Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Psychology

Automated audiovisual depression analysis

Jeffrey M Girard and Jeffrey F Cohn

Analysis of observable behavior in depression primarily relies on subjective measures. New computational approaches make possible automated audiovisual measurement of behaviors that humans struggle to quantify (e.g., movement velocity and voice inflection). These tools have the potential to improve screening and diagnosis, identify new behavioral indicators of depression, measure response to clinical intervention, and test clinical theories about underlying mechanisms. Highlights include a study that measured the temporal coordination of vocal tract and facial movements, a study that predicted which adolescents would go on to develop depression based on their voice qualities, and a study that tested the behavioral predictions of clinical theories using automated measures of facial actions and head motion.

Addresses
Department of Psychology, University of Pittsburgh, Sennott Square,
210 South Bouquet Street, Pittsburgh, PA 15260, USA

Corresponding author: Cohn, Jeffrey F (jeffcohn@pitt.edu)

Current Opinion in Psychology 2015, **4**:75–79

This review comes from a themed issue on **Depression**

Edited by **Christopher G Beevers**

For a complete overview see the Issue and the Editorial

Available online 19th December 2014

http://dx.doi.org/10.1016/j.copsyc.2014.12.010

Introduction

Depression has salient, observable behavioral symptoms pertaining to general psychomotor functioning, the expression of affective states, and the negotiation of interpersonal situations. Current methods for the diagnosis and assessment of depression rely on subjective measures of behavior, such as self-report or family-report and clinical interviews. Such measures are useful only to the extent that they can be explicitly defined and reliably assessed. Automated methods for behavior analysis — the product of recent advances in computer vision, signal processing, and affective computing — have the potential to powerfully inform assessment and understanding of depression. Efforts in this direction are underway.

The current article reviews empirical studies from 2013 to 2014 that use automated methods to analyze depression from audiovisual data captured using telephones, microphones, and video cameras. These studies and the methods they promote impact one or more of four applications: (1) identifying behavioral indicators of depression, (2) screening and diagnosis, (3) measuring response to intervention, and (4) testing clinical theories about underlying mechanisms. Some of these applications have been more researched than others, but all have the potential (and are beginning) to contribute to advances in clinical science and practice. After providing an overview of contemporary automated methods for audiovisual behavior analysis, the current article reviews their contribution to each application area in turn. Finally, future directions and ongoing challenges are outlined.

Overview of automated methods

Visual behavior analysis

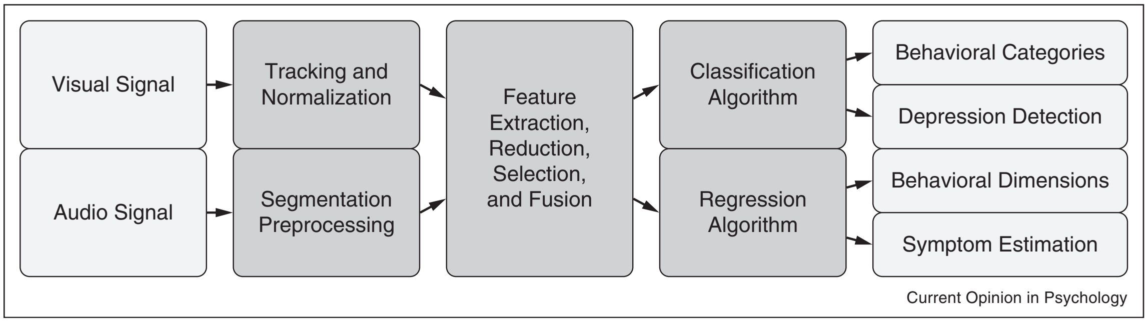
Facial expressions, eye gaze, head and body movements, posture, and gesture are communicated visually. Although numerous approaches exist for the automated analysis of such behaviors [1] and new developments are currently underway, most researchers have converged on the same basic structure of analysis (Figure 1).

First, the relevant body parts (e.g., head, torso, or hands) are detected within each video frame. This process is typically achieved by searching within the video frame for regions that match previously learned models of specific body parts. Next, features are extracted from these regions that quantify their shape and/or appearance. These features are frequently extracted from the body part models themselves or from orientation-sensitive filters applied to the regions in a process similar to primate vision [2]. To control for variation in head pose and size, features are usually registered to a common view. Finally, an algorithm is developed to interpret the features. This process is typically achieved through *supervised learning*, wherein human-verified examples are provided to a classification or regression algorithm, which learns a generalizable mapping between the features and various behavioral categories or dimensions; novel video frames can then be interpreted by extrapolating from this learned mapping. A recent study found that, when applied to individual facial actions, such methods are robust to changes in participant gender and ethnicity, as well as to the range of head pose and illumination changes common in spontaneous data [3].

Acoustic behavior analysis

Speech, back-channeling, vocal pauses, and voice quality are communicated through audio signals. Although some researchers are working on automated analysis of the lexical, syntactic, and semantic content of audio signals, we focus on *paralinguistic*, or *prosodic*, features. These are

Figure 1



Standard structure of analysis for automated audiovisual behavior analysis.

perceived by listeners in terms of pitch, loudness, speaking rate, rhythm, voice quality, and articulation. They can be measured from recordings of spontaneous or scripted speech and quantified using a variety of parameters, such as cepstral, glottal, and spectral features. The most common prosodic features are *intra*-personal (e.g., pauses between utterances within a speaking turn); however, recent research has begun to focus on *inter*-personal features (e.g., switching pauses between two speakers) as well [4•].

Before extracting prosodic features from an audio signal, it is useful to segment participant speech from periods of silence, noise, and the speech of other parties. Some studies accomplish segmentation automatically (or semi-automatically) through transcription and forced alignment, while others manually segment the audio or record only specific segments. The type of segmentation used impacts how deployable an automated system is, as well as which features it can analyze. For instance, many interpersonal features are only analyzable when using multiparty segmentation.

Automated depression analysis

Three main approaches to analyzing depression from audiovisual information have been proposed. The first approach compares individual behaviors between groups defined by diagnosis or symptom severity, typically employing null hypothesis significance testing to compare group means. The second approach uses classification algorithms to assign participants to two or more mutually exclusive groups using high dimensional audiovisual features. Finally, the third approach uses regression algorithms to estimate participants' symptom severity using high dimensional audiovisual features.

Within each approach, some studies examined clinical samples diagnosed with depressive disorders, while others measured depressive symptoms in non-clinical samples. Studies using diagnostic inclusion criteria have better specificity for depression than those that use symptom-rating measures to define depression, and those that study clinical samples may have better generalizability to patient populations than those that include only non-clinical samples.

Identifying behavioral indicators of depression

The DSM-5 describes a range of audiovisual indicators of depression [5, pp. 160–164]. These include tears or crying for depressed mood; facial expression and demeanor for sadness; inability to sit still, pacing, hand-wringing, or pulling or rubbing the skin (i.e., self-adaptors) for psychomotor agitation; and slowed speech or body movements, longer vocal pauses, and decreased volume and inflection for psychomotor retardation.

Automated measurement has a vital role to play in operationalizing these behaviors, identifying which ones reliably indicate depression and its symptoms, and identifying distributions of typical and atypical behavior. Studies using the mean-comparison approach to depression analysis are well-suited to this application, as they illuminate the differences between various groups in terms of specific behaviors.

Such studies have identified potential indicators of depression that can be measured automatically. Recent examples from the visual channel include smaller average distance between eyelids and shorter duration of blinks [6], slower head movements [7,8••], less head motion [7,8••,9,10], longer duration of looking down [7,11•], decreased smiling [8••,11•], decreased frowning, and increased mouth dimpling [8••]. Recent examples from the acoustic channel include increased voice tension [11•], decreased coordination among formant frequencies and cepstral channels [12], longer and more variable switching pauses [4•], and decreased dyadic synchrony [13].

Several studies have begun to explore how specific behaviors are related to individual depressive symptoms or subindices of self-reported symptomatology [12,14•,15]. Such analyses are useful for evaluating issues of specificity