Automatic Behavior Descriptors for Psychological Disorder Analysis

Stefan Scherer, Giota Stratou, Marwa Mahmoud, Jill Boberg,
Jonathan Gratch, Albert (Skip) Rizzo, Louis-Philippe Morency

***Abstract*— We investigate the capabilities of automatic nonverbal behavior descriptors to identify indicators of psychological disorders such as depression, anxiety, and post-traumatic stress disorder. We seek to confirm and enrich present state of the art, predominantly based on qualitative manual annotations, with automatic quantitative behavior descriptors. In this paper, we propose four nonverbal behavior descriptors that can be automatically estimated from visual signals. We introduce a new dataset called the Distress Assessment Interview Corpus (DAIC) which includes 167 dyadic interactions between a confederate interviewer and a paid participant. Our evaluation on this dataset shows correlation of our automatic behavior descriptors with specific psychological disorders as well as a generic distress measure. Our analysis also includes a deeper study of self-adaptor and fidgeting behaviors based on detailed annotations of where these behaviors occur.**

I. Introduction

The recent progress in facial feature tracking and articulated body tracking [2], [25], [34] has opened the door to new applications for automatic nonverbal behavior analysis. One promising direction for this technology is the medical domain where computer vision algorithms can assist clinicians and health care providers in their daily activities. For example, these new perceptual software can assist doctors during remote telemedicine sessions that lack the communication cues provided in face-to-face interactions. Automatic behavior descriptors can further add quantitative information to the interactions such as behavior dynamics and intensities. These quantitative data can improve both post-session and online analysis. Proper sensing of nonverbal cues can also provide support for an interactive virtual coach able to offer advice based on perceived indicators of distress or anxiety.

A key challenge when building such nonverbal perception technology is to develop and validate robust descriptors of human behaviors that are correlated with psychological disorders such as depression, anxiety or post-traumatic stress disorder (PTSD). These descriptors should be designed to support the diagnosis or treatment performed by a clinician; no descriptor is completely diagnostic by itself, but they show tendencies in people's behaviors. A promising result in this direction is the recent work of Cohn and colleagues who studied facial expressions and vocal patterns related to depression [27], [9].

The authors are with the University of Southern California Institute for Creative Technologies, Playa Vista, CA, 90094, USA; Marwa Mahmoud is with the University of Cambridge, UK; the corresponding author is Stefan Scherer scherer@ict.usc.edu

This work is supported by DARPA under contract (W911NF-04-D-0005) and U.S. Army Research, Development, and Engineering Command and. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

In this paper, we present and validate automatic behavior descriptors related to depression, anxiety and/or PTSD. We introduce a new dataset, called the Distress Assessment Interview Corpus, which consists of 70+ hours of dyadic interviews designed to study the verbal and nonverbal behaviors correlated with psychological disorders. We describe our approach to automatically assessing indicators of psychological disorders from head pose, eye gaze and facial expressions (smiles). We also present a detailed study of the fidgeting and self-adaptor gestures happening during these interviews.

The following section presents previous work studying the relationship between nonverbal behaviors and psychological disorders. Section III introduces the research goals of this work. In Section IV we describe the procedure for data acquisition, the used psychological measures, as well as the recorded population. Section V presents the multimodal behavior analysis platform MultiSense. The manual annotation scheme is introduced in Section VI, and the observed results of the automatic and manual analysis are presented and discussed in Section VII. Finally, Section VIII concludes the paper and introduces future directions of our work.

II. Related Work

A large body of research has examined the relationship between nonverbal behavior and clinical conditions. Most of this research resides in clinical and social psychology and, until very recently, the vast majority relied on manual annotation of gestures and facial expressions. Despite at least forty years of intensive research, there is still surprisingly little progress on identifying clear relationships between patient disorders and expressed behavior. In part, this is due to the difficulty in manually annotating data, inconsistencies in how both clinical states and expressed behaviors are defined across studies, and the wide range of social contexts in which behavior is elicited and observed. Despite these complexities, there is general consensus on the relationship between some clinical conditions (especially depression and social anxiety) and associated nonverbal cues. These general findings inform our search for automatic nonverbal behavior descriptors, so we first review these key findings. Some nonverbal behaviors associated with psychological disorders are summarized in Table I.

Gaze and mutual attention are critical behaviors for regulating conversations, so it is not surprising that a number of clinical conditions are associated with atypical patterns of gaze. Depressed patients have a tendency to maintain significantly less mutual gaze [33], show nonspecific gaze,

such as staring off into space [29] and avert their gaze, often together with a downward angling of the head [26]. The pattern for depression and PTSD is similar, with patients often avoiding direct eye contact with the clinician.

Emotional expressivity, such as the frequency or duration of smiles, is also diagnostic of clinical state. For example, depressed patients frequently display flattened or negative affect including less emotional expressivity [26], [7], fewer mouth movements [13], [29], more frowns [13], [26] and fewer gestures [15], [26]. Some findings suggest it is not the total quantity of expressions that is important, but their dynamics. For example, depressed patients may frequently smile, but these are perceived as less genuine and often shorter in duration [18] than what is found in non-clinical populations. Social anxiety and PTSD share some of the features of depression also have a tendency for heightened emotional sensitivity and more energetic responses including hypersensitivity to stimuli: e.g., more startle responses, and greater tendency to display anger [18], or shame [24].

Finally, certain gestures are seen with greater frequency in clinical populations. Fidgeting is often reported. This includes gestures such as tapping or rhythmically shaking hands or feet and is seen in both anxiety and depression [13]. Depressed patients also often engage in "self-adaptors" [11], such as rhythmically touching, hugging or stroking parts of the body or self-grooming, such as repeatedly stroking the hair [13].

One recent brewing controversy within the clinical literature is whether the specific categories of mental illness (e.g., depression, PTSD, anxiety, and schizophrenia) reflect discrete and clearly separable conditions or, rather, continuous differences along some more general underlying dimensions [28]. This parallels controversies in emotion research as to whether emotions reflect discrete and neurologically distinct systems in the brain, or if they are simply labels we apply to differences along broad dimensions such as valence and arousal. Indeed, when it comes to emotion recognition, some meta-reviews suggest that dimensional approaches may lead to better recognition rates than automatic recognition techniques based on discrete labels.

The broad dimension receiving the most support in clinical studies is the concept of general distress. For example, [12] examined a large number of clinical diagnostic interviews and found that diagnoses of major depression and PTSD were better characterized by considering only a single dimension of general distress. Several other researchers have statistically re-examined the standard scales and interview protocols used to diagnose depression, anxiety and PTSD and found they highly correlate and are better seen as measuring general distress [4], [23], [1]. For this reason, we will investigate if general distress may be a more appropriate concept for recognizing clinical illness in addition to the more conventional discrete categories.

III. Research Goals

We seek to investigate the following research goals:

Authors	Nonverbal behavior	Disorder
Fairbanks, et al. 1982	↓ mouth movements	depression
	↓ *smiling*	
	↑ *self-grooming*	
	↑ *turning head away*	
	↑ *fidgeting*	anxiety
Hall, et al. 1995	↓ gestures	depression
	↓ speech	
	↑ long pauses	
Kirsch and Brunnhuber 2007	↑ anger	PTSD
	↓ *genuine joy*	
Perez and Riggio 2003	↑ *gaze down*	depression
	↑ *gaze aversion*	
	↓ emotional expressivity	
	↓ gestures	
	↑ frowns	
Schelde 1998	↑ *nonspecific gaze*	depression
	↓ mouth movements	
	↓ interaction	
Waxer 1974	↓ *mutual gaze*	depression

TABLE I

Summary of nonverbal behaviors found in the literature. Nonverbal behaviors written in italics are part of the analysis in the present work.

1) **Automatic gaze descriptors:** As discussed in [29], [33], [26], subjects with psychological disorders show increased averted gaze and nonspecific gazing behavior based on manual annotations. Within our analysis we both seek to confirm these findings with automatic descriptors and investigate quantitatively the dynamics of both the head as well as eye gaze during dyadic conversations. In particular, we study the downward angling of the head and the eye gaze for subjects with psychological disorders.
2) **Automatic smile descriptors:** Additionally, findings in [13] support that a reduced number of smiles can be observed in subjects with psychological disorders. However, this could not be confirmed for the number of smiles and laughter of depressed subjects in [27], but an increased amount of masking was observed. Further, [18] found less genuine smiles in PTSD patients. Again, we seek to further analyze these findings by analyzing smiling behaviors quantitatively and dynamically. In particular, we analyze if a reduced average duration of smiles as well as a reduced intensity of smiles can be observed for subjects with psychological disorders, due to increased amount of masking and a reduced amount of genuine smiles.
3) **Manual self-adaptors annotation:** An additional research goal of this work is to better study the typical regions of self-adaptors (i.e. self-touches) for people with psychological disorders. These were observed in [13] for people with depression and anxiety. Through manual annotations we seek to better understand the type of fidgeting and self-adaptors displayed by people with psychological disorders. We are particularly interested in the behaviors of people with PTSD, as this population was relatively understudied in the past.

IV. Distress Assessment Interview Corpus

In this section we discuss the procedure for data acquisition of the Distress Assessment Interview Corpus (DAIC). We further introduce the employed psychological measures, and the overall size and characteristics of the corpus.

A. Procedure

For the recording of the dataset we adhered to the following procedure: After a short explanation of the study and giving consent, participants were left alone to complete a series of questionnaires at a computer. These included the following: the PTSD Checklist-Civilian version (PCL-C), the Patient Health Questionnaire, depression module (PHQ-9), the Spielberger State-Trait Anxiety Inventory (STAI-T), the Balanced Inventory of Desirable Responding (BIDR), the Big Five Inventory (BFI), the Reading the Mind in the Eyes (RME) scale, and the Positive and Negative Affect Schedule (PANAS). The following section describes the main three questionnaires used in this paper. This process took from 30-60 minutes, depending on the participant.

Upon completion of the questionnaires, the participants were asked to sit down in a chair facing the interviewer directly. Both of them were video recorded with an HD webcam and a depth sensor (i.e. Kinect). The participant and interviewer were about seven feet apart. This distance was required for the Kinect to record depth information for the whole body of the subject/interviewer. This was not a problem for the participants, as only 5% said that it had a large effect on their interaction and only about 9% were uncomfortable or very uncomfortable with the distance.

Lavaliere microphones were attached to the lapel of the subject, and the recording was started. The interviewer then began a series of semi-structured questions. The questions were based partly on answers given by the participant during the questionnaire phase about their diagnosis and symptoms of PTSD or depression. The initial questions were neutral, but became more specific about possible symptoms and traumatic events as the interview progressed and as the participants willingness to talk dictated. Interviews lasted between 30 and 60 minutes.

Finally, the participant was asked to complete the final set of questionnaires, which included a second PANAS, situational motivation questions, and questions about the participant's reactions to the interviewer. This phase took between 10 and 20 minutes. Participants were then debriefed, paid $25 to $35, and escorted out.

B. Measures

Standard clinical screening measures were used to assess PTSD, anxiety, and depression. Further, we introduce and motivate a measure of general distress based on the observed correlation between these three measures.

Post-traumatic stress disorder (PTSD). The PTSD Checklist-Civilian version (PCL-C) [5] is a self-report measure that evaluates all 17 PTSD criteria using a 5-point Likert scale. It is based on the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association, 1994). Scores range from 17-85, and PTSD severity is reflected in the size of the score, with larger scores indicating greater severity. Sensitivity and specificity are reportedly 0.82 and 0.83, respectively for detecting DSM PTSD diagnoses. The PCL-C is scored based on the DSM-IV schema, with symptomatic responses (moderately or above) to at least six items from three categories. The scores are added to assess the severity of symptoms.

State/Trait Anxiety Inventory (STAI). The State/Trait Anxiety Inventory (STAI) [31], [3] is another commonly used self-report questionnaire that can be used in the formulation of a clinical diagnosis, to help differentiate anxiety from depression, for psychological and health research, and for the assessment of clinical anxiety in patients. The STAI is a validated 20-item self-report assessment scale which includes separate measures of transient (state) and enduring (trait) levels of anxiety. Many reliability and validity tests have proven evidence that the STAI is an appropriate and adequate assessment for studying anxiety [30]. Trait Anxiety is assessed by adding up all scores and using the experimental STAI-T population mean of 34.84+SD(9.21) for a total cut-off of 44.

Patient Health Questionnaire-Depression 9 (PHQ-9). The Patient Health Questionnaire-Depression 9 (PHQ-9) is a ten-item self-report measure based directly on the diagnostic criteria for major depressive disorder in the DSM-IV [20]. The PHQ-9 is typically used as a screening tool for assisting clinicians in diagnosing depression as well as selecting and monitoring treatment. Further, it has been shown to be a reliable and valid measure of depression severity [21]. Scores range from 0-27, with higher scores indicating higher depression severity. Due to IRB requirements, we used a 9-question PHQ-9 instrument, leaving off question 9 about suicidal thoughts. When scoring the PHQ-9, response categories 2-3 (More than half the days or above) are treated as symptomatic and responses 0-1 (Several days or below) as non-symptomatic. At least five of the first eight questions must be checked as symptomatic, including at least one of the first two questions. Additionally, Question 10 must be checked as at least somewhat difficult. Severity is calculated by totaling the answers to all of the questions. A PHQ-9 score of at least 10 was used to determine a positive assessment, in addition to the previous requirements.

General distress. We observed significant correlations between the disorders (i.e. PTSD, anxiety, and depression), with a significance level of $p < 0.01$. Diagnosis for depression correlated with PTSD with $\rho = 0.64$, using Pearson's correlation, diagnosis for depression and anxiety correlated with $\rho = 0.40$, and PTSD with anxiety correlated with $\rho = 0.43$.

When directly considering the scalar severity measure of the three inventories, we found even stronger correlations with $\rho > 0.8$, as seen in Figure 1. Based on this analysis, and several findings in the literature that confirmed these co-morbidities [8], [23], we decided to combine the three measures using factor analysis to that of *general distress*. We performed factor analysis on all three metrics and kept

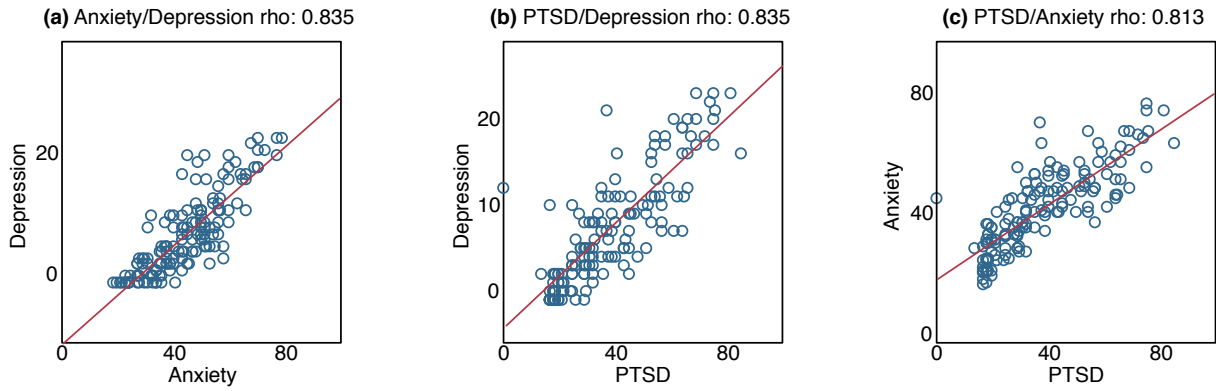

Fig. 1. Scatterplots showing the correlations between the conditions (a) anxiety and depression, (b) depression and PTSD, and (c) anxiety and PTSD. It is clearly seen that strong correlations are found ($\rho > 0.8$) for all combinations. The regression line fit to the data is shown in red.

the first three components. The population was separated into tertiles: the upper third was considered as "distressed", the middle third was "unclear" and was discarded, and the lower third was labeled "not distressed".

C. Participants

The DAIC was recorded on two sites, comprising three conditions. At a US Vets site in California, 57 subjects were interviewed face-to-face. At the USC Institute for Creative Technologies, 110 subjects were interviewed, 54 face-to-face and 56 over a teleconferencing set-up. The sessions all followed the general procedure introduced in Section IV-A.

The population of subjects who were interviewed at the Institute was recruited off of Craigslist. One ad asked for participants who had been previously diagnosed with depression, PTSD, or traumatic brain injury, while another asked for any subjects between the ages of 18 and 65. All subjects who met requirements (age, adequate eyesight) were accepted. Each subject was randomly assigned to either the teleconferencing or face-to-face condition. Some also were connected to a BIOPAC to measure psychophysiological signals.

The population at the US Vets site was recruited from among the resident and visiting population there. The resident population consists entirely of veterans. Some spouses and veterans who had completed one or more programs or were in a non-resident program were among the subjects.

For this paper, only the participants that were assigned to the face-to-face, non-BIOPAC condition were considered, due to possible impact of cables to the behavior. Of those, 54 were those recruited from Craigslist, and 57 were recruited from the US Vets population.

When participants were asked about their history of particular psychological disorders, 59.4% reported depression, and 29.5% PTSD. Following the assessment using the inventories introduced in Section IV-B, 29% were positive for depression, 32% for PTSD, and 62% for anxiety.

V. Automatic Behavior Analysis

In this section we describe the automatic analysis conducted in this paper in more detail. The goals of the automatic behavior analysis utilizing current state of the art behavior descriptors is two-fold: first, we would like to confirm findings from previous work that have identified several nonverbal behaviors that are characteristic of psychological disorders; and second, we seek to enrich previous findings, which have until recently predominantly relied on manual behavior annotations, with the quantitative analysis of the behavior dynamics. In the following we introduce our automatic analysis system *MultiSense* and the automatic behavior descriptors analyzed in the present study.

A. Automatic Analysis System

For the automatic analysis we employ a multimodal sensor fusion framework called *MultiSense*. This is a flexible framework that was based on the Social Signal Interpretation framework (SSI) by [32] and it is created as a platform to integrate and fuse sensor technologies and develop probabilistic models for human behavior recognition. The modular setup of MultiSense allows us to integrate multiple sensing technologies including the following: *CLM-Z FaceTracker* by [2] for facial tracking (66 facial feature points), *GAVAM HeadTracker* by [25] for 3D head position and orientation, OMRON's *OKAO Vision* for the eye gaze signal, smile level, and face pose and skeleton tracking by Microsoft Kinect SDK. It also includes RGB video capture via webcam device, synchronized audio capture and depth image capture via Microsoft Kinect sensor. MultiSense utilizes a multithreading architecture enabling all these different technologies to run in parallel and in realtime. Moreover MultiSense's synchronization schemes allow for inter-module cooperation and information fusion. We employ fusion of the different tracker results to create a multimodal feature set that can be used to infer higher level information on perceived human behavioral state such as attentiveness, emotional state, agitation, and agreement by building probabilistic models for these states.

B. Automatic Behavior Descriptors

Based on our research goals (cf. Section III) and our tracking technology we designed a few key behavioral descriptors that are informative for the psychological disorders, namely general distress, anxiety, depression, and PTSD. According

to literature presented in Section II and summarized in Table I, gaze and head turns are important features to observe (gaze aversion, gaze down and head turning are some of those behaviors associated with these features), as well as overall smile level (amount of smiling, and expression of genuine joy are associated with this feature). We seek to confirm these findings and add quantitative evidence to them by utilizing the automatic behavior description processes described above. To extract the features for our dataset we used the output from MultiSense to estimate the head orientation, the eye-gaze direction, smile level, and smile duration. The following are the behavior descriptors we analyzed in detail:

- **Vertical Head Gaze**: This is a measure of how much the person is facing up or down during the conversation. MultiSense returns the 3D head orientation per video frame in radians [25]. The average head rotation is measured based on the x-axis (i.e. pitch).
- **Vertical Eye Gaze**: This is a measure of the gaze vertical direction of the subject during the conversation. MultiSense returns the vertical gaze direction that can range in the span: [-60,60] degrees. We are measuring the average vertical gaze.
- **Smile Intensity**: This is the average smile level of the subject during the conversation. MultiSense returns the smile level, which can range in the span: [0,100], where 0 is the absence of smile and 100 a strong smile. Since MultiSense returns not only the existence but also the intensity of the smile in every frame, averaging that signal over the whole conversation includes the factors of how frequent, how strong, and how long the subject is smiling.
- **Smile Duration**: This is the average duration of the smiles of the subject during the conversation. It is again extracted using MultiSense. In this case, the smile level signal was thresholded to leave only instances where the smile level is greater than 60. We proceeded with a small window smoothing process to get a binary smile pulse signal that allows us to count the number of strong smiles and approximate the duration of each. Based on the literature [27], these are factors that can help differentiate between genuine and non genuine smiles.

The MultiSense signals that we extracted provide a confidence level for their output. We used the average confidence over the whole session as a screening measure to discard noisy videos. We analyze and discuss the results in terms of our research goals in Section VII.

VI. Manual behavior annotation

As mentioned in Section III, one of the goals of this work is to identify the typical regions of self-adaptors and fidgeting behaviors, which were found to be correlated with psychological disorders as stated in [13]. As there are no automatic behavior descriptors currently available that robustly detect these behaviors, we complement the capabilities of our automatic descriptors with manual annotations. In the future, we plan to develop and train automatic descriptors for those behaviors based on the annotations. Particular interest was directed to the behaviors of people with PTSD, as this population is relatively understudied. The cues that were selected were divided into the following two tiers:

- **Hands self-adaptors:** For this tier self-adaptors were annotated along with hand fidgets. These include hand tapping, stroking, grooming, playing with fingers or the hair, and similar fidgeting behaviors. These self-adaptors were separated into three distinct regions, namely *head*, *torso*, and *hands*. We split the manual annotation into these regions in order to be able to later disambiguate the regions on the body where these self-adaptors predominantly occur. We then compare the average durations of self-adapters to either (Self-adaptors Head) the head, face and hair region, (Self-adaptors Hands) the hands touch, or (Self-adaptors Torso) the arms and torso, in Section VII-C.
- **Legs fidgeting:** Similarly to the hand fidgets, we annotated leg fidgets that include behaviors such as leg shaking and foot tapping. In our evaluation in Section VII-C, we then compare the average length of the subjects tapping or shaking their legs.

In total, four student annotators were recruited to carry out the full annotation. Each pair got one tier assigned to them and went through a training phase. Both sets of annotators showed great agreement between annotations. Self-adaptors resulted after training in a Krippendorff's alpha of $\alpha = 0.77$; for the leg fidgets $\alpha = 0.84$ was observed [19]. These manual annotations, were performed using ELAN [22].

After the training phase, each annotator started to annotate videos separately. To monitor the reliability of the coding in the post-training full annotations phase, every 10-15 videos each pair got assigned the same video to annotate without knowledge that the other teammate was also annotating the same video, and inter-rater agreement was re-checked. Since findings suggest that annotators perform better when they know that their reliability is being assessed [35], [16], annotators were informed that their reliability was measured but did not know which of the videos they worked on were used for cross-checking.

VII. Evaluation and Discussion

In this section we report the results of our investigations. The results are separated into two parts: the automatic behavior analysis using MultiSense and the manual nonverbal behavior annotation. In sections VII-A and VII-B, we report the results of the automatic nonverbal behavior descriptors. We analyze *Vertical Head Gaze*, the overall vertical directionality of the gaze direction, as well as *Vertical Eye Gaze*, the overall vertical directionality of the gaze direction. Further, we compare *Smile Intensity*, the average intensity of smiles as well as *Smile Duration*, the average duration of a smile. Finally, we report some supplementary findings based on the manual annotations in Section VII-C.

	Tier	Condition μ (σ)	No-Condition μ (σ)	p	g
Distress	**VHead Gaze**	0.14 (0.11)	0.19 (0.10)	**0.04**	-0.46
	VEye Gaze	8.93 (7.92)	13.65 (6.30)	**0.01**	-0.64
	Smile Int.	12.31 (10.09)	23.76 (18.30)	**<0.01**	-0.75
	Smile Dur.	2.49 (0.87)	3.43 (1.85)	**0.01**	-0.63
Depression	**VHead Gaze**	0.15 (0.11)	0.17 (0.09)	0.20	-0.19
	VEye Gaze	9.83 (7.35)	11.56 (6.69)	0.16	-0.25
	Smile Int.	12.81 (11.14)	19.94 (16.85)	**0.04**	-0.45
	Smile Dur.	2.59 (0.87)	3.02 (1.69)	0.15	-0.27
Anxiety	**VHead Gaze**	0.15 (0.10)	0.19 (0.10)	0.06	-0.36
	VEye Gaze	10.05 (7.87)	12.87 (4.04)	**0.04**	-0.41
	Smile Int.	14.77 (13.33)	23.52 (18.15)	**<0.01**	-0.56
	Smile Dur.	2.66 (1.25)	3.33 (1.87)	**0.03**	-0.44
PTSD	**VHead Gaze**	0.14 (0.11)	0.18 (0.09)	**0.04**	-0.39
	VEye Gaze	9.37 (8.12)	11.86 (6.11)	0.07	-0.36
	Smile Int.	12.25 (10.78)	20.85 (17.11)	**0.01**	-0.55
	Smile Dur.	2.37 (0.81)	3.17 (1.73)	**0.02**	-0.52

TABLE II

EVALUATION OF THE AUTOMATIC NONVERBAL BEHAVIOR ANALYSIS. SEE SECTION V FOR DETAILS ABOUT DESCRIPTORS. WITH VHEAD GAZE THE VERTICAL HEAD GAZE, VEYE GAZE THE VERTICAL EYE GAZE, SMILE INT. THE SMILE INTENSITY AND SMILE DUR. THE SMILE DURATION.

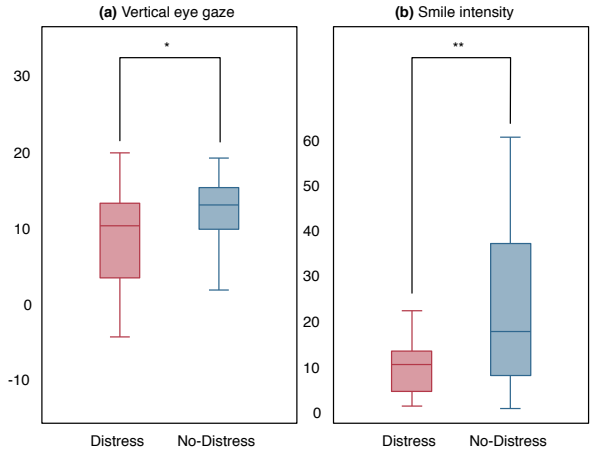

Fig. 2. Example of two automatic behavior descriptors. Boxplots show the significantly stronger overall downward angle of the (a) eye gaze ($p < 0.05$) and a significantly lowered average (b) smile intensity ($p < 0.01$) of subjects in the condition distress vs. no-distress, as measured by MultiSense.

A. Automatic gaze descriptors

The results of the automatic analysis are summarized in Table II. We present the statistics of both the condition and no-condition subjects for all four evaluated groups: distress, depression, anxiety, and PTSD. The column μ denotes the mean, and σ the standard deviation. Additionally, we present the p values of one-tailed t-tests and Hedges' g value as a measure of the effect size found in the data. The tail was chosen according to findings in the literature as stated in Section III. The g value denotes the required shift of the mean of one set to match the mean of the other in magnitude of standard deviations [17].

The vertical gaze measurements provided by MultiSense show significant results for the condition distress vs. no-distress. Based on these measures, distressed subjects tend to gaze downwards more over the whole interview than non-distressed subjects. Head gaze as measured by MultiSense is on average at 0.14 for distressed subjects and 0.19 for non-distressed subjects. The one-tailed t-test shows a significant difference with $p < 0.05$. Similarly, the overall eye gaze as measured by MultiSense shows a downward trend on average with 8.93 for distressed subjects and 13.65 for non-distressed subjects. Again, the one-tailed t-test shows a significant difference with $p < 0.05$. The observations have the same tendency and trend for the three other conditions (depression, anxiety, and PTSD), with strong trends and significant results for anxiety and PTSD. For the condition depression vs. no-depression, no significant results could be found with respect to the overall vertical gaze angle.

These results add to the rich literature corpus on nonverbal behavior indicators of psychological disorders, as the automatic behavior descriptors yield precise measures of general gaze directionality of the face as well as the eyes. An analysis on such a granular level is manually only possible by investing great effort. Further, the robustness of the results was confirmed by two independent gaze trackers that estimated face as well as eye gaze in parallel.

B. Automatic smile descriptors

We utilized MultiSense to measure the average smile intensity ($\in [0, 100]$) and found that over all four conditions subjects exhibit less intense smiles. Table II shows the evaluation results of our two smile descriptors. In particular, distressed subjects smile less intensely than non-distressed subjects (distressed: 12.31 vs. non-distressed: 23.76; $p < 0.01$). The strongest effect for the three remaining conditions is observed for the condition anxiety. Also, the average smile duration was significantly smaller for subjects in the condition distress than for non-distressed subjects (distressed: 2.49 vs. non-distressed: 3.43; $p = 0.01$) as well as in the conditions anxiety and PTSD. Again, only depression shows no significant difference, but a similar trend.

Hence, based on our findings using the automatic behavior descriptors to estimate smile intensity and smile duration, we can confirm that our quantitative analysis of the smiling behavior is indeed correlated with psychological disorders of subjects. In particular, the automatic detection of decreased average intensity of smiles has strong benefits over traditional manual annotation approaches, as the coding of expression intensities can prove to be a very tedious and time consuming procedure.

These findings correspond to those in [7], where significantly attenuated positive emotional reactions were confirmed in a large meta analysis across self-reported, physiological, and behavioral emotional reactivity in major depressive disorders studies. Even though we observed reduced smile intensities and reduced smile durations for subjects with psychological disorders, the nonverbal behavior of smiling might require some further analysis. For example, it is stated in [27] that an increase in masking behaviors of

	Tier	Condition μ (σ)	No-Condition μ (σ)	p	g
Distress	**SelfAd Head**	2.19 (1.92)	1.67 (0.90)	0.17	0.34
	SelfAd Hand	3.99 (2.03)	2.52 (0.92)	**<0.01**	0.93
	SelfAd Torso	2.45 (1.53)	2.17 (1.01)	0.28	0.21
	Leg fidgeting	3.61 (1.76)	2.68 (1.48)	**0.03**	0.56
Depression	**SelfAd Head**	2.17 (1.93)	1.57 (0.79)	0.07	0.47
	SelfAd Hand	3.85 (1.99)	3.02 (1.38)	**0.05**	0.51
	SelfAd Torso	2.46 (1.53)	2.12 (0.84)	0.17	0.31
	Leg fidgeting	3.78 (1.60)	3.06 (2.13)	0.09	0.36
Anxiety	**SelfAd Head**	1.85 (1.48)	1.61 (0.90)	0.27	0.18
	SelfAd Hand	3.71 (1.82)	2.61 (0.96)	**0.01**	0.70
	SelfAd Torso	2.28 (1.16)	2.14 (0.99)	0.33	0.13
	Leg fidgeting	3.52 (2.12)	2.69 (1.52)	0.07	0.41
PTSD	**SelfAd Head**	2.26 (1.82)	1.50 (0.80)	**0.03**	0.59
	SelfAd Hand	3.95 (2.04)	2.94 (1.27)	**0.02**	0.63
	SelfAd Torso	2.51 (1.45)	2.08 (0.86)	0.11	0.39
	Leg fidgeting	3.55 (1.90)	3.11 (2.06)	0.20	0.22

TABLE III

EVALUATION OF THE MANUAL ANNOTATIONS BASED ON THE VARIOUS TIERS DESCRIBED IN SECTION VI. WITH SELFAD DENOTING SELF-ADAPTORS WITH THE CORRESPONDING REGION.

smiles was observed for depressed subjects. These masking behaviors might be of further interest in future analysis. Hence, we plan to annotate such masking behaviors (e.g. AU14 or AU12 of the facial action coding scheme [10]) in a further annotation effort in order to confirm the hypothesis of [27] and to create training examples for the training of future automatic behavior descriptors.

C. Manual annotation evaluation

As introduced in Section VI, we manually annotated the recordings on two tiers self-adaptors and leg fidgeting. Here, we report several results and indicators based on these. The results of the manual annotation are summarized in Table III.

The average durations of hand self-adaptors all follow the same trend towards longer durations observed for subjects with psychological disorders, similarly to that observed in [13]. Self-touches in the head region (i.e. head, hair, and face) are significantly longer in a one tail t-test with $p = 0.03$, with an average duration of 2.26s for subjects with PTSD and 1.50 for subjects with no sign of PTSD. For the other conditions the results follow the same trend. Self-touches of the hands are significantly longer for all four conditions. For example, in the distress condition subjects exhibit longer self-adaptors with 3.99s on average and 2.52s for non-distressed subjects ($p < 0.01$). The results for self-adaptors and fidgets in the hand region are also further visualized in Figure 3 (a) for the conditions distress and no-distress. Significant results are marked with brackets and * for $p < 0.05$ and ** for $p < 0.01$. The self-touches in the region of the torso show no significant differences for all the four categories.

Further, leg fidgets are significantly longer for distressed subjects with 3.61s on average than for non-distressed subjects with 2.68s ($p = 0.03$). Figure 3 (b) visualizes this result. The other three conditions show no significant differences but follow the same trend.

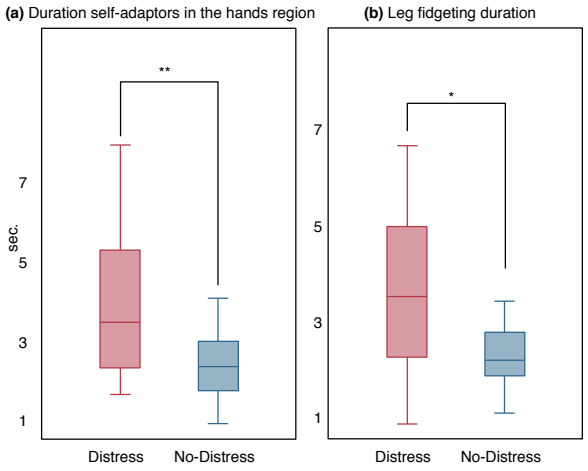

Fig. 3. Example of two manual behavior annotations. Boxplots show significantly increased average duration of self-adaptors and fidgets in the hand region (a) and leg fidgets (b), with $p < 0.05$ in a one-tailed t-test, for the condition distress vs. no-distress.

Due to the granularity of our manual annotations we are able to disambiguate the self-adaptor regions. The torso region does not seem to be statistically different for any of the four conditions. However, the average duration of self-adaptors in the hand region is significantly longer for all four conditions. Further, self-adaptors in the head region are significantly longer for subjects with PTSD. These findings add to those of [13], where general grooming was identified to be correlated with thought disorders.

Further, our results confirmed the correlation between the longer durations of hands/legs fidgeting and psychological distress. In [13], hand tappings[1] were identified to be correlated with anxiety/depression disorders. In our analysis we could also confirm significantly longer leg fidgets for the condition distress. As part of our future work, we plan to develop an automatic descriptor for such behaviors so that they can be automatically detected and further investigated in future analysis.

VIII. CONCLUSION

In this study we analyzed a large dataset, namely the Distress Assessment Interaction Corpus (DAIC), of face-to-face interactions with a confederate interviewer and a paid participant. Within the DAIC we investigated the nonverbal behaviors of subjects with psychological disorders (i.e. depression, anxiety, PTSD, and distress) using both automatic behavior descriptors and manual annotations.

We focused our efforts on the behaviors, vertical gaze directionality, smile intensity and average duration, and self-adaptors and leg fidgeting. The gaze and smile behaviors were both analyzed using automatic behavior descriptors, whereas the fidgets were analyzed using manual annotations, as there are no current robust automatic descriptors for such behaviors available.

[1]This behavior falls in our analysis under the general term of hand fidgeting.

As reported in Section VII, we found several statistically significant differences in the nonverbal behavior of subjects in all four conditions (i.e. depression, anxiety, PTSD, and distress). Based on the three research goals stated in Section III we could identify the three main findings: (1) There are significant differences in the automatically estimated gaze behavior of subjects with psychological disorders. In particular, an increased overall downwards angle of the gaze could be automatically identified using two separate automatic measurements, for both the face as well as the eye gaze; (2) using automatic measurements, we could identify on average significantly less intense smiles for subjects with psychological disorders as well as significantly shorter average durations of smiles; (3) based on the manual analysis, subjects with psychological conditions exhibit on average longer self-touches and fidget on average longer with both hands (e.g. rubbing, stroking) and legs (e.g. tapping, shaking).

Whereas, we mainly analyzed the subject's behavior in the present study, for future work we plan to investigate audiovisual dyadic behaviors and patterns between the interviewer and the participant, in order to reveal additional indicators for both the presence and severity evaluation of psychological conditions. In [6], for example it was found that the clinician's behavior was strongly correlated with the patient's condition. Additionally, in [14] nonverbal attunement and entrainment was a strong predictor for the subsequent improvement of the patient's condition.

REFERENCES

[1] P. A. Arbisi, M. E. Kaler, S. M. Kehle-Forbes, C. R. Erbes, M. A. Polusny, and P. Thuras. The predictive validity of the ptsd checklist in a nonclinical sample of combat-exposed national guard troops. *Psychological Assessment*, page No Pagination Specified, 2012.

[2] T. Baltrusaitis, P. Robinson, and L.-P. Morency. 3D constrained local model for rigid and non-rigid facial tracking. In *IEEE Computer Vision and Pattern Recognition (CVPR 2012)*, Providence, RI, 2012.

[3] A. Beck, N. Epstein, G. Brown, and R. Steer. An inventory for measuring clinical anxiety: psychometric properties. *Journal of Consulting and Clinical Psychology*, 56:893–89, 1988.

[4] P. J. Bieling, M. M. Antony, and R. P. Swinson. The state–trait anxiety inventory, trait version: structure and content re-examined. *Behaviour Research and Therapy*, 36(7–8):777–788, 1998.

[5] E. B. Blanchard, J. Jones-Alexander, T. Buckley, and C. Forneris. Psychometric properties of the ptsd checklist (pcl). *Behaviour Research and Therapy*, 34(8):669–673, 1996.

[6] A. L. Bouhuys and R. H. van den Hoofdakker. The interrelatedness of observed behavior of depressed patients and of a psychiatrist: an ethological study on mutual influence. *Journal of Affective Disorders*, 23:63–74, 1991.

[7] L. M. Bylsam, B. H. Morris, and J. Rottenberg. A meta-analysis of emotional reactivity in major depressive disorder. *Clinical Psychology Review*, 28:676–691, 2008.

[8] D. Campbell, B. Felker, C.-F. Liu, E. Yano, J. Kirchner, D. Chan, L. Rubenstein, and E. Chaney. Prevalence of depression–ptsd comorbidity: Implications for clinical practice guidelines and primary care-based interventions. *Journal of General Internal Medicine*, 22:711–718, 2007.

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

[10] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978.

[11] P. Ekman and W. V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1:49–98, 1969.

[12] J. D. Elhai, L. de Francisco Carvalho, F. K. Miguel, P. A. Palmieri, R. Primi, and B. Christopher Frueh. Testing whether posttraumatic stress disorder and major depressive disorder are similar or unique constructs. *Journal of Anxiety Disorders*, 25(3):404–410, 2011.

[13] L. A. Fairbanks, M. T. McGuire, and C. J. Harris. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of Abnormal Psychology*, 91(2):109–119, 1982.

[14] E. N. Geerts, A. L. Bouhuys, and R. H. van den Hoofdakker. Nonverbal attunement between depressed patients and an interviewer predicts subsequent improvement. *Journal of Affective Disorders*, 40(1-2):15–21, 1999.

[15] J. A. Hall, J. A. Harrigan, and R. Rosenthal. Nonverbal behavior in clinician-patient interaction. *Applied and Preventive Psychology*, 4(1):21–37, 1995.

[16] F. Harris and B. Lahey. Recording system bias in direct observational methodology: A review and critical analysis of factors causing inaccurate coding behavior. *Clinical Psychology Review*, 2(4):539–556, 1982.

[17] L. V. Hedges. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128, 1981.

[18] A. Kirsch and S. Brunnhuber. Facial expression and experience of emotions in psychodynamic interviews with patients with ptsd in comparison to healthy subjects. *Psychopathology*, 40(5):296–302, 2007.

[19] K. Krippendorff. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112, 2011.

[20] K. Kroenke and R. L. Spitzer. The phq-9: A new depression and diagnostic severity measure. *Psychiatric Annals*, 32:509–521, 2002.

[21] K. Kroenke, R. L. Spitzer, and J. B. W. Williams. The phq-9. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[22] H. Lausberg and H. Sloetjes. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior research methods*, 41(3):841–849, 2009.

[23] G. N. Marshall, T. L. Schell, and J. N. V. Miles. All ptsd symptoms are highly associated with general distress: ramifications for the dysphoria symptom cluster. *Journal of Abnormal Psychology*, 119(1):126–135, 2010.

[24] R. Menke. *Examining nonverbal shame markers among post-pregnancy women with maltreatment histories*. PhD thesis, Wayne State University, 2011.

[25] L.-P. Morency, J. Whitehill, and J. Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *8th IEEE International Conference on Automatic Face Gesture Recognition (FG08)*, pages 1–8, 2008.

[26] J. E. Perez and R. E. Riggio. *Nonverbal social skills and psychopathology*, pages 17–44. Nonverbal behavior in clinical settings. Oxford University Press, 2003.

[27] L. I. Reed, M. Sayette, and J. F. Cohn. Impact of depression on response to comedy: A dynamic facial coding analysis. *Journal of Abnormal Psychology*, 116:804–809, 2007.

[28] J. A. Russell and L. F. Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805–819, 1999.

[29] J. T. M. Schelde. Major depression: Behavioral markers of depression and recovery. *The Journal of Nervous and Mental Disease*, 186(3):133–140, 1998.

[30] A. Sesti. State trait anxiety inventory in medication clinical trials. *Quality of Life Newsletter*, 25:15–16, 2000.

[31] C. D. Spielberger, R. L. Gorsuch, and L. R. E. *Manual for the State-Trait Anxiety Inventory*. Consulting Psychologists Press, 1970.

[32] J. Wagner, F. Lingenfelser, N. Bee, and E. André. Social signal interpretation (ssi). *KI - Kuenstliche Intelligenz*, 25:251–256, 2011.

[33] P. Waxer. Nonverbal cues for depression. *Journal of Abnormal Psychology*, 83(3):319–322, 1974.

[34] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:2106–2111, 2009.

[35] B. Wildman, M. Erickson, and R. Kent. The effect of two training procedures on observer agreement and variability of behavior ratings. *Child Development*, pages 520–524, 1975.