Image and Vision Computing 32 (2014) 648–658

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Automatic audiovisual behavior descriptors for psychological disorder analysis☆

Stefan Scherer [a,*], Giota Stratou [a], Gale Lucas [a], Marwa Mahmoud [a,b], Jill Boberg [a], Jonathan Gratch [a], Albert (Skip) Rizzo [a], Louis-Philippe Morency [a]

[a] University of Southern California Institute for Creative Technologies, Playa Vista, CA 90094, USA
[b] University of Cambridge, UK

ARTICLE INFO

Article history:
Received 17 June 2013
Received in revised form 25 April 2014
Accepted 13 June 2014
Available online 20 June 2014

Keywords:
Psychological distress
Depression
Post-traumatic stress disorder
Anxiety
Nonverbal behavior
Automatic assessment
Audiovisual

ABSTRACT

We investigate the capabilities of automatic audiovisual nonverbal behavior descriptors to identify indicators of psychological disorders such as depression, anxiety, and post-traumatic stress disorder. Due to strong correlations between these disordersas measured with standard self-assessment questionnaires in this study, we focus our investigations in particular on a generic distress measure as identified using factor analysis. Within this work, we seek to confirm and enrich present state of the art, predominantly based on qualitative manual annotations, with automatic quantitative behavior descriptors. We propose a number of nonverbal behavior descriptors that can be automatically estimated from audiovisual signals. Such automatic behavior descriptors could be used to support healthcare providers with quantified and objective observations that could ultimately improve clinical assessment. We evaluate our work on the dataset called the Distress Assessment Interview Corpus (DAIC) which comprises dyadic interactions between a confederate interviewer and a paid participant. Our evaluation on this dataset shows correlation of our automatic behavior descriptors with the derived general distress measure. Our analysis also includes a deeper study of self-adaptor and fidgeting behaviors based on detailed annotations of where these behaviors occur.

1. Introduction

The recent progress in facial feature tracking and articulated body tracking [1–3] has opened the door to new applications for automatic nonverbal behavior analysis.[1] 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 softwares 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 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 [5,6].

In this paper, we present and validate automatic behavior descriptors related to depression, anxiety and/or PTSD and in particular to a more generic distress measure introduced in Section 4.2. We introduce a new dataset, called the Distress Assessment Interview Corpus, which consists of 70 + h of dyadic interviews designed to study the verbal and nonverbal behaviors correlated with psychological disorders. We describe our approach in automatically assessing indicators of psychological disorders from head pose, eye gaze, facial expressions (smiles), and acoustic measures capturing the voice quality and monotonicity of the speech. We also investigate fidgeting and self-adaptor gestures occurring during these interviews.

The next section presents a previous work studying the relationship between nonverbal behaviors and psychological disorders. Section 3 introduces the research goals of this work. In Section 4 we describe the procedure for data acquisition, the used psychological measures, as well

☆ This paper has been recommended for acceptance by Qiang Ji.

* Corresponding author at: University of Southern California Institute for Creative Technologies 12015 Waterfront Dr. 90094, Playa Vista, CA.

E-mail addresses: scherer@ict.usc.edu (S. Scherer), Stratou@ict.usc.edu (G. Stratou), Lucas@ict.usc.edu (G. Lucas), emmam3@cam.ac.uk (M. Mahmoud), Boberg@ict.usc.edu (J. Boberg), gratch@ict.usc.edu (J. Gratch), rizzo@ict.usc.edu (A.(S.) Rizzo), morency@ict.usc.edu (L.-P. Morency).

[1] This work is an extension of the work in [4] originally published in the proceedings of the IEEE Automatic Face and Gesture Recognition Conference (FG) 2013.

http://dx.doi.org/10.1016/j.imavis.2014.06.001
0262-8856/

as the recorded population. Section 5 presents the multimodal behavior analysis platform MultiSense. The manual annotation scheme is introduced in Section 6, and the observed results of the automatic and manual analysis are presented and discussed in Section 7. Finally, Section 8 concludes the paper and introduces future directions of our work.

2. 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 on 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. Several research programs around the globe study these relationships, including a project funded by the Australian Research Council [7–12], a Department of Defense funded project [13–15], and the DARPA funded project that the present work is funded on [4,16–23]. General findings from these research programs and other investigations 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 1.

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 [24], show nonspecific gaze, such as staring off into space [25] 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.

Table 1
Summary of nonverbal behaviors found in the literature. Nonverbal behaviors written in italics are part of the analysis in the present work.

Authors	Nonverbal behavior	Disorder
Fairbanks et al., 1982	↓ Mouth movements	Depression
	↓ *Smiling*	
	↑ *Self-grooming*	
	↑ *Turning head away*	
	↑ *Fidgeting*	Anxiety
Girard et al., 2013	↓ Smiles	Depression
	↑ Smile controls	
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
Darby et al., 1984	↓ *Pitch variability*	Depression
	↓ *Loudness variability*	
	↑ *Harsh voice*	
	↑ *Speech monotonicity*	
Flint et al., 1993	↑ *Vocal tension*	Depression

Emotional expressivity, such as the frequency or duration of smiles, is also indicative of an underlying clinical state. For example, depressed patients frequently display flattened or negative affect including less emotional expressivity [26,27], fewer mouth movements [28,25], more frowns [28,26] and fewer gestures [29,26]. Some findings suggest that 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 [30] than what is found in non-clinical populations. Social anxiety and PTSD share some of the features of depression and 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 [30], or shame [31].

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 [28]. Similarly, “self-adaptors”, such as rhythmically touching, hugging or stroking parts of the body or self-grooming, e.g. repeatedly stroking the hair [28], have been identified to be of interest in this field of research [32].

Also acoustic indicators for depression were investigated and reduced speech variability and monotonicity in loudness and pitch were found [33,34]. Further, depressed speech was found to show increased tension in the vocal tract and the vocal folds [35,19]. In particular, increased tense voice quality characteristics within a comparable recording setup to the one investigated in the present work were found [19]. In this work a virtual human interview setup was investigated [19], whereas here we investigate human to human interviews.

These findings of indicative speech parameters have led to additional classification experiments [36]; the analysis involved glottal flow features as well as prosodic features for the discrimination of depressed read speech. The authors identified glottal flow features to be chosen by the feature selection algorithm for the majority of the classifiers as well as energy-based features for female speakers. Several spectral and energy based features were investigated for their discriminative capabilities of read speech using Gaussian mixture models, with Mel frequency cepstral coefficients and the first three formants yielding promising results [8]. Also, acoustic spectral measures associated with psychomotor retardation at different time resolutions are investigated in an international challenge to identify depression severity in the subject's voice characteristics [13].

More recently vocal and facial expressions were found as indicators of depression severity using within-subject analysis of longitudinal data. Both participants' and interviewers' vocal timing and fundamental frequency were found to correlate with Hamilton Rating Scale for Depression scores [37]. For facial expressions within the same longitudinal dataset, both manually and automatically coded facial action coding scheme action units (AU) varied markedly with depressive symptom severity. Significantly lower overall AU 12 activity (i.e. fewer smiles), significantly higher overall AU 14 activity (associated with contempt), and significantly more AU 14 activity during smiling, i.e. smile controls were observed [38].

Few multimodal studies are found in the literature with [6] being one of the exceptions. Facial action units and variability of fundamental frequency (f_0) as well as latency to respond to questions have been investigated [6]. Both approaches, yield promising discriminative power with about 80% accuracy for each modality.

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 [39]. 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 todifferences along broad dimensions such as valence and arousal. Indeed, when it comes to emotion recognition, some meta-reviews