our protocol combines the structured presentation of audio-visual stimuli in a controlled environment, with open-ended questions designed to elicit the self-reporting of traumatic and stressful experiences. The kiosk collects data from multiple modalities: neuro-physiological (electro-encephalogram (EEG), electrocardiogram (EGC) and galvanic-skin response (GSR)) and audio-visual signals.

• We demonstrate that the systematic combination of multiple modalities monotonically increases prediction and diagnosis performance of PTSD and MLS. To the best of our knowledge, this is the first study to address such categorization using combined neuro-physiological, acoustic, and lexical data.

• We identify the most informative stimuli and modality for PTSD and MLS prediction from the analysis of data recorded on individuals screened by trained specialists and psychologists.

From an application and protocol design perspective, our work has connections to dialogue-based systems for assessment [9] and treatment of PTSD [10]. The main difference is that our protocol relies on a predefined sequence of trauma-related and generic audio-visual stimuli, and open-ended self-report questions to rule out potential inconsistencies present in open-ended dialogues. While there is existing research that addresses PTSD assessment using different modalities, such as heart rate [11], heart rate and GSR [12], EEG [13], speech [14] and voice quality [15], the emphasis of our work is the analysis of modality combinations and the identification of the most informative stimuli and modalities. Finally, our
work belongs to a set of efforts that analyze statistical properties of neurophysiological response measures to various stimuli (images, audio-visual, cognitive tasks or during selected cognitive tasks [16]) for different subject groups (depression, PTSD, anxious, healthy [13, 17] etc.). We go beyond statistical analysis of responses for different subject groups (healthy vs. PTSD positive) and present results on prediction of PTSD, high and low stress from subjects’ responses.

The remainder of the paper is organized as follows. In Section 2 we describe the response elicitation scenario and the collected dataset. In Section 3 we describe extracted feature sets. In Section 4 we present the PTSD classification results and performance analysis for different modalities, feature sets and stimuli. We close with conclusions and future work directions in Section 5.

2. KIOSK AND PROTOCOL

An overview of our kiosk in and response elicitation protocol are presented in Fig. 1. In order to elicit responses informative for PTSD and MLS diagnosis we designed a set of response elicitation scenarios. The scenario consists of five segments. In the first segment participants are presented with two standard self-report questionnaires, Holmes-Rahe major life stressor scale (MLS) [8] and Clinician Administered PTSD Scale (CAPS) [18], followed by two open-ended questions prompting user to talk about the traumatic experience and its effect on different aspects of daily life. In segments two, three, and four, participants are respectively presented with blocks of, positive, negative and neutral images, selected from the International Affective Pictures System (IAPS) [19]. Each image is displayed for 5 seconds, and followed-up with a debriefing screen on which subject self-report perceived distress level of the stimuli and briefly talk about thoughts triggered by the stimuli. Finally, in segment five, subject are presented with blocks of stimuli, containing images (5), audio clips (5), combination of images and audio clips (5) and video clips (5) related to the trauma category selected in the first segment. The total duration of the response elicitation scenario varies between 30 and 60 minutes depending on subject.

The response elicitation scenarios are delivered to subjects in a kiosk via response elicitation tool that synchronously presents stimuli and collects data from multiple sensors during participants interaction with the tool. The tool synchronously records 20-channel EEG, ECG and head motion signals using Advanced Brain Monitoring X-24 headband [20], GSR using Affectivas Q-sensor [21], speech using close talk-microphone and the high-definition frontal face video (Fig. 1).

3. FEATURE EXTRACTION

In this section we describe features we extracted from different modalities. In order to compare the discriminative potential of different modalities and stimuli types, we extract a set of neurophysiological features with record of successful application on related tasks and augment it with features from the non-intrusive audio and video modalities.

For EEG, EKG, GSR and head motion signals we extracted an exhaustive set of features in the time and spectral domain on different time intervals (Fig. 2). In particular, we extracted features on:

- Segment-level: Five intervals including the full duration of each scenario segment;
- Stimuli-level: Answers to the two open-ended self-reporting questions in the first segment, 15 intervals corresponding to stimuli presentations in segments two, three, and four, 20 intervals corresponding to stimuli presentations in segment five.

Motivated by successful application in recognition of af-
fective states from EEG [22], we extracted additional sub-
stimuli level EEG features on 2-second sub-intervals with 1-
second overlap within each interval. These sub-stimuli fea-
tures include spectral features (power spectral densities and
filter bank power coefficients). As in [23], we also compute
measures of “brain state” that capture alertness and workload.

We extracted two types of features from the raw ECG,
GSR, motion and sub-stimuli EEG feature time series on in-
tervals of interest: signal statistics and signal entropy (Fig. 2).
Signal statistical features such as the signal’s minimum, max-
umum, range, mean, standard deviation, skewness, kurtosis,
zero crossings, fuse signal samples within each interval. This
is standard practice for transforming variable length signals
into fixed-length feature vectors [24]. Additionally, we com-
pute signal entropy features that measure signal complexity
and include Hjorth mobility parameters, approximate entropy,
sample entropy, and SVD entropy [25, 26].

In the spectral domain, we extracted an additional set of
EEG features on full intervals: the power spectral density
and derived multiple features from it, filter bank power co-
eficients for the bands theta (4-8Hz), alpha-low (8-10Hz),
alpha-high (10-12Hz), beta (12-30), and gamma(30-40Hz),
spectral edge, total power and bandwidth, alpha peak, in-
tensity weighted frequency and bandwidth and spectral en-
tropy [27, 26].

We extracted acoustic and ASR-based distress-related
features on intervals that correspond to spoken responses to
the open-ended questions in the self-report segment. These
interval-level acoustic features were obtained as statistical
functionals (max, min, range, higher order moments) of the
frame-level descriptors (pitch, intensity, formants, voice qual-
ity related jitter and shimmer, MFCCs, delta and acceleration
MFCCs) extracted on 25ms processing frames with 10ms
frame shift. In order to mitigate effects of variability in
speaker characteristics and recording conditions we perform
cepstral mean normalization of MFCCs for each participant.

In prior work [28], we designed a coding scheme based on
PTSD diagnostic criteria, trained and evaluated classifiers that
discover 70 PTSD codes (combat exposure, sleep problems,
affective states, etc.) from text. We leverage this work by
running 70 PTSD code classifiers trained on PTSD forum text
data [28] on the ASR outputs. The obtained 70-dimensional
binary vectors form an intermediate representation directly
related to our classification task and we appended them to
the unigram features derived from the ASR output to create
an ASR feature set (Fig. 3(b)).

4. EXPERIMENTAL RESULTS

We evaluate the performance of our kiosk on a group of
30 individuals belonging to three cohorts, PTSD, MLS and
Healthy, with the last two cohorts corresponding to individu-
als who scored high (MLS) and low (Healthy) respectively
in the Holmes-Rahe stress instrument [8]. For the identifica-
tion of PTSD individuals, a trained psychologist interviewed
and diagnosed each of the candidates. All individuals were
sampled from the general population according to a multi-
step screening process that included the evaluation of several
instruments such as the Beck Depression Inventory (BDI),
the Profile of Mood States (POMS) the State-Trait Anxiety
Inventory (STAI), the Center for Epidemiologic Studies De-
pression Scale (CES-D), and the NEO Personality Inventory.
The selection excluded individuals with medical conditions
that made them ineligible to withstand stressful stimuli or
perform demanding neuro-cognitive tasks.

Prior to data acquisition, all subjects completed a se-
quence of neurocognitive tasks designed to assess their levels
of memory, attention, mental workload and learning [23].
Once all data was collected, we investigated the following
problems: (1) Does it help to combine modalities for di-
gnosis? (2) What modality is most informative for PTSD
and MLS diagnosis? and (3) What types of stimuli are most
informative?

We research these questions in the context of multi-label
classification performance using support vector machines
(SVMs) with a grid search on kernel and parameter space
using an early fusion scheme (concatenation of features)
followed by Principal Component Analysis (PCA). An im-
portant general challenge in neuro-clinical pattern recognition
research is that datasets consist of a low number of observa-
tions [29] and are often high-dimensional. These conditions
can lead to optimistically biased estimates of generalization
performance. To address this risk, we propose two solutions.
First, rather than reporting the best performing model using
standard K-fold cross validation, we resort to repeated boot-
strapping (100 samples) and compute the full distribution
of out-of-bag results across all bootstrap samples. Second,
we evaluate every testing condition against two control ex-
periments, one using using stratified random classification,
and a second one using optimal classifiers fitted to features
randomly generated of similar dimensionality.

For scoring each solution, we use the Area Under the
Curve of the Receiver Operator Characteristic (AUC-ROC)
as the scoring selection method to evaluate our kiosk in every
testing condition. Since AUC-ROC is formally defined for
binary classification, we compute the macro average of AUC
across all labels mAUC as:

\[
mAUC = E_L \left[ \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}[p_i > p_j] \right],
\]

where \(E_L\) denotes the expectation under the empirical out-
of-bag label distribution \(L\), \(m\) and \(n\) are the number of true
positives and negatives respectively for a given label, \(\mathbb{1}\) is the
indicator function, and \(p_i\) is the classifier score on instance \(i\).

**Which modality is most informative? Does combining
modalities help?** We evaluate all possible combinations of
modalities in our kiosk; EEG, GSR, EKG, Speech and Head
motion, by running independent early fusion classification experiments on the powerset of modalities, for a total of 64 evaluations. We distinguish between the information added by speech content vs prosody/acoustic aspects of the speech, and show them as different modalities (ASR and Acoustic). To evaluate the value added by each modality, we study the average improvement in performance that each modality adds when combined with other modalities. To assess the value of fusing modalities, we group the results of the powerset experiments by the cardinality of the modalities involved in each run (one to six modalities per run). The results are conclusive. As we show in Fig. 3(a), adding modalities helps with prediction, monotonically, suggesting that modalities complement each other, even though most modalities provide similar performance when tested in isolation as shown in Fig. 3(b), with the sole exception of EEG which is particularly informative. We note that this is not a trivial result, since every modality adds hundreds of dimensions to a feature vector that is already high-dimensional in relation to the number of instances.

Which type of stimuli is most informative? As discussed in Section 1, our stimuli vary in the emotional response they try to elicit (positive, negative, neutral and trauma-specific time segments), and the way they are delivered (images, audio, images and audio and video). We study the value of each stimulus according to both criteria. In Fig. 3(b) we show the distribution of scores for each time segment. The stronger the emotional response elicited, the more discriminative the features become, with the traumatic segment winning over all other stimuli. In Fig. 3(d) we report the scores for the type of stimulus, averaging across all repetitions (5) of the traumatic stimuli. The results show that the multi-modal combination of features from images and audio show the best performance across all our experiments.

5. CONCLUSIONS AND FUTURE WORK

We present, to the best of our knowledge, the first multi-modal kiosk and protocol for PTSD and major life stress prediction. Our results show that (1) eliciting responses with trauma-specific stimuli made of still images and audio is particularly discriminative, and that (2) combining modalities results in a systematic monotonic improvement in performance, despite the fact that the performance of each modality in isolation is relatively weak. We note that building classifiers at the individual stimulus level, as in the experiments in Figs. 3(c) and 3(d) yields stronger performance on average than when working with features computed from the full data collection session as in Figs 3(a) and 3(b). We speculate that grouping time-series from different stimuli in long sessions mixes and dilutes information that could be captured by our features.

The kiosk technicians noted during our experiments that subjects with severe PTSD found it hard to complete the trauma-specific segments, even though they yielded the most discriminative results. Future work includes the design, and optimization of elicitation kiosks that avoid the presentation of traumatic stimuli without sacrificing performance.
6. REFERENCES


