Multimodal Detection of Depression in Clinical Interviews

Hamdi Dibeklioğlu[1,*], Zakia Hammal[2,*], Ying Yang[3], and Jeffrey F. Cohn[2,4]
[1]Pattern Recognition and Bioinformatics Group, Delft University of Technology, Delft, The Netherlands
[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[3]Center for Cognitive Brain Imaging, Carnegie Mellon University, Pittsburgh, PA, USA
[4]Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA
h.dibeklioglu@tudelft.nl, zakia_hammal@yahoo.fr, yingyang02@gmail.com, jeffcohn@cs.cmu.edu

ABSTRACT

Current methods for depression assessment depend almost entirely on clinical interview or self-report ratings. Such measures lack systematic and efficient ways of incorporating behavioral observations that are strong indicators of psychological disorder. We compared a clinical interview of depression severity with automatic measurement in 48 participants undergoing treatment for depression. Interviews were obtained at 7-week intervals on up to four occasions. Following standard cut-offs, participants at each session were classified as remitted, intermediate, or depressed. Logistic regression classifiers using leave-one-out validation were compared for facial movement dynamics, head movement dynamics, and vocal prosody individually and in combination. Accuracy (remitted versus depressed) for facial movement dynamics was higher than that for head movement dynamics; and each was substantially higher than that for vocal prosody. Accuracy for all three modalities together reached 88.93%, exceeding that for any single modality or pair of modalities. These findings suggest that automatic detection of depression from behavioral indicators is feasible and that multimodal measures afford most powerful detection.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications; J.4 [**Social and Behavioral Sciences**]: Psychology

Keywords

Depression; Facial Movement; Head Movement; Vocal Prosody

1. INTRODUCTION

Depression is one of the most common and recurrent psychological disorders and a leading cause of disease burden [23]. The World Health Organization predicts that depression will become the leading cause of disease burden within the next 15 years [21]. Reliable assessment is critical to understand mechanisms, discover and implement more effective treatment, and ameliorate this trend.

Many symptoms of depression are observable. They profoundly influence the behavior of affected individuals and the reactions of their family members and others. DSM-5 [4] describes a range of audiovisual indicators. These include facial expression and demeanor, inability to sit still, pacing, hand-wringing and other signs of psychomotor agitation, slowed speech and body movements, reduced interpersonal responsiveness, and decreased vocal intensity and inflection in psychomotor retardation. Yet, screening and diagnosis of depression fail to exploit these signs and symptoms. They rely instead almost entirely on patients' verbal reports. Recent advances in automatic analysis of face, voice, and body movement could play a vital role in the clinical science of depression, screening efforts, treatment, and community intervention.

Automatic assessment of depression from behavioral cues is of increasingly interest to both psychologists and computer scientists. The latter use signal processing, computer vision, and pattern recognition methodologies. From the computer-science perspective, efforts have sought to identify depression from vocal utterances [10], facial expression [9, 25], head movements/pose [1, 20, 25], body movements [20], and gaze [2]. While most research is limited to a single modality, there is increasing interest in multimodal approaches to depression detection [22, 26]. Multimodal assessment raises several issues. One is whether one or another modality is more informative? Ekman [13] claimed that facial expression is less revealing than body and was equivocal about face relative to voice. Alternatively, one could imagine that high redundancy across channels would render modest any potential gain of a multimodal approach. Comparative studies are needed to explore these questions. Two is choice of context. Audio/Visual Emotion Challenge (AVEC) explored multimodal expression during an individual task, for which audience effects would likely be absent. Research by Fridlund and others [16] suggests that when other people are present, signal strength of nonverbal behavior increases. Nonverbal reactions to and from others present additional sources of information. Three, feature selection and approaches to fusion are critical. Several studies are relevant to these issues.

We investigated the discriminative power of three modalities – facial movement dynamics, head movement dynamics, and vocal prosody– individually and in combination. Instead of using a large number of descriptors, or selecting informative features individually, an optimum feature set was obtained by maximizing the combined mutual information. In contrast to previous work, we focused on both individual and interpersonal behavior. Informed by the psychology literature, we anticipated that interpersonal context would provide powerful detection of depression. To capture aspects of psychomotor retardation and agitation, we include dynamic measures of expressive behavior. Additionally, in contrast to previous work, all participants met DSM criteria for major depression as

*These authors contributed equally.

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.

DOI: http://dx.doi.org/10.1145/2818346.2820776.

determined by diagnostic interview, and symptom severity and recovery were ground-truth using state-of-the-art depression severity interviews. Diagnostic criteria are important because many non-depressive disorders are confusable with depression. By using diagnostic criteria, we were able to rule out participants that may have had PTSD or other related disorder.

2. MATERIALS

2.1 Participants

Fifty-seven depressed participants (34 women, 23 men) were recruited from a clinical trial for treatment of depression. They ranged in age from 19 to 65 years (μ=39.65) and were Euro- or African-American (46 and 11, respectively). At the time of study intake, all met DSM-IV [3] criteria for Major Depressive Disorder [14]. Although not a focus of this report, participants were randomized to either anti-depressant treatment with a selective serotonin re-uptake inhibitor or Interpersonal Psychotherapy. Both treatments are empirically validated for treatment of depression [18]. Data from 48 participants were available for analysis. Participant loss was due to change in original diagnosis, severe suicidal ideation, and methodological reasons (e.g., missing audio/video).

2.2 Observational Procedures

Symptom severity was evaluated on four occasions at 1, 7, 13, and 21 weeks by ten female clinical interviewers. Interviewers were not assigned to specific participants, and they varied in the number of interviews they conducted. Four interviewers were responsible for the bulk of the interviews. Interviews were conducted using the Hamilton Rating Scale for Depression (HRSD), which is a criterion measure of depression severity [17]. Interviewers were expert in the HRSD and reliability was maintained above 0.90. HRSD scores $>$14 are considered to indicate moderate to severe depression; scores $<$8 indicate return to normal [15]. These cut-off scores were used to define depression and remission status. Sessions with scores between 7 and 15 were removed from analyses.

Video was obtained from a VGA camera positioned about 15° from frontal view. Audio from participants and interviewers was digitized at 48 kHz and later down-sampled to 16 kHz for speech processing. Missing data occurred from missed appointments and technical problems. To be included for analysis, we required a minimum of 20 speaker turns of 3 seconds minimum duration and at least 50 seconds of vocalization in total. The final sample was 95 sessions from 48 participants; 58 depressed and 37 remitted.

3. METHODS

Since depression can impact a wide range of behavior, we propose a multimodal assessment from facial movement, head movement, and vocal prosody.

3.1 Tracking of Landmarks and Head Pose

To track and align facial shape and head pose, we used a person-independent approach (ZFace) [19], which accomplishes 3D registration from 2D video without requiring person-specific training (for details, see [19]). Forty-nine facial landmarks (fiducial points on the regions of eyebrows, eyes, nose, and mouth), and the three degrees of rigid head movements (i.e., pitch, yaw, and roll) were tracked.

3.1.1 Facial Movement Dynamics

Previous research has found that facial movement dynamics are discriminative for facial expressions [8, 12]. We investigated the reliability of depression severity assessment from facial movement dynamics. To control for variation due to rigid head movement, fiducial points were normalized by removing translation, rotation, and scale. The movement of the normalized 98 time series (49 fiducial points $\times$ x and y coordinates) was then smoothed by the 4253H-twice method [27]. For a compact representation, principal component analysis was used to reduce the 98 time series to 15 time series components $\mathcal{D}_i$, where $i = \{1, 2, 3, \ldots, 15\}$. These 15 components accounted for 95% of the variance; they are referred to as amplitude of facial movement. The velocity ($\mathcal{V}_i = \frac{d\mathcal{D}_i}{dt}$) and acceleration ($\mathcal{A}_i = \frac{d^2\mathcal{D}_i}{dt^2}$) of change in the extracted 15 time series were computed as the derivatives of the corresponding amplitude. Based on previous research [11], each time series is divided into increasing ($^+$) and decreasing ($^-$) segments for dynamic feature extraction. Twenty one features were then used to measure the dynamics of facial displacement, velocity, and acceleration, respectively, as described in Table 1. By concatenating the features extracted from the 15 time series components, 315-dimensional (21 features $\times$ 15 time series components) dynamic features were obtained.

Table 1: Twenty one dynamic features for facial fiducial points and head pitch, yaw, and roll.

Feature	Definition
Maximum Ampl.:	$\max(\mathcal{D})$
Mean Amplitude:	$\left[\frac{\sum \mathcal{D}}{\eta(\mathcal{D})} , \frac{\sum \mathcal{D}^+}{\eta(\mathcal{D}^+)} , \frac{\sum \mathcal{D}^-}{\eta(\mathcal{D}^-)} \right]$
STD of Amplitude:	$\text{std}(\mathcal{D})$
Maximum Speed:	$\left[\max(\lvert\mathcal{V}\rvert) , \max(\lvert\mathcal{V}^+\rvert) , \max(\lvert\mathcal{V}^-\rvert) \right]$
Mean Speed:	$\left[\frac{\sum \mathcal{V}}{\eta(\mathcal{V})} , \frac{\sum \mathcal{V}^+}{\eta(\mathcal{V}^+)} , \frac{\sum \mathcal{V}^-}{\eta(\mathcal{V}^-)} \right]$
STD of Speed:	$\text{std}(\mathcal{V})$
Maximum Accel.:	$\left[\max(\lvert\mathcal{A}\rvert) , \max(\lvert\mathcal{A}^+\rvert) , \max(\lvert\mathcal{A}^-\rvert) \right]$
Mean Accel.:	$\left[\frac{\sum \mathcal{A}}{\eta(\mathcal{A})} , \frac{\sum \mathcal{A}^+}{\eta(\mathcal{A}^+)} , \frac{\sum \mathcal{A}^-}{\eta(\mathcal{A}^-)} \right]$
STD of Accel.:	$\text{std}(\mathcal{A})$
$^+/^-$ Frequency:	$\left[\frac{\tau^+}{\eta(\mathcal{A}^+)} , \frac{\tau^-}{\eta(\mathcal{A}^-)} \right]$

3.1.2 Head Movement Dynamics

Head angles in the horizontal, vertical, and lateral directions were used to measure head movement. These directions correspond to head nods (i.e. pitch), head turns (i.e. yaw), and lateral head inclinations (i.e. roll). Similarly to fiducial points, head angles were first smoothed using the 4253H-twice method [27]. The three normalized time series of pitch, yaw, and roll were then used to measure head movement amplitudes. Head movement velocity and acceleration for pitch, yaw, and roll were computed as the derivative of the amplitude and velocity, respectively. Twenty one features were then used to measure the magnitude of variation of the amplitude, velocity, and acceleration of head movement (see Table 1). By concatenating the features extracted from the 3 time series components, 63-dimensional (21 features $\times$ 3 time series components) dynamic features were extracted.

3.2 Vocal Prosody

Previous research has shown that switching pause duration and vocal fundamental frequency (f_0) are discriminative features for depression severity assessment [9, 28]. We used both features for depression assessment.

Since audio was recorded in a clinical office setting rather than laboratory setting, some acoustic noise was unavoidable. An in-

termediate level of 40% noise reduction was used to achieve the desired signal-to-noise ratio without distorting the original signal. Each pair of recordings was transcribed manually using Transcriber software [6], force-aligned using CMU Sphinx III [7], and post-processed using Praat [5]. Because session recordings exceeded the memory limits of Sphinx, it was necessary to segment recordings prior to forced alignment. We segmented recordings at transcription boundaries; that is, whenever a change in speaker occurred, resulting in speaker-specific segments. Forced alignment produced a matrix of four columns: speaker (which encoded both individual and simultaneous speech), start time, stop time, and utterance (see [28] for details). The forced alignment timings were used to identify speaker-turns and speaker diarization for the subsequent automatic feature extraction.

3.2.1 Switching Pause Duration

Switching pause, or latency to speak, is defined as the pause duration between the end of one speaker's utterance and the start of an utterance by the other. Switching pauses were identified from the matrix output of Sphinx. So that back channel utterances would not confound switching pauses, overlapping voiced frames were excluded. Switching-pauses were aggregated to yield mean duration and coefficient of variation (CV) for both participants and interviewers. The CV is the ratio of standard deviation to the mean. It reflects the variability of switching pauses when the effect of mean differences in duration is removed. In order to characterize the participants latency to speak, mean, variance, and variation coefficient of switching pause durations were computed.

3.2.2 Fundamental Frequency (f_0)

For each participant's utterance, f_0 was computed automatically using the autocorrelation function in Praat [5] with a window shift of 10 ms. Since microphones were not calibrated for intensity, intensity measures were not considered. To measure dynamic changes in f_0, mean amplitude, variation coefficient of amplitude, mean speed, and mean acceleration of f_0 were extracted.

3.3 Feature Selection and Classification

Min-Redundancy Max-Relevance (mRMR) algorithm [24] was used to select the most relevant features based on mutual information. Let S_{m-1} be the set of selected $m-1$ features, then the m^{th} feature can be selected from the set $\{F - S_{m-1}\}$ as:

$$\max_{f_j \in F - S_{m-1}} \left[I(f_j, c) - \frac{1}{m-1} \sum_{f_i \in S_{m-1}} I(f_j, f_i) \right], \quad (1)$$

where I is the mutual information function and c is the target class. F and S denote the original feature set, and the selected sub set of features, respectively. Eq. 1 is used to determine which feature is selected at each iteration of the algorithm. The size of the selected feature set is determined based on the validation error.

Due to notable overlap in the feature space, density-based models would be an efficient choice for distinguishing between depression and remission. Logistic regression classifiers using leave-one participant-out cross-validation were employed for depression assessment from facial movement dynamics, head movement dynamics, and vocal prosody.

4. RESULTS

We seek to discriminate depression (HRSD <8) and remission (HRSD >14) using facial movement dynamics, head movement dynamics, and vocal prosody, separately and in combination. We use a two-level leave-one participant-out cross-validation scheme:

Table 2: Accuracy (%) of using the different modalities.

	Modality	Remission	Depression	Mean
Current Study	Facial Movement Dynamics	78.38	84.49	81.44
	Head Movement Dynamics	72.97	86.21	79.59
	Vocal Prosody	75.68	63.79	69.73
Previous Studies	Facial Movements [9]	75.68	81.03	78.35
	Head Mov. GMM [1]	74.97	65.07	70.02
	Head Mov. Functionals [1]	83.71	74.16	78.94
	Vocal Prosody [28]	56.76	79.31	68.03

Table 3: Accuracy (%) of different feature selection methods.

	Modality	Feature Selection		
		none	t-test based	mRMR
Current Study	Facial Movement Dynamics	71.83	73.56	**81.44**
	Head Movement Dynamics	72.09	68.03	**79.59**
	Vocal Prosody	**69.73**	68.52	**69.73**
Previous Studies	Facial Movements [9]	71.34	71.71	**78.35**
	Head Mov. GMM [1]	58.67	68.97	**70.02**
	Head Mov. Functionals [1]	70.97	77.89	**78.94**
	Vocal Prosody [28]	68.03	**68.52**	68.03

Each time a test fold is separated, a leave-one participant-out cross-validation is used to train the system, and parameters are optimized without using the test partition. Minimum classification error on a separate validation set is used to determine the most discriminative set of features. For the fusions, whole set of features are fused into one low-abstraction vector and feature selection is applied afterwards to optimize the informativeness of the feature combinations.

4.1 Assessment of Modalities

Accuracy differed between modalities ($F_{2,282} = 3.16$, p=0.04). Both facial movement dynamics and head movement dynamics performed better than vocal prosody for depression status (p=0.02, t =2.24, df=94 and p=0.03, t=2.24, df=94, respectively). No difference was found between facial movement dynamics and head movement dynamics (p=0.83, t=2.20, df=94). Overall, visual information performed better for depression status than vocal prosody.

To further assess the quality of the feature sets, we compared the obtained results with recently proposed methods using facial movement [9], head movement [1], and prosodic features [28] for depression detection. For a fair comparison, we evaluated all methods using the same feature selection algorithm (i.e., mRMR), and the same classification procedure (i.e., logistic regression).

As shown in Table 2, the proposed features outperformed their counterparts for each modality. The accuracy of the proposed facial movement dynamics is approximately 3% higher than that of facial movement features in [9]. Proposed head movement dynamics performed better than representing head movements by Gaussian Mixture Model (GMM) parameters or functionals as in [1]. A small increase (1.7%) was obtained with the proposed prosodic features compared to their counterpart in [28].

4.2 Assessment of Feature Selection

To evaluate the reliability and effectiveness of mRMR feature selection, we compare mRMR results to the t-test threshold based feature selection method proposed in [1], and without any feature selection step (see Table 3). Because the number of samples per class was unbalanced, average accuracies for remission and depression are reported instead of total correct classification rate.

Feature selection using mRMR performed better for all but voice features in [28]. The t-test based method provided a 0.5% accuracy improvement over the use of raw features. This finding may be explained by the carefully defined feature sets with a limited dimen-

Table 4: Accuracy (%) of fusing different modalities.

Modality	Remission	Depression	Mean
Face + Head	81.08	89.66	85.37
Face + Voice	78.38	87.93	83.15
Head + Voice	83.78	81.03	82.41
Head + Face + Voice	**86.49**	**91.38**	**88.93**

Note: Face and Head refer to the facial movement dynamics and head movement dynamics, respectively (see Table 1 and Section 3.1). Voice refers to the prosodic features (see Section 3.2).

sionality (4 vocal features in [28], 7 vocal features in the current study). Overall, the results suggest the value of maximizing the combined mutual information of individual features.

4.3 Multimodal Fusion

Because individual features may provide additional discrimination power, we concatenated features of different modalities prior to feature selection. As shown in Table 4, the combination of facial and head movement dynamics performed best, followed by facial movement dynamics and vocal prosody. Fusing of all modalities gave best result. Interestingly, correct classification patterns for each fusion combination were very different.

5. CONCLUSIONS

We proposed an automatic, multimodal assessment of depression using facial, postural, and vocal behavioral measures in participants undergoing treatment for depression. We systematically investigated the discriminative power of facial movement dynamics, head movement dynamics, and vocal prosody, individually and in combination. Face and head movement dynamics outperformed vocal prosody. Best performances were obtained when all modalities were combined.

Acknowledgements

Supported in part by NIH award 51435 to U. Pittsburgh.

6. REFERENCES

[1] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear. Head pose and movement analysis as an indicator of depression. ACII, pages 283–288, 2013.

[2] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear. Eye movement analysis for depression detection. ICIP, pages 4220–4224, 2013.

[3] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. Washington, DC, 1994.

[4] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. Washington, DC, 2013.

[5] P. Boersma and D. Weenink. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10):341-345, 2001.

[6] K. Boudahmane, M. Manta, F. Antoine, S. Galliano, and C. Barras. TranscriberAG, 2011.

[7] CMU Sphinx: Open source toolkit for speech recognition. `http://cmusphinx.sourceforge.net/`.

[8] J.F. Cohn and K.L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *Int. Journal of Wavelets, Multires. and Information Processing*, 2(2):121–132, 2004.

[9] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Yang, M.H. Nguyen, M.T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. ACII, 2009.

[10] N. Cummins, S. Scherer, J. Krajewskid, S. Schniederd, J. Eppsa, and T.F. Quatierif. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[11] H. Dibeklioğlu, A.A. Salah, and T. Gevers, Are you really smiling at me? Spontaneous versus posed enjoyment smiles. ECCV, pages 525–538, 2012.

[12] H. Dibeklioğlu, A.A. Salah, and T. Gevers. Recognition of genuine smiles. *IEEE Transactions on Multimedia*, 17(3):279–294, 2015.

[13] P. Ekman. *Telling lies*. New York, NY: Norton, 2009.

[14] M.B. First, R.L. Spitzer, M. Gibbon, and J. B.W. Williams. *Structured clinical interview for DSM-IV axis I disorders: Patient Edition* (SCID-I/P, version 2.0). Biometrics Research Department, New York State Psychiatric Institute, New York, 1995.

[15] J.C. Fournier, R.J. DeRubeis, S.D. Hollon, S. Dimidjian, J.D. Amsterdam, R.C. Shelton, and J. Fawcett. Antidepressant drug effects and depression severity: A patient-level meta-analysis. *JAMA*, 303(1):47–53, 2010.

[16] A.J. Fridlund. The behavioral ecology and sociality of human faces. In M.S. Clark (Ed.), *Emotion*, pages 90–121, Sage Publications, 1992.

[17] M. Hamilton. A rating scale for depression. *Journal of Neurology and Neurosurgery*, 23:56–61, 1960.

[18] S.D. Hollon, M.E. Thase, and J.C. Markowitz. Treatment and prevention of depression. *Psychological Science in the Public Interest*, 3(2):38–77, 2002.

[19] L.A. Jeni, J.F. Cohn, T. Kanade. Dense 3D face alignment from 2D videos in real-time. AFGR, 2015.

[20] J. Joshi, R. Goecke, M. Breakspear, and G. Parker. Can body expressions contribute to automatic depression analysis? AFGR, 2013.

[21] J.P. Lepine, and M. Briley. The increasing burden of depression. *Neuropsychiatric Disease and Treatment*, 7(Suppl 1):3, 2011.

[22] G.M. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou. Towards an affective interface for assessment of psychological distress. ACII, 2015.

[23] C.D. Mathers, D. Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine*, 3(11) e442, 2006.

[24] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on PAMI*, 27(8):1226–1238, 2005.

[25] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences. ACII, pages 147–152, 2013.

[26] M.F. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S.Schnieder, R. Cowie, and M. Pantic. Int. Workshop on AVEC, pages 3–10, 2013.

[27] P.F. Velleman. Definition and comparison of robust nonlinear data smoothing algorithms. *J. Amer. Statist. Assoc.*, 75(371):609–615, 1980.

[28] Y. Yang, C. Fairbairn, and J.F. Cohn. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 4(2):142-150, 2013.