Automated Audiovisual Depression Analysis

Jeffrey M. Girard[a], Jeffrey F. Cohn[a,*]

[a]*Department of Psychology, University of Pittsburgh*
Sennott Square, 210 S. Bouquet Street, Pittsburgh, PA, USA 15260

Abstract

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.

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- 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 2013–2014 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.

*Corresponding author. T: 1-412-624-8826; F: 1-412-624-2023
Email addresses: jmg174@pitt.edu (Jeffrey M. Girard), jeffcohn@pitt.edu (Jeffrey F. Cohn)

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

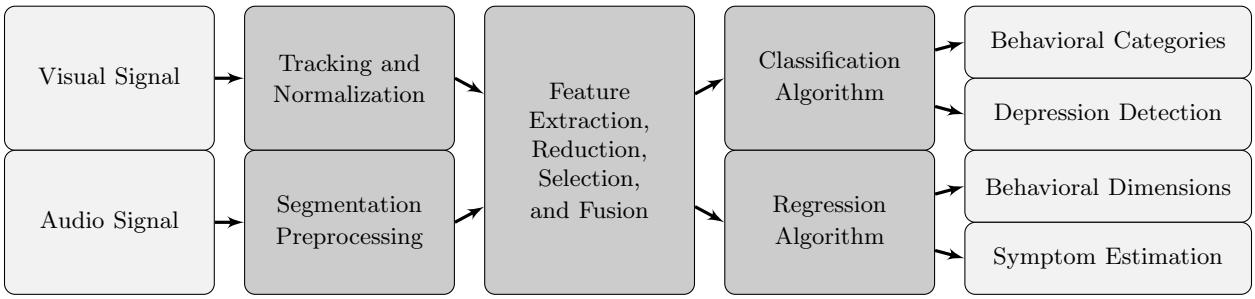

Figure 1: Standard structure of analysis for automated audiovisual behavior analysis

lexical, syntactic, and semantic content of audio signals, we focus on *paralinguistic*, or *prosodic*, features. These are 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 [7–10], 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 sub-indices of self-reported symptomatology [12, 14, 15]. Such analyses are useful for evaluating issues of specificity and individual differences. Several studies have also begun to report the distributions (e.g., box plots) of individual behaviors across different groups [9–11], which is the first step towards establishing behavioral "norms" that novel data can be compared against.

Aiding in the screening and diagnosis of depression

The most common application in recent years has been automated depression recognition (i.e., screening). This

popularity is related to several organized "challenges" that provide common data for multiple research groups to analyze and compare results [16, 17]. While the approaches best suited to this application are group classification and symptom severity estimation, mean differences can also be informative.

Many studies have attempted to use automated behavioral measures to detect depression and estimate symptom severity. Numerous techniques for feature extraction and supervised learning have been proposed and evaluated. Some used visual features only [6, 18, 19], while others used acoustic features only [4, 12–15, 20–26]. Incremental gains have been found by fusing information from both modalities [27, 28] and many studies have begun to do so [27–37].

Because the performance metrics for supervised learning are often sensitive to the distribution of classes [38], they can be difficult to compare between studies. That said, classification accuracy for detecting depression has been above chance, typically falling between 70% and 80%, with some studies reporting as high as 90% accuracy. One study of particular interest is Ooi et al. [24], which compared the behavior of adolescents who would go on to develop depression against those who would not; using acoustic features, the authors were able to distinguish these groups with 73% accuracy. For the estimation of symptom severity, results have also been well above chance. For example, the winner of the 2013 Depression Recognition Challenge [16] scored a mean absolute error of 5.75, far better than the baseline score of 10.28 [12].

Measuring change over time and response to intervention

In addition to identifying those at risk for depression, automated methods can be used to track behavioral symptoms over time (e.g., during intervention studies and ongoing clinical practice). Within-subject comparisons also circumvent the confounding influence of stable characteristics that correlate with depression risk and independently influence behavior (e.g., high neuroticism and low extraversion [39]).

Relative to the other applications, measuring change over time and response to intervention has been less explored. This is likely due to the high cost and time commitment of collecting longitudinal data. Two studies using same clinical sample have examined this issue directly. These studies followed depressed participants over the course of treatment and examined the relation between symptom severity and behavior during clinical interviews. Girard et al. [8] found that, when severely depressed, participants showed reduced head motion, reduced smiling and frowning, and increased mouth "dimpling." Yang et al. [4] found that participants showed longer and more variable switching pauses when more severely depressed. Interviewer behavior also changed when interacting with more depressed participants, becoming lower and more variable in pitch with longer and more variable switching pauses.

Exploring clinical theories and underlying mechanisms

Several clinical theories attempt to explain the relationship between depression and its behavioral manifestations through reference to underlying mechanisms in the affective, neurobiological, and psychosocial domains. Researchers using automated methods for behavior analysis can make important contributions to clinical science by formulating and testing the behavioral hypotheses of these and novel theories.

Studies exploring the relationship between behavioral measures and clinical ratings of psychomotor retardation are well-suited to exploring neurobiological theories [12–15]. Of particular interest is work by Williamson et al. [12, 37], which found that depression is related to decreased coordination between acoustic parameters (i.e., formant frequencies and cepstral channels) and facial actions. Attenuated correlation within behavioral channels may reflect dysregulation in central and autonomic control of the body (e.g., the vocal track and facial muscles).

Studies exploring interpersonal measures are well-suited to exploring psychosocial theories; findings that depression is related to decreased dyadic synchrony [13] and longer and more variable switching pauses [4] suggest that depression interferes with interpersonal functioning in a measurable way. Of particular interest is an article by Girard et al. [8], which tested the hypotheses of three clinical theories and, using automated analyses of facial expressions and head motion, found the strongest support for a novel psychosocial theory that nonverbal behavior serves to facilitate social withdrawal during periods of severe depression.

Conclusions

Automated methods for behavior analysis can provide measurements that are difficult for humans to quantify (e.g., velocity and inflection) and have the benefit of high repeatability. While clinicians may vary in their degree of accuracy and consistency, anyone implementing the same automated system (within its operational parameters) can be confident that it will perform consistently. These tools have the potential to validate behavioral indicators of depression, aid in screening and diagnosis, measure response to intervention, and test clinical theories about underlying mechanisms.

Future directions for the field include exploring additional behavioral indicators and nonverbal modalities, as well as temporal dynamics, dyadic coordination, and contextual effects. Eye gaze, body movements, posture, and gesture are all ripe for automated audiovisual analysis. Exploring the temporal dynamics of behavior may also prove

important for understanding depression and psychomotor retardation; these topics have been relatively untapped, especially in the visual domain. Exploring measures of dyadic coordination, such as synchrony and accommodation could prove invaluable in understanding depression and social withdrawal. Finally, the influence of social and experimental context is likely to be critical in understanding the nuanced functions and consequences of depressive behavior.

Several salient challenges remain. It is of paramount importance to determine and enhance the generalizability of methods and results to diverse samples, recording conditions, cultures, ethnicities, ages, and genders. In service of this goal, the field must also work to enhance the comparability of its results. This means more shared datasets and open challenges, as well as increased consistency in terms of the diagnostic criteria, clinical instruments, and performance metrics used. Finally, the issue of clinical specificity must be grappled with, as there are many psychiatric and medical disorders that co-occur with or mimic depression.

We hope this review has offered insight into the current state of automated audiovisual depression analysis. The past two years have seen an explosion of interest and activity in these methods; the next two years are sure to usher in unprecedented gains in their development, validation, and deployment.

Acknowledgments

This work was supported in part by US National Institutes of Health grant MH096951 to the University of Pittsburgh. Neither the National Institutes of Health nor any other funding source was involved in the planning and writing of the article or in the decision to submit the article for publication.

[1] Cohn JF, De la Torre F: Automated face analysis for affective computing. In *Handbook of affective computing*, Edited by Calvo RA, D'Mello SK, Gratch J, Kappas A. Oxford New York; 2014.

[2] Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T: A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *Artificial Intelligence* 2005:1–130.

[3] Girard JM, Cohn JF, Sayette MA, Jeni LA, De la Torre F: Spontaneous facial expression in unscripted social interactions can be measured automatically. *Behavior Research Methods* In Press.

[4] Yang Y, Fairbairn CE, Cohn JF: Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing* 2013, 4:142–150.

> ∗ The authors demonstrate that naive listeners can perceive depression severity from vocal recordings of patients and clinical interviewers. Then, using automated analyses and a longitudinal sample, they explore the relationship between symptom severity and both intrapersonal and interpersonal prosodic measures. This study demonstrates that depression can affect the behavior of interacting parties (i.e., the interviewers).

[5] American Psychiatric Association: *Diagnostic and statistical manual of mental disorders*. Washington, DC, 5th edition, 2013.

[6] Alghowinem S, Goecke R, Wagner M, Parker G, Breakspear M: Eye movement analysis for depression detection. In *IEEE International Conference on Image Processing* 2013:4220–4224.

[7] Alghowinem S, Goecke R, Wagner M, Parker G, Breakspear M: Head pose and movement analysis as an indicator of depression. In *Humaine Association Conference on Affective Computing and Intelligent Interaction* 2013:283–288.

[8] Girard JM, Cohn JF, Mahoor MH, Mavadati SM, Hammal Z, Rosenwald DP: Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and Vision Computing* 2014, 32:641–647.

> ∗∗ Using both manual and automated analyses of facial expressions and head motion, the authors find support for a novel psychosocial theory: that nonverbal behavior serves to facilitate social withdrawal during episodes of severe depression. This is the first study to demonstrate that automated analyses can find the same results as manual analyses and is one of the few to compare behavior within a longitudinal, clinical sample.

[9] Stratou G, Scherer S, Gratch J, Morency LP: Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences. In *Humaine Association Conference on Affective Computing and Intelligent Interaction* 2013:147–152.

[10] Scherer S, Stratou G, Morency LP: Audiovisual behavior descriptors for depression. In *ACM International Conference on Multimodal Interaction* 2013:135–140.

[11] Scherer S, Stratou G, Lucas G, Mahmoud M, Boberg J, Gratch J, Rizzo AS, Morency LP: Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing* 2014, 32:648–658.

> ∗ Using automated analyses, the authors compare the nonverbal behavior of participants with high and low levels of psychological distress, including depressive symptomatology. This study analyses a multitude of behaviors from both the visual and acoustic channels (i.e., head gaze, eye gaze, speech prosody, smile intensity and duration) and provides explicit distributions for the behavior of both groups.

[12] Williamson JR, Street W, Quatieri TF, Helfer BS, Horwitz R, Yu B: Vocal Biomarkers of Depression Based on Motor Incoordination. In *ACM International Workshop on Audio/Visual Emotion Challenge* 2013:41–47.

[13] Scherer S, Hammal Z, Yang Y, Morency LP, Cohn JF: Dyadic behavior analysis in depression severity assessment interviews. *Proceedings of the 16th International Conference on Multimodal Interaction* 2014:112–119.

[14] Horwitz R, Quatieri TF, Helfer BS, Yu B, Williamson JR, Mundt J: On the relative importance of vocal source, system, and prosody in human depression. In *IEEE International Conference on Body Sensor Networks* 2013:2–7.

> ∗ The authors explore the effect of depression on three aspects of speech: source, system, and melody. Automatically-measured acoustic parameters are shown to correlate with overall symptom severity, as well as with individual symptom measures. Contextual effects of free-response versus read speech are also explored.

[15] Cummins N, Epps J, Ambikairajah E: Spectro-temporal analysis of speech affected by depression and psychomotor retardation. In *IEEE International Conference on Acoustics, Speech and Signal Processing* 2013:7542–7546.

[16] Valstar M, Schuller B, Smith K, Eyben F, Jiang B, Bilakhia S, Schneider S, Cowie R, Pantic M: AVEC 2013: The continuous audio/visual emotion and depression recognition challenge.

In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge* 2013:3–10.

[17] Valstar M, Schuller BW, Krajewski J, Cowie R, Pantic M: AVEC 2014: the 4th international audio/visual emotion challenge and workshop. *Proceedings of the ACM International Conference on Multimedia* 2014:1243–1244.

[18] Joshi J, Dhall A, Goecke R, Cohn JF: Relative body parts movement for automatic depression analysis. In *Humaine Association Conference on Affective Computing and Intelligent Interaction* 2013:492–497.

[19] Joshi J, Goecke R, Parker G, Breakspear M: Can body expressions contribute to automatic depression analysis? In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* 2013:1–7.

[20] Alghowinem S, Goecke R, Wagner M, Epps J, Gedeon T, Breakspear M, Parker G: A comparative study of different classifiers for detecting depression from spontaneous speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing* 2013:8022–8026.

[21] Cummins N, Sethu V, Epps J, Krajewski J: Probabilistic acoustic volume analysis for speech affected by depression. In *Annual Conference of the International Speech Communication Association* 2014:1238–1242.

[22] Giannakopoulos T, Smailis C, Perantonis S, Spyropoulos C: Realtime depression estimation using mid-term audio features. In *International Workshop on Artificial Intelligence and Assistive Medicine* 2014:41–46.

[23] Lopez-Otero P, Docio-Fernandez L, Garcia-Mateo C: A study of acoustic features for the classification of depressed speech. *Proceedings of the International Convention on Information and Communication Technology, Electronics and Microelectronics* 2014:1331–1335.

[24] Ooi KEB, Lech M, Allen NB: Multichannel weighted speech classification system for prediction of major depression in adolescents. *IEEE Transactions on Biomedical Engineering* 2013, 60:497–506.

** The authors use four types of automatically-measured acoustic parameters (i.e., prosodic, spectral, glottal, and Teager's energy operator) to predict which nondepressed adolescents will go on to develop clinical depression within the next two years. Using a novel multichannel classification method, they achieve a classification accuracy of 73%. This is the first study to attempt the important task of prospective screening.

[25] Lopez-Otero P, Docio-Fernandez L, Garcia-Mateo C: A study of acoustic features for depression detection. In *International Workshop on Biometrics and Forensics* 2014:1–6.

[26] Mitra V, Shriberg E, McLaren M: The SRI AVEC-2014 Evaluation System. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* 2014:93–101.

[27] Joshi J, Goecke R, Alghowinem S, Dhall A, Wagner M, Epps J, Parker G, Breakspear M: Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces* 2013, 7:1–18.

[28] Espinosa HP, Escalante HJ, Villaseñor pineda L, Montes-y gómez M, Pinto-Avedaño D, Reyes-Meza V: Fusing affective dimensions and audio-visual features from segmented video for depression recognition. In *ACM International Workshop on Audio/Visual Emotion Challenge* 2014.

[29] Cummins N, Joshi J, Dhall A, Sethu V, Goecke R, Epps J: Diagnosis of depression by behavioural signals: A multimodal approach. In *ACM International Workshop on Audio/Visual Emotion Challenge* 2013:11–20.

[30] Gupta R, Malandrakis N, Xiao B, Guha T, Van Segbroeck M, Black MP, Potamianos A, Narayanan SS: Multimodal prediction of affective dimensions and depression in human-computer interactions. In *International Audio/Visual Emotion Challenge and Workshop* 2014.

[31] Kaya H, Salah AA: Eyes whisper depression. In *ACM International Workshop on Audio/Visual Emotion Challenge* 2013:1–4.

[32] Meng H, Hunag D, Wang H, Yang H, Al-shuraifi M, Wang Y: Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *ACM International Workshop on Audio/Visual Emotion Challenge* 2013:21–30.

[33] Jan A, Meng H, Gaus YFA, Zhang F, Turabzadeh S: Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression Categories and Subject Descriptors. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* 2014:73–80.

[34] Jain V, Crowley JL, Dey AK, Lux A: Depression Estimation Using Audiovisual Features and Fisher Vector Encoding. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* 2014:87–91.

[35] Sidorov M, Minker W: Emotion Recognition and Depression Diagnosis by Acoustic and Visual Features: A Multimodal Approach. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* 2014:81–86.

[36] Senoussaoui M, Sarria-Paja M, Santos JaF, Falk TH: Model Fusion for Multimodal Depression Classification and Level Detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* 2014:57–63.

[37] Williamson JR, Street W, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD: Vocal and Facial Biomarkers of Depression Based on Motor Incoordination and Timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* 2014:65–72.

** The authors estimate self-reported depressive symptom severity with high accuracy using the timing and coordination of automatically-measured speech parameters and facial actions. This work demonstrates that much can be learned about depression from the interactions and timing of different behavioral channels. These findings reveal evidence of degraded coordination among the neurocognitive subsystems of expression.

[38] Jeni LA, Cohn JF, De la Torre F: Facing imbalanced data: Recommendations for the use of performance metrics. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* 2013:245–251.

[39] Kotov R, Gamez W, Schmidt F, Watson D: Linking big personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin* 2010, 136:768–821.