Published in final edited form as:

IEEE J Biomed Health Inform. 2018 March ; 22(2): 525–536. doi:10.1109/JBHI.2017.2676878.

Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding

Hamdi Dibeklioğlu* **[Member IEEE]**,
Pattern Recognition and Bioinformatics Group, Delft University of Technology, The Netherlands

Zakia Hammal* **[Member IEEE]**, and
Robotics Institute, Carnegie Mellon University, USA

Jeffrey F. Cohn [Member IEEE]
Department of Psychology, University of Pittsburgh, USA, and also with the Robotics Institute, Carnegie Mellon University, USA

Abstract

Depression is one of the most common psychiatric disorders worldwide, with over 350 million people affected. Current methods to screen for and assess depression depend almost entirely on clinical interviews and self-report scales. While useful, such measures lack objective, systematic, and efficient ways of incorporating behavioral observations that are strong indicators of depression presence and severity. Using dynamics of facial and head movement and vocalization, we trained classifiers to detect three levels of depression severity. Participants were a community sample diagnosed with major depressive disorder. They were recorded in clinical interview (Hamilton Rating Scale for Depression, HRSD) at 7-week intervals over a period of 21 weeks. At each interview, they were scored by HRSD as moderately to severely depressed, mildly depressed, or remitted. Logistic regression classifiers using leave-one-participant-out validation were compared for facial movement, head movement, and vocal prosody individually and in combination. Accuracy of depression severity measurement from facial movement dynamics was higher than that for head movement dynamics; and each was substantially higher than that for vocal prosody. Accuracy using all three modalities combined only marginally exceeded that of face and head combined. These findings suggest that automatic detection of depression severity from behavioral indicators in patients is feasible and that multimodal measures afford most powerful detection.

Index Terms

Depression severity; Facial movement dynamics; Head movement dynamics; Vocal prosody; Multimodal fusion

I. Introduction

Depression is one of the most common psychological disorders and a leading cause of disease burden [1]. Nearly 14.8 million people in the United States suffer from depression.

*Co-first authorship: The first and second listed authors contributed equally.

The social and personal costs of major depression and related unipolar disorders are substantial. Depression increases the risk of suicide some 20-fold. Economic losses approach 40 billion dollars per year. The World Health Organization predicts that depression will become the leading cause of disease burden (mortality plus morbidity) within the next 15 years [2]. Reliable, objective, and efficient screening and measurement of depression severity are critical to identify individuals in need of treatment and to evaluate treatment response.

Many symptoms of depression are observable. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [3], the standard for psychiatric diagnosis in the U.S., describes a range of audiovisual depression 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, decreased vocal intensity, and neuromotor disturbances [3], [4], [5]. Yet, often these indicators are not taken into account in screening, diagnosis, and evaluation of treatment response. Depression assessment relies almost entirely on patients' verbally reported symptoms in clinical interviews (e.g., the clinician-administered Hamilton Rating Scale for Depression [6]) and self-report questionnaires (e.g., Beck Depression Inventory [7]). These instruments, while useful, fail to include visual and auditory indicators that are powerful indices of depression. Recent advances in computer vision and signal processing for automatic analysis and modeling of human behavior could play a vital role in overcoming this limitation.

A. Related Work

Psychomotor symptoms such as gross motor activity, facial expressiveness, body movements, and speech timing differ between depressed and normal comparison groups [3], [4], [5]. Consequently, an automatic and objective assessment of depression from behavioral signals is of increasing interest to clinical and computer scientists. The latter use signal processing, computer vision, and pattern recognition methodologies. From the computer-science perspective, research has sought to identify depression from vocal utterances [8], [9], [10], [11], [12], [13], facial expression [14], [15], [16], [17], head movements/pose [18], [16], [19], body movements [18], and gaze [20]. While most research is limited to a single modality, there is increasing interest in multimodal approaches to depression detection [21], [22].

A challenge for automatic measurement of depression severity is the lack of available, suitable audio-video archives of behavioral observations of individuals that have clinically relevant depression. Well-labeled unscripted audio-visual recordings of clinically relevant variation in depression severity are necessary to train classifiers. Because the confidentiality of patient data must be protected, clinical databases such as the one used in this paper are not generally available.

One option so far was to recruit participants that have a range of depressive symptoms without regard to whether they meet DSM-5 diagnostic criteria. The recent Audio/Visual Emotion Challenge (AVEC) is a leading example. AVEC explored automatic measurement of the behavior of non-clinical participants partaking in an individual Human-Computer Interaction (HCI) task. The objective of the challenge was to automatically predict the level

of participants' self-reported depression on the Beck Depression Inventory-II (BDI) [21]. AVEC provided common data for multiple research groups to analyze and compare their results.

The AVEC depression database is composed of audio-video recordings of 300 participants with a wide range of BDI scores. For each recording, the database includes self-reported BDI, spatiotemporal Local Gabor Binary Patterns (LGBP-TOP) video features, and a large set of voice features (a set of low level voice descriptors and functionals extracted using the freely open-source openEAR [23] and openSMILE [24]).

Using AVEC and a few non-publicly available resources [25], audiovisual detection of depression has been proposed [26], [27] [28], [29], [30], [31], [32], [33]. In [28] for instance, visual bag-of-words (BoW) features computed from space time interest points (STIP), were combined with melfrequency cepstral coefficients (MFCCs) features. Extracted audiovisual features were then fused at the feature level and modeled using support vector regressors (SVRs) to measure self-reported BDI [28].

Using a similar approach, Joshi and colleagues [27], combined STIPs and MFCCs with other visual features (e.g., Spatiotemporal Local Binary Patterns (LBP-TOP) and audio features (such as fundamental frequency, loudness, and intensity). BoW features were then learned for each of the extracted audiovisual feature sets using SVMs. Feature and decision level fusion strategies were compared for the automatic audiovisual detection of depression.

In related work, Jain and colleagues [29], combined visual LBP-TOP features with Dense Trajectories and low level audio descriptors provided in [21]. The extracted audiovisual features were encoded using a Fisher Vector representation and a linear SVR was used to learn BDI score classification. In [31], visual Motion History Histogram (MHH) features were measured from three different visual texture features (Local Binary Patterns, Edge Orientation Histogram, and Local Phase Quantization) and combined with low-level audio descriptors provided in [21]. Partial Least Square (PLS) and Linear regression algorithms were used to model the mapping between the extracted features and BDI scores for face and voice features separately, followed by a decision based combination. In [32], the authors combined two regression models using LGBP-TOP video features with another single regression model based on acoustic i-vectors to compute a final BDI score. In another contribution, [26] combined temporal patterns of head pose and eye/eyelid movements. A hybrid fusion method of the scores obtained from individual modalities and their combination was used to detect presence from absence of depression.

In all but a few cases, such as [34], previous efforts have relied on high dimensional audiovisual descriptors to detect self-reported depression severity. In contrast, using AVEC data, Williamson and his colleagues [34] investigated the specific changes in coordination, movement, and timing of facial and vocal signals as potential symptoms for self reported BDI scores. They proposed a multi-scale correlation structure and timing feature sets from video-based facial action units (AUs) and audio-based vocal features. They combined the extracted complementary features using a Gaussian mixture model and extreme learning machine classifiers to predict BDI scores.

Despite increasing efforts, the current state of the art has not yet achieved the goal of automatic, reliable, and objective measurement of depression severity from behavioral indicators of affected individuals. Multimodal measurement of depression severity raises several issues. Many of these are shared with other applications of automatic and multimodal human behavior analysis.

- One is whether one or another modality is more informative. Ekman [35] for instance proposed that for affect recognition, facial expression is more revealing than body; he was equivocal about face relative to voice. Alternatively, one could imagine that high redundancy across channels would render modest any potential gain provided by a multimodal approach for depression severity measurement. Comparative studies are needed to explore this issue.
- A second issue is choice of context. AVEC explored audiovisual expression in the context of an individual human-machine interaction task, for which audience effects would likely be absent. However, research by Fridlund and others [36] suggests that when an audience is present, signal strength of nonverbal behavior increases. Nonverbal reactions to and from others present additional sources of information. For instance, switching-pause or turn-taking latency can only be measured in social interaction. Context influences what behaviors occur and their intensity.

B. Proposed Contribution

Reduced reactivity is consistent with many evolutionary theories of depression and highlights the symptoms of psychomotor retardation [37], [38], [39]. To capture aspects of psychomotor retardation and agitation in clinically relevant participants, we used dynamic measures of expressive behavior.

- In contrast to previous work, all participants met DSM-4 or DSM-5 criteria for major depression as determined by diagnostic interview. Diagnostic criteria matter for at least two reasons. First, many non-depressive disorders are confusable with depression. Post-traumatic stress disorder (PTSD) and generalized anxiety disorder, for instance, share overlapping symptoms with depression. Second, people with history of depression may differ from those without depression in personality factors or in other non-specific ways [40]. By using diagnostic criteria and focusing on change in depression severity, we were able to rule out other sources of influence.
- Compared with previous efforts, we focused on an interpersonal context, clinical interviews. Informed by the psychology literature on depression, we anticipated that the interpersonal nature of clinical interviews would heighten discriminability across modalities.
- We investigated the discriminative power of three modalities – facial movement dynamics, head movement dynamics, and vocal prosody – individually and in combination to measure depression severity. Symptom severity was ground-truthed using state-of-the-art depression severity interviews.

- Instead of using a large number of descriptors or selecting informative features individually, we investigated the selection of an optimum feature set by maximizing the combined mutual information for depression severity measurement.

Part of this work has been presented at the ACM International Conference on Multimodal Interaction [41].

II. Materials

A. 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 (mean = 39.65) and were Euro- or African-American (46 and 11, respectively). At the time of the study, all met DSM-4 criteria [42] for Major Depressive Disorder (MDD). DSM-4 (since updated to DSM-5) is the standard in the U.S. and much of the world. Although not a focus of this study, participants were randomized to either anti-depressant treatment with a selective serotonin reuptake inhibitor (SSRI) or Interpersonal Psychotherapy (IPT). Both treatments are empirically validated for the treatment of depression [43]. Data from 49 participants were available for analysis. Participant loss was due to change in original diagnosis, severe suicidal ideation, and methodological reasons (e.g., missing audio or video).

B. Observational Procedures

Symptom severity was evaluated on up to four occasions at 1, 7, 13, and 21 weeks post diagnosis and intake by ten clinical interviewers (all female). Interviewers were not assigned to specific participants. Four interviewers were responsible for the bulk of the interviews but the number of interviews per interviewer varied. The median number of interviews per interviewer was 14.5; four conducted six or fewer.

Interviews were conducted using the Hamilton Rating Scale for Depression (HRSD) [6]. HRSD is a clinician-rated multiple item questionnaire to measure depression severity and response to treatment. The HRSD rates the severity of depression by probing mood, feelings of guilt, suicide ideation, insomnia, agitation or retardation, anxiety, weight loss, and somatic symptoms. Each item is scored on a 3- or 5-point Likert type scale, depending on the item, and the total score is compared to the corresponding descriptor. Interviewers were expert in the HRSD and reliability was maintained above 0.90. Variation in HRSD scores is used as a guide to evaluate recovery by detecting ordinal ranges of depression severity. HRSD scores of 15 or higher are generally considered to indicate moderate to severe depression; scores between 8 and 14 indicate mild depression; and scores of 7 or lower indicate remission [44].

Interviews were recorded using three hardware-synchronized analogue cameras and two unidirectional microphones (see Fig. 1). Two cameras were positioned approximately 15° to the participant's left and right. One camera recorded the participant's face and one camera recorded a full body view (see Fig. 1 left). A third camera recorded the interviewer's shoulders and face from approximately 15° to the interviewer's right (see Fig. 1 right).

Audio-visual data from the camera and microphone to the participant's right were used in this study.

Missing data occurred due to missed appointments or technical problems. Technical problems included failure to record audio or video, occurrence of audio or video artifacts, and insufficient data. The distribution of the data (i.e., number and mean duration of sessions per HRSD score) from the beginning of the first question to the end of the interview is reported in Fig. 2. Videos were digitized into a resolution of 640×480 pixels at a rate of 29.97 frames per second (see section III-A.) Audio was digitized at 48 kHz and later down-sampled to 16 kHz for speech processing. To be included for audio analysis, we required a minimum of 20 speaker turns of at least 3 seconds in length and at least 50 seconds of vocalization in total (see section III-B) [11]. Using these data and the cut-off scores described above, we defined three ordinal depression severity classes: moderate to severe depression, mild depression, and remission (i.e., recovery from depression). The final sample was 130 sessions from 49 participants: 58 moderate to severely depressed, 35 mildly depressed, and 37 remitted.

III. Audiovisual Feature Extraction

Depression alters the timing of nonverbal behavior [3]. To capture changes in visual and auditory modalities, we focused on the dynamics of facial and head movement, and vocal fundamental frequency and switching pauses. We thus include both visual and auditory measures.

A. Visual Measures

1) Automatic Tracking of Facial Landmarks and Head Pose—Previous research has used person-dependent Active Appearance Models (AAMs) to track the face and facial features (e.g., [14], [19]). Because AAMs must be pre-trained for each participant, they are not well suited for clinical applications or large numbers of participants. We used a fully automatic, person-independent, generic approach that is comparable to AAMs to track the face and facial features, referred to as ZFace [45]. The robustness of this method for 3D registration and reconstruction from 2D video has been validated in a series of experiments (for details, see [45], [46]).

ZFace performs 3D registration from 2D video with no pre-training. This is done using a combined 3D supervised descent method [47], where the shape model is defined by a 3D mesh and the 3D vertex locations of the mesh [45]. ZFace registers a dense parameterized shape model to an image such that its landmarks correspond to consistent locations on the face. We used ZFace to track 49 facial landmarks (fiducial points) and 3 degrees of out-of-plane rigid head movements (i.e., pitch, yaw, and roll) from 2D videos (see Fig. 3).

2) Preprocessing of the Extracted Facial Landmarks and Head Pose—Most previous work in affect analysis uses holistic facial expressions, action units, or valence. Because our interest is the dynamics rather than the configuration of facial expression, we used only facial and head movement dynamics.

Facial movement dynamics was represented using the time series of the coordinates of the 49 tracked fiducial points. To control for variation due to rigid head movement, fiducial points were first normalized by removing translation, rotation, and scale. To reduce tracking errors that could happen, the movement of the normalized 98 time series (49 fiducial points × x and y coordinates) was smoothed by the 4253H-twice method [48] and used to measure the dynamics of facial movement between clinical interviews.

Likewise, head movement dynamics was represented using the time series of the 3 degrees of freedom of out-of-plane rigid head movement. These movements correspond to head nods (i.e., pitch), head turns (i.e., yaw), and lateral head inclinations (i.e., roll). Similar to fiducial points, head angles were smoothed using the 4253H-twice method [48] prior to analysis.

3) Per-Frame Encoding of Facial and Head Movement—Our goal is to automatically estimate depression severity scores from the moment-to-moment changes (i.e., per frame changes) of the smoothed measures of head and facial movement. To achieve this goal, we faced three main challenges: (1) The movements of individual fiducial points and head pose orientations are highly correlated. (2) Both facial and head movement measures include redundant information and complex relations that cannot be revealed by linear methods, such as the conventional Principle Component Analysis (PCA) and Canonical Correlation Analysis [49]. (3) Only a single label (i.e., Remission, Mild, and Moderate to Severe) is available for each session. Per-frame class labels are not available. To meet these challenges, we used deep learning based methods [50], [51] and, in particular, Stacked Denoising Autoencoders (SDAE) [52]. SDAE has emerged as one of the most successful unsupervised methods to discover unknown non-linear mappings between features (which in our case are the face and head movement dynamics) and outcomes (which in our case are depression severity scores) while coping with high dimensionality and redundancy.

SDAE is a deep network based on stacking layers of denoising autoencoders. They are locally trained to learn representations that are insensitive to small irrelevant changes in the inputs (see Fig. 4.b). Each hidden layer (i.e., denoising autoencoder) of the resulting deep network learns efficient representations of the corresponding inputs. In order to force the hidden layer to discover robust features instead of simply learning the input's identity, the autoencoder is trained to reconstruct the input from a “corrupted version” of the input [53], [52]. A stochastic corruption process randomly sets some of the elements of the input $\mathbf{x}$ to 0 resulting in a corrupted version $\tilde{\mathbf{x}}$ [53], [52]. The corruption is only used for the training process. A different corrupted version of $\mathbf{x}$ is generated each time the training example $\mathbf{x}$ is presented. Each hidden layer is then learned using a denoising autoencoder, which maps a corrupted version $\tilde{\mathbf{x}}$ of input $\mathbf{x} \in \mathbb{R}^p$ to a latent representation $\mathbf{y} \in \mathbb{R}^q$, and then maps it back to the original space $\mathbf{z} \in \mathbb{R}^p$, where p and q denote the sizes of the input $\mathbf{x}$ and the latent representation $\mathbf{y}$, respectively. The denoising autoencoder is trained by minimizing the reconstruction error $\|\mathbf{x} - \mathbf{z}\|^2$.

The first hidden layer of the SDAE is trained to reconstruct the input data, and the following hidden layers are trained to reconstruct the states of the layers below, respectively (see Fig. 4.b). Transformation weights are initialized at random and then optimized by gradient descent. Once a stack of encoders has been built, the entire deep autoencoder is then trained

to fine-tune all the parameters together to obtain an optimal reconstruction using gradient-based backpropagation [51].

In the current contribution, we used a separate 3-layer SDAE deep network architecture (i.e., SDAE with 3 hidden layers) to encode efficient per-frame representations of both facial movement and head movement (see Fig. 4.b). Each SDAE was trained using the normalized and smoothed 49 facial landmark coordinates and smoothed 3 head pose orientations, respectively (see section III-A2 and Fig. 4.a). For each SDAE deep network architecture (face and head), the number of units per each hidden layer, and other hyperparameters of SDAE were determined during training by minimizing the prediction error (the difference between estimated and actual depression severity scores). The list of the investigated hyperparameters is given in Table I. For a compact representation of facial features, the number of units of the 3^{rd} hidden layer (i.e., dimension of the learned facial representation) of the corresponding SDAE deep network was empirically set to 15. The number of units of the 3^{rd} hidden layer for the head SDAE (see Fig. 4.b) was automatically set to 5 for most folds of the cross-validation (see section V).

After training, each SDAE deep network learned a transformation of the extracted per-frame features to an effective representation. By applying the learned per-frame encoding to each frame of the video, the SDAE-based outputs were combined into an effective $n \times d_{\text{final}}$ time series representation $\mathcal{D}$ (n being the number of frames of a given video and d_{final} the dimension of the learned features) describing the video.

The obtained SDAE-based time series $\mathcal{D} \in \mathrm{R}^{n \times d_{\text{final}}}$ encoded how the SDAE-based representation of input elements (facial landmark coordinates and head angles, respectively) changed over time across the video (see Fig. 4.c). Considering that each column $\mathcal{D}_i$ ($i \in \{1, 2, \ldots, d_{\text{final}}\}$) of data matrix $\mathcal{D}$ corresponds to a specific movement, we computed the dynamic changes of these movements over time. The velocity of change of the extracted d_{final} time series was then computed as the derivative of the corresponding values as $\mathcal{V}_i = \frac{d\mathcal{D}_i}{dt}$, measuring the velocity of change of the per-frame facial (or head pose) features from one frame to the next (see Fig. 4.c). Similarly, the acceleration of change of the perframe facial (or head pose) features was computed as the derivative of the corresponding velocities as $\mathcal{A}_i = \frac{d^2\mathcal{D}_i}{dt^2}$. For the purpose of alignment of the three time series ($\mathcal{D}$, $\mathcal{V}$, and $\mathcal{A}$), the first two frames of videos were discarded from all analyses. For simplicity, $\mathcal{D}$, $\mathcal{V}$, and $\mathcal{A}$ will hereafter be referred to as amplitude, velocity, and acceleration, respectively (see Fig. 4.c).

4) Per-Video Encoding of Facial and Head Movement—Because videos of interviews varied in length, the extracted time series features of different videos varied in length. It was thus useful to encode the extracted time series descriptors with fixed length per-video descriptors (see Fig. 4.d). To this end, we used two different representations to describe the videos with a fixed length representation: 1) Improved Fisher Vector (IFV) coding [54], and 2) Compact Dynamic Feature Set (DFS) [41].

In the IFV-based representation, amplitude, velocity, and acceleration measures were first concatenated for each frame. Using a Gaussian mixture model (GMM) with 64 Gaussian distributions, these combined measurements were encoded into a $384 \times d_{\text{final}}$ ($64 \times 2 \times 3 \times d_{\text{final}}$) dimensional IFV for each video. The resulting feature vectors were then normalized by power normalization [54] and l^2-norm.

The compact DFS-based representation corresponds to 21 features extracted for each of the elements of the d_{final} time series as described in Table II, yielding $21 \times d_{\text{final}}$ dimensional descriptor per video. Based on previous research [55], DFS comprises key measurements of amplitude, velocity, and acceleration, as well as taking into account the direction of the change in the extracted time series. This is done by dividing each time series into increasing ($^+$) and decreasing ($^-$) segments (see Fig. 5 and Table II). In Table II, signals symbolized with superindex ($^+$) and ($^-$) denote the segments of the related signal with continuous increase and continuous decrease in amplitude, respectively. η defines the length (number of frames) of a given time series. τ^+ and τ^- denote the number of increasing and decreasing amplitude segments in the time series sequence, respectively. In some cases, the features cannot be calculated; for instance, if we extract features from a continuously increasing times series, no decreasing segment can be detected [$\eta(\mathcal{D}^-) = 0$]. In such conditions, all the features describing the related segments are set to zero.

Using IFV coding and DFS, 5760 (384×15) dimensional IFV and 315 (21×15) dimensional DFS representations were obtained as per-video facial features. Similarly, $384 \times d_{\text{final}}$ dimensional IFV and $21 \times d_{\text{final}}$ dimensional DFS representations were obtained as per-video head features.

B. Vocal Measures

1) Preprocessing—Because audio was recorded in a clinical office setting rather than laboratory setting, some acoustic noise was unavoidable. To reduce noise level and equalize intensity, Adobe Audition II [11] was used. An intermediate level of 40% noise reduction was used to achieve the desired signal-to-noise ratio without distorting the original signal [11]. Each pair of recordings was transcribed manually using Transcriber software [56], then force-aligned using CMU Sphinx III [57], and post-processed using Praat [58]. Because session recordings exceeded the memory limits of Sphinx, it was necessary to segment recordings prior to forced alignment. While several approaches to segmentation were possible, we segmented recordings at transcription boundaries; that is, whenever a change in speaker occurred [11]. Except for occasional overlapping speech, this approach resulted 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. To assess the reliability of the forced alignment, audio files from 30 sessions were manually aligned and compared with the segmentation yielded by Sphinx [11]. Mean error(s) for onset and offset were 0.097 and 0.010 seconds for participants, respectively. The forced alignment timings were used to identify speaker-turns and speaker diarization for the subsequent automatic feature extraction [11].

2) Vocal Features—Previous investigations have revealed that compared to non-depressed participants, depressed participants presented reduced speech variability and monotonicity in loudness and pitch [59], [60], [61], [62], reduced speech [63], reduced articulation rate [64], and increased pause duration [11], [65]. Consistent with alternative methods, and because we were interested in severity assessment and not in diagnostic, in preliminary work we investigated a number of possible vocal features for the measurement of depression severity. We considered both frequency and timing features such as fundamental frequency (f_0), Maxima Dispersion Quotient (MDQ), Peak Slope (PS), Normalized Amplitude Quotient (NAQ), Quasi Open Quotient (QOQ), and switching pause durations [11], [12]. However, preliminary results showed that only switching pause durations and f_0 were correlated with depression severity [11], [12]. For this reason, we used only these measures.

Switching pause (SP), or latency to speak, is defined as the pause duration between the end of one speaker's utterance and the start of the other speaker's utterance. SPs were identified from the matrix output of Sphinx [11]. So that back channel utterances would not confound SPs, overlapping voiced frames were excluded. SPs 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 [11]. It reflects the variability of SPs when the effect of mean differences in duration is removed. To characterize the participants' latency to speak, mean, variance, and CV of SP durations were computed over the whole session and used for automatic measurement of depression severity.

Vocal fundamental frequency (f_0) for each utterance was computed automatically using the autocorrelation function in Praat [58] with a window shift of 10 ms [11]. To measure dynamic changes in the fundamental frequency, mean amplitude, variation coefficient of amplitude, mean speed, and mean acceleration of f_0 over the whole session were extracted and used for automatic measurement of depression severity. Since microphones were not calibrated for intensity, intensity measures were not considered.

IV. Feature Selection and Depression Severity Estimation

To reduce redundancy and select the most discriminative audiovisual feature set, the Min-Redundancy Max-Relevance (mRMR) algorithm [66] was used for feature selection. Compared to the closely related Canonical Correlation Analysis (CCA) based feature selection, which optimizes the mutual correlation between labels and feature set, mRMR is an incremental method for minimizing redundancy while selecting the most relevant features based on mutual information. The efficiency of mRMR to select the best set of individual features by maximizing the combined mutual information was established in previous research (e.g., [67], [68], [69]). We used it for the first time for audiovisual depression severity measurement.

More specifically, 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:

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

where I is the mutual information function and c is a 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 (probabilistic) models would be an efficient choice for distinguishing between depression severity scores. Given this consideration, logistic regression classifiers using leave-one-participant-out cross-validation were employed for depression severity measurement from facial movement dynamics, head movement dynamics, and vocal prosody, separately and in combination. Each regression model describes the distribution over a class y as a function of features ϕ as follows:

$$p(y|\phi) = \frac{1}{1+\exp(-y\, w^T \phi)} \quad (2)$$

The model can be trained by maximizing the log-likelihood of the training data under the model with respect to the model parameters w. Then, unseen features can be classified by maximizing the above equation over the trained classes (see section V).

V. Experimental Results

We seek to discriminate three levels of depression severity (moderate-to-severe, mild, and remitted) from facial movement dynamics, head movement dynamics, and vocal prosody separately and in combination. To do so, we used a two-level leave-one-participant-out cross-validation scheme. For each iteration a test fold was first separated. A leave-one-participant-out cross-validation was then used to train the whole system (i.e., SDAE and logistic regression) and optimize the corresponding hyperparameters without using the test partition. The optimized parameters included the regularization hyperparameter of the logistic regression classifier, number of features selected by mRMR, and the hyperparameters of the SDAE (i.e., number of units per hidden layer, fixed learning rate, number of pre-training epochs, and corruption noise level). The optimized parameters were then used to measure the classification error on the test set. This process was repeated for all participants.

For the fusion of modalities, whole sets of features were first combined into one low-abstraction vector; feature selection was then applied to optimize the informativeness of the feature combinations. Thus, each modality could effectively contribute to the selected set of features even though the numbers of features (whole set) of facial, head pose, and vocal modalities were different.

Performance was quantified two ways. One was the mean accuracy over the three levels of severity. The other was weighted kappa [70]. Weighted kappa is the proportion of ordinal agreement above what would be expected to occur by chance [70].

A. Assessment of Visual Features

We investigated the discriminative power of the proposed per-frame features (using the stacked denoising autoencoders, see section III-A3) and per-video based features (using IFV and DFS, see section III-A4) to measure depression severity from facial and head movement dynamics (see Table III). The feature selection step was included for both approaches. Raw data for head movement dynamics reported in Table III correspond to the time series of the 3 degrees of freedom of the smoothed rigid head movement (see section III-A2). Likewise, raw data for facial movement dynamics were represented using the time series of the registered and smoothed coordinates of the 49 tracked fiducial points (see section III-A2). For a compact representation, principal component analysis was used to reduce the 98 time series to 15 time series components that account for 95% of the variance (see Table III).

As shown in Table III, for both facial movement and head movement, the SDAE-based per-frame encoding together with IFV-based per-video encoding performed best. In all conditions, SDAE-based per-frame encoding achieved higher performance than than did raw features. Similarly, IFV-based per-video encoding performed best compared with the DFS-based per-video encoding (see Table III). Given these results, the SDAE-based per-frame encoding together with IFV-based per-video encoding was used in the remaining experiments.

B. Assessment of Modalities

Accuracy varied between modalities (Table IV). Facial movement dynamics and head movement dynamics performed significantly better than vocal prosody (28.15% higher, t=-4.15, df=387, $p \leq 0.001$ and 20.81% higher, t=-3.00, df=387, $p \leq 0.01$, respectively). Facial movement dynamics and head movement dynamics failed to differ from each other significantly (7.34% higher, t=-1.22, df=387, p>0.1). Overall, visual information performed better for depression severity than vocal prosody.

To further assess the quality of the proposed dynamic feature sets, we compared them with the dynamic features proposed in our earlier work [41] and with alternative dynamic features that include facial movement [14], head movement [19], or prosodic features [11]. For a fair and accurate comparison between the proposed dynamic features and alternative methods, it was necessary to re-implement the alternative methods for: (1) the more challenging problem of measurement of 3-levels of depression severity (as compared to 2-classes classification), and (2) evaluating them on our clinical data. Thus, we re-implemented previous methods as well as could be done from their description in the corresponding papers (including adapting our own work [41] to the problem of 3-levels of depression severity).

In our earlier work [41], facial and head movement dynamics were modeled using the per-frame raw features described in section V-A and the DFS-based per-video features. The extracted per-video features were fed to a logistic regression classifier. In [14], the AAMs

based fiducial point time series were fed into a PCA resulting to 10 time series components accounting for a total of 95% of variance. The velocity of movement of the extracted 10 time series was computed and segmented into contiguous 10s intervals. The mean, median, and standard deviation of velocities were computed for each interval. The extracted statistics were concatenated for each interview and used for final representation. In [19], head movements were tracked by AAMs and modeled by Gaussian mixture models with seven components. Mean, variance, and component weights of the learned GMMs were used as features. Additionally, a set of head pose functionals was proposed, such as the statistics of head movements and duration of looking in different directions. In [11], fundamental vocal frequency, and switching pause duration were shown to be informative for depression detection. Mean value and coefficient of variation were used for depression assessment. For a fair and accurate comparison, for all re-implemented dynamic features, we used the same feature selection algorithm (i.e., mRMR), the same classification procedure (i.e., logistic regression), and the same accuracy measures.

As shown in Table IV, the proposed features outperformed their counterparts for each modality. The accuracy of the proposed facial movement dynamics was 7.6% and 13.1% higher than that of facial movement dynamics in our earlier work [41], and facial movement features in [14], respectively. Likewise, the proposed head movement dynamics performed better than our earlier work [41] (9.19% higher) and better than Gaussian Mixture Model (GMM) in [19] (17.49% higher). A small increase (2%) was obtained with the proposed prosodic features compared to their counterpart in [11]. With the exception of our own previous method [41], we reimplemented previously published approaches. It is possible that had the original algorithms been used, results for alternative approaches may have been different. On the other hand, for comparison with our previous work [41], we had benefit of the original code. The new proposed features outperformed the previous one as shown in Table IV. To enable other researchers to compare their own algorithms directly with ours, we have arranged to release a version of the database (see section VII).

C. Assessment of Feature Selection

To evaluate the reliability and effectiveness of mRMR feature selection (see section IV), we compare mRMR results to no feature selection (see Table V).

As shown in Table V, feature selection using mRMR algorithm performed better than no feature selection in all but voice features. This finding may be explained by the carefully defined feature sets with a limited dimensionality (seven features were used for voice in the current paper). Overall, the results explicitly indicate the usefulness of maximizing the combined mutual information of individual features for depression severity measurement.

D. Multimodal Fusion

We evaluated the informativeness of the combination of different modalities for depression severity measurement (see Table VI). Because the combination of individual features may provide additional discrimination power, we concatenated features of different modalities prior to feature selection. Decision-level fusion strategies (i.e., SUM rule, PRODUCT rule,

and VOTING [71]) were also evaluated in our preliminary experiments, yet did not perform as accurate as feature-level fusion and were not included in the paper.

Results for multimodal fusion are presented in Table VI. Highest performance was achieved by fusion of all modalities (78.67%), followed by the combination of facial and head movement dynamics (77.77%), and then by facial movement dynamics and vocal prosody (73.16%). Lowest performance was achieved by fusion of head movement dynamics and vocal prosody (67.25%). Accuracy of facial movement dynamics and head movement dynamics together was significantly greater than head movement dynamics (12.52% higher, t=2.09, df=516, $p \leq 0.05$) but did not significantly differ from facial movement dynamics alone (5.18% higher, t=0.88, df=516, p>0.1). While the fusion of all modalities significantly improved the accuracy of individual use of head movement dynamics (13.42% higher, t=2.24, df=516, $p \leq 0.05$) and prosodic features (34.23% higher, t=5.04, df=516, $p \leq 0.001$), the performance improvement found over facial movement dynamics was moderate (6.08% higher, t=1.03, df=516, p>0.1). Combining the proposed prosodic features with facial and head movement dynamics increased accuracy only minimally. It is possible that had we considered additional measures of prosody, prosody might have contributed more to accuracy. We considered a large number of vocal features and selected for inclusion those that significantly correlated with depression severity. Further research will be needed to explore this issue.

In related work, we found that depression severity is associated with reduced head and lower body movement [72], [18] and reduced vowel space [73], [74]. These findings are consistent with observations in clinical psychology and psychiatry of psychomotor retardation in depression; that is, a slowing and attenuation of expressive behavior. Evolutionary perspectives on depression [75] similarly propose that depression in this way serves to decrease involvement with other persons. That is what we found in our previous work [72].

Motivated by these findings and theory, the current study extends previous efforts by integrating dynamic measures of face, head, and voice to measure depression severity. We found strong evidence that dynamic features reveal depression severity. A limitation of these findings is that while rooted in dynamic measures they are unable to reveal how dynamics change with respect to depression. Deep learning, which we used, while powerful learning tool, suffers from lack of explainability. Future work is needed to address this limitation.

This may be the first time that depression severity rather than only presence-absence of depression has been measured. From a clinical perspective, it is critical to measure change over time in course of depression and its treatment. In clinical trials for treatment of depression, response to treatment is quantified as 50% decrease in symptom severity. The ability to detect magnitude of change is subject of current work. Interventions can only be assessed when severity of symptoms is measured reliably.

VI. Conclusion

We proposed an automatic, multimodal approach to detect depression severity in participants undergoing treatment for depression. Deep learning based per-frame coding and pervideo

Fisher-vector based coding were used to characterize the dynamics of facial and head movement. Statistical criteria were used to select vocal features. For each modality, selection among features was performed using combined mutual information, which improved accuracy relative to blanket selection of all features regardless of their merit. For individual modalities, facial and head movement dynamics outperformed vocal prosody. For combinations, fusing the dynamics of facial and head movement was more discriminative than head movement dynamics and more discriminative than facial movement dynamics plus vocal prosody and head movement dynamics plus vocal prosody.

VII. Distribution of Clinical Depression Interviews

To promote research on automated measurement of depression severity and enable other researchers to compare algorithms directly with ours, de-identified features from the 130 clinical interviews are available for academic research use. For each video frame, the distribution includes summary vocal features (see section III-B), normalized 2D coordinates of the tracked 49 facial fiducial points (see Fig. 3), 3 degrees of freedom of head pose (see Fig. 3), and deep autoencoded frame-based representations of head and facial movement (see section III-A3). The data are available for non-commercial research from http://www.pitt.edu/~emotion/depression.htm.

Acknowledgments

Research reported in this publication was supported in part by the National Institute of Mental Health of the National Institutes of Health under Award Number MH096951. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Medicine. 2006; 3(11):e442. [PubMed: 17132052]
2. Lépine JP, Briley M. The increasing burden of depression. Neuropsychiatric Disease and Treatment. 2011; 7(Suppl 1):3–7. [PubMed: 21750622]
3. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. Washington, DC: 2013.
4. Sobin C, Sackeim HA. Psychomotor symptoms of depression. The American journal of psychiatry. 1997; 154(1):4. [PubMed: 8988952]
5. Caligiuri MP, Ellwanger J. Motor and cognitive aspects of motor retardation in depression. Journal of affective disorders. 2000; 57(1):83–93. [PubMed: 10708819]
6. Hamilton M. A rating scale for depression. Journal of neurology, neurosurgery, and psychiatry. 1960; 23(1):56–61.
7. Beck A, Ward C, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry. 1961; 562:53–63.
8. Cummins N, Epps J, Breakspear M, Goecke R. An investigation of depressed speech detection: Features and normalization. in Interspeech. 2011:2997–3000.
9. Trevino AC, Quatieri TF, Malyska N. Phonologically-based biomarkers for major depressive disorder. EURASIP Journal on Advances in Signal Processing. 2011; 2011(1):1–18.
10. Scherer S, Stratou G, Gratch J, Morency LP. Investigating voice quality as a speaker-independent indicator of depression and PTSD. Interspeech. 2013:847–851.
11. Yang Y, Fairbairn C, Cohn JF. Detecting depression severity from vocal prosody. IEEE Transactions on Affective Computing. 2013; 4(2):142–150. [PubMed: 26985326]

12. Scherer S, Hammal Z, Yang Y, Morency LP, Cohn JF. Dyadic behavior analysis in depression severity assessment interviews. in ACM International Conference on Multimodal Interaction. 2014:112–119.
13. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J. Analysis of acoustic space variability in speech affected by depression. Speech Communication. 2015; 75:27–49.
14. Cohn JF, Kruez TS, Matthews I, Yang Y, Nguyen MH, Padilla MT, Zhou F, La Torre FD. Detecting depression from facial actions and vocal prosody. International Conference on Affective Computing and Intelligent Interaction. 2009
15. Maddage NC, Senaratne R, Low LSA, Lech M, Allen N. Video-based detection of the clinical depression in adolescents. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2009:3723–3726.
16. 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.
17. Joshi J, Dhall A, Goecke R, Breakspear M, Parker G. Neural-net classification for spatio-temporal descriptor based depression analysis. International Conference on Pattern Recognition. 2012:2634–2638.
18. Joshi J, Goecke R, Parker G, Breakspear M. Can body expressions contribute to automatic depression analysis? IEEE International Conference on Automatic Face and Gesture Recognition. 2013
19. Alghowinem S, Goecke R, Wagner M, Parkerx G, Breakspear M. Head pose and movement analysis as an indicator of depression. Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013:283–288.
20. Alghowinem S, Goecke R, Wagner M, Parker G, Breakspear M. Eye movement analysis for depression detection. IEEE International Conference on Image Processing. 2013:4220–4224.
21. Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, Cowie R, Pantic M. AVEC 2014: 3D dimensional affect and depression recognition challenge. ACM International Workshop on Audio/Visual Emotion Challenge. 2014:3–10.
22. Cohn, JF., Cummins, N., Epps, J., Goecke, R., Joshi, J., Scherer, S. Multimodal assessment of depression and related disorders based on behavioral signals. In: Oviatt, S.Schuller, B.Cohen, P., Sonntag, D., editors. Handbook of Multi-Modal Multi-Sensor Interfaces. Morgan and Claypool; 2017.
23. Eyben F, Wöllmer M, Schuller B. Openear - Introducing the Munich open-source emotion and affect recognition toolkit. International Conference on Affective Computing and Intelligent Interaction. 2009
24. Eyben F, Wöllmer M, Schuller B. Opensmile: The Munich versatile and fast open-source audio feature extractor. ACM International Conference on Multimedia. 2010:1459–1462.
25. Black Dog Institute. [Online] Available: http://www.blackdoginstitute.org.au
26. Alghowinem S, Goecke R, Cohn JF, Wagner M, Parker G, Breakspear M. Cross-cultural detection of depression from nonverbal behaviour. IEEE International Conference on Automatic Face and Gesture Recognition. 2015
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(3):217–228.
28. Cummins N, Joshi J, Dhall A, Sethu V, Goecke R, Epps J. Diagnosis of depression by behavioural signals: A multimodal approach. ACM International Workshop on Audio/Visual Emotion Challenge. 2013:11–20.
29. Jain V, Crowley JL, Dey AK, Lux A. Depression estimation using audiovisual features and fisher vector encoding. ACM International Workshop on Audio/Visual Emotion Challenge. 2014:87–91.
30. Sidorov M, Minker W. Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach. ACM International Workshop on Audio/Visual Emotion Challenge. 2014:81–86.

31. Jan A, Meng H, Gaus YFA, Zhang F, Turabzadeh S. Automatic depression scale prediction using facial expression dynamics and regression. ACM International Workshop on Audio/Visual Emotion Challenge. 2014:73–80.
32. Senoussaoui M, Sarria-Paja M, Santos JF, Falk TH. Model fusion for multimodal depression classification and level detection. ACM International Workshop on Audio/Visual Emotion Challenge. 2014:57–63.
33. Meng H, Huang D, Wang H, Yang H, AI-Shuraifi M, Wang Y. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. ACM International Workshop on Audio/Visual Emotion Challenge. 2013:21–30.
34. Williamson JR, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD. Vocal and facial biomarkers of depression based on motor incoordination and timing. ACM International Workshop on Audio/Visual Emotion Challenge. 2014:65–72.
35. Ekman, P. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. New York, NY: WW Norton & Company; 2009.
36. Fridlund, AJ. The behavioral ecology and sociality of human faces. In: Clark, MS., editor. Emotion. Sage Publications; 1992. p. 90-121.
37. Klinger E. Consequences of commitment to and disengagement from incentives. Psychological Review. 1975; 82(1):1–25.
38. Fowles, DC. A motivational theory of psychopathology. In: Spaulding, WD., Simon, HA., editors. Integrative Views of Motivation, Cognition, and Emotion. University of Nebraska Press; 1994. p. 181-238.
39. Nesse RM. Is depression an adaptation? Archives of General Psychiatry. 2000; 57(1):14–20. [PubMed: 10632228]
40. 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(5):768–821. [PubMed: 20804236]
41. Dibeklioğlu* H, Hammal* Z, Yang Y, Cohn JF. Multimodal detection of depression in clinical interviews. ACM International Conference on Multimodal Interaction. 2015:307–310. *Equal contribution.
42. First, MB., Spitzer, RL., Gibbon, M., Williams, JB. Structured clinical interview for DSM-IV axis I disorders - Patient edition (SCID-I/P, Version 2.0). New York, NY: Biometrics Research Department, New York State Psychiatric Institute; 1995.
43. Hollon SD, Thase ME, Markowitz JC. Treatment and prevention of depression. Psychological Science in the public interest. 2002; 3(2):39–77. [PubMed: 26151569]
44. Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J. Antidepressant drug effects and depression severity: A patient-level meta-analysis. Journal of the American Medial Association. 2010; 303(1):47–53.
45. Jeni LA, Cohn JF, Kanade T. Dense 3D face alignment from 2D videos in real-time. IEEE International Conference on Automatic Face and Gesture Recognition. 2015
46. Jeni LA, Cohn JF, Kanade T. Dense 3D face alignment from 2D video for real-time use. Image and Vision Computing. 2017; 58:13–24.
47. Xiong X, Torre F. Supervised descent method and its applications to face alignment. IEEE Conference on Computer Vision and Pattern Recognition. 2013:532–539.
48. Velleman PF. Definition and comparison of robust nonlinear data smoothing algorithms. Journal of the American Statistical Association. 1980; 75(371):609–615.
49. Hotelling H. Relations between two sets of variates. Biometrika. 1936; 28(3/4):321–377.
50. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006; 313(5786):504–507. [PubMed: 16873662]
51. Hinton GE. Deep belief networks. Scholarpedia. 2009; 4(5):5947.
52. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research. 2010; 11:3371–3408.

53. Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. in International Conference on Machine learning. 2008:1096–1103.
54. Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification. European Conference on Computer Vision. 2010:143–156.
55. Dibeklioğlu H, Salah AA, Gevers T. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. European Conference on Computer Vision. 2012:525–538.
56. Boudahmane, K., Manta, M., Antoine, F., Galliano, S., Barras, C. TranscriberAG. [Online] Available: http://transag.sourceforge.net/
57. CMU Sphinx: Open source toolkit for speech recognition. [Online] Available: http://cmusphinx.sourceforge.net
58. Boersma, P., Weenink, D. Praat: Doing phonetics by computer. [Online] Available: http://www.fon.hum.uva.nl/praat
59. Nilsonne A. Speech characteristics as indicators of depressive illness. Acta Psychiatrica Scandinavica. 1988; 77(3):253–263. [PubMed: 3394527]
60. Darby JK, Simmons N, Berger PA. Speech and voice parameters of depression: A pilot study. Journal of Communication Disorders. 1984; 17(2):75–85. [PubMed: 6725627]
61. Leff J, Abberton E. Voice pitch measurements in schizophrenia and depression. Psychological Medicine. 1981; 11(4):849–852. [PubMed: 7323240]
62. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Transactions on Biomedical Engineering. 2000; 47(7):829–837. [PubMed: 10916253]
63. Hall JA, Harrigan JA, Rosenthal R. Nonverbal behavior in clinician-patient interaction. Applied and Preventive Psychology. 1996; 4(1):21–37.
64. Cannizzaro M, Harel B, Reilly N, Chappell P, Snyder PJ. Voice acoustical measurement of the severity of major depression. Brain and Cognition. 2004; 56(1):30–35. [PubMed: 15380873]
65. Alpert M, Pouget ER, Silva RR. Reflections of depression in acoustic measures of the patient's speech. Journal of Affective Disorders. 2001; 66(1):59–69. [PubMed: 11532533]
66. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and minredundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(8):1226–1238. [PubMed: 16119262]
67. Dibeklioğlu H, Salah AA, Gevers T. Recognition of genuine smiles. IEEE Transactions on Multimedia. 2015; 17(3):279–294.
68. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology. 2005; 3(2):185–205. [PubMed: 15852500]
69. Alshamlan H, Badr G, Alohali Y. mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. BioMed Research International. 2015; 2015
70. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin. 1968; 70(4):213–220. [PubMed: 19673146]
71. Kuncheva LI. Fusion of continuous-valued outputs. Combining Pattern Classifiers: Methods and Algorithms. 2004:151–188.
72. 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(10):641–647. [PubMed: 25378765]
73. Scherer S, Morency LP, Gratch J, Pestian J. Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2015:4789–4793.
74. Scherer S, Lucas GM, Gratch J, Rizzo AS, Morency LP. Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. IEEE Transactions on Affective Computing. 2016; 7(1):59–73.
75. Allen NB, Badcock PB. The social risk hypothesis of depressed mood: Evolutionary, psychosocial, and neurobiological perspectives. Psychological Bulletin. 2003; 129(6):887. [PubMed: 14599287]

Biographies

Hamdi Dibeklioğlu received the Ph.D. degree from the University of Amsterdam, The Netherlands, in 2014. He is currently a Post-doctoral Researcher in the Pattern Recognition & Bioinformatics Group at Delft University of Technology, The Netherlands. Earlier, he was a Visiting Researcher at Carnegie Mellon University, University of Pittsburgh, and Massachusetts Institute of Technology. His research interests include Affective Computing, Intelligent Human-Computer Interaction, Computer Vision, and Pattern Recognition. Dr. Dibeklioğlu was a Co-chair for the Netherlands Conference on Computer Vision 2015, and a Local Arrangements Co-chair for the European Conference on Computer Vision 2016. He served on the Local Organization Committee of the eNTERFACE Workshop on Multimodal Interfaces, in 2007 and 2010.

Zakia Hammal PhD is a Research Associate at the Robotics Institute at Carnegie Mellon University. Her research interests include computer vision and signal and image processing applied to human-machine interaction, affective computing, and social interaction. She has served as the organizer of several workshops including CBAR 2012, CBAR 2013, CBAR 2015, and CBAR 2016. She is a Review Editor in Human-Media Interaction section of Frontiers in ICT, served as the Publication Chair of the ACM International Conference on Multimodal Interfaces in 2014 (ICMI 2014), and is an Area Chair at IEEE FG 2017. Her honors include an outstanding paper award at ACM ICMI 2012, an outstanding reviewer award at IEEE FG 2015, and a Best Paper Award at IEEE ACII 2015.

Jeffrey F. Cohn is Professor of Psychology and Psychiatry at the University of Pittsburgh and Adjunct Professor of Computer Science at the Robotics Institute at Carnegie Mellon University. He leads interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis and synthesis of facial expression and prosody and applies those tools to research in human emotion, social development, nonverbal communication, psychopathology, and biomedicine. He co-chaired the IEEE International Conference on Automatic Face and Gesture Recognition (FG2017, FG2015, and FG2008), the International Conference on Affective Computing and Intelligent Interaction (ACII 2009), and the International Conference on Multimodal Interfaces (ACM 2014). He is a co-editor of IEEE Transactions in Affective Computing (TAC) and has co-edited special issues on affective computing for the Journal of Image and Vision Computing, Pattern Recognition Letters, Computer Vision and Image Understanding, and ACM Transactions on Interactive Intelligent Systems.

Fig. 1. Face-to-face clinical interview setup

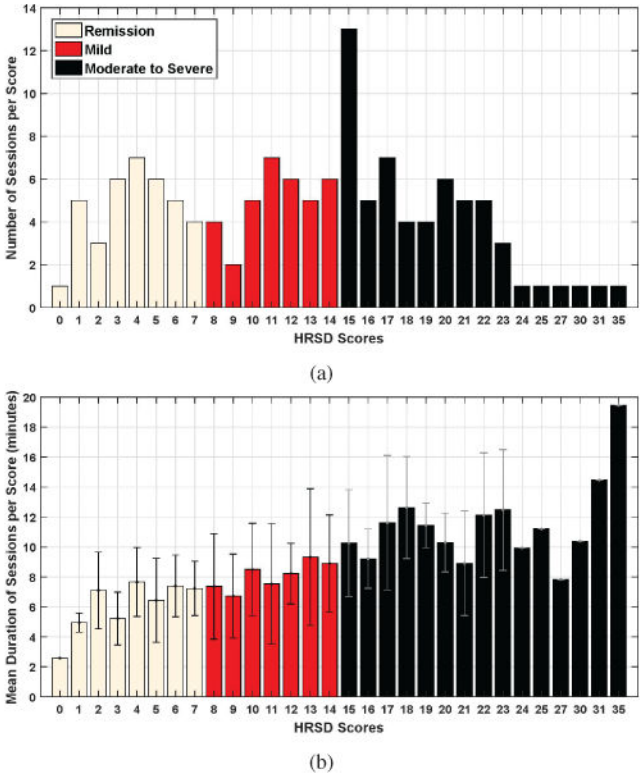

Fig. 2.

(a) Number of sessions per HRSD score and (b) mean duration (with standard deviation) of the interviews (per HRSD score) from the beginning of the first question to the end of the interview.

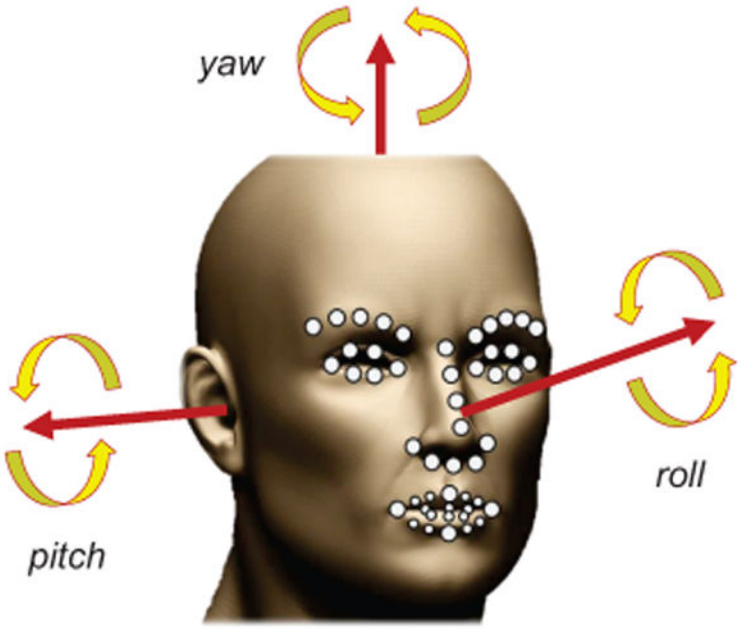

Fig. 3. The automatically tracked 3 degrees of freedom of head pose and the 49 facial landmarks

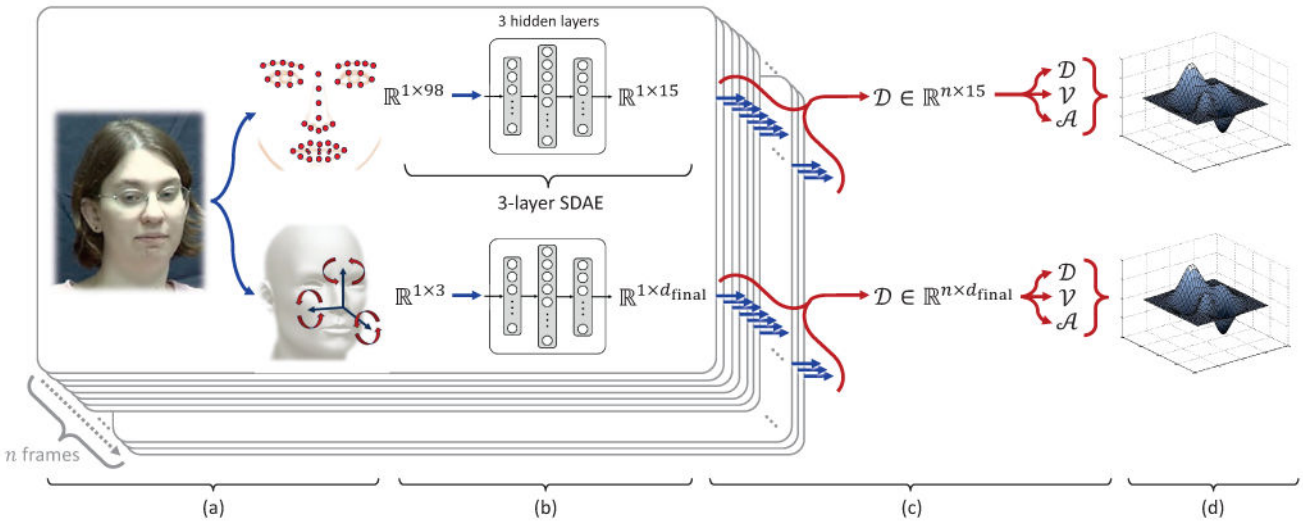

Fig. 4.

Overview of the proposed approach: (a) Tracking of facial landmarks and head pose, (b) per-frame encoding through Stacked Denoising Autoencoders, (c) extraction of per-frame dynamics (amplitude, velocity, and acceleration), and (d) per-video encoding through Improved Fisher Vector coding or Compact Dynamic Feature Set.

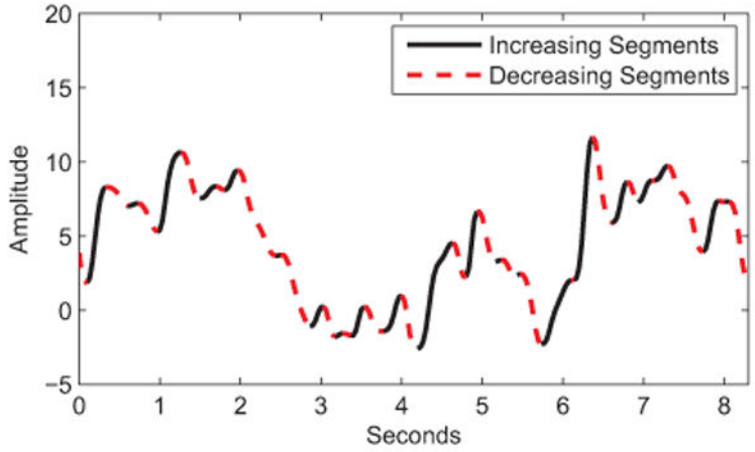

Fig. 5. Increasing and decreasing segments on an amplitude signal

TABLE I

The list of the considered hyperparameters of SDAEs.

Hyperparameter	Considered values
Number of units per hidden layer	$\{\lceil\frac{d}{4}\rceil, \lceil\frac{d}{2}\rceil, d, \lceil\frac{3d}{2}\rceil, 2d\}$
Fixed learning rate	{0.001, 0.01}
Number of pre-training epochs	{30, 50}
Corruption noise level	{0.1, 0.2, 0.4}

Note: d is the per frame features' dimensionality of the input data. Noise level corresponds to the fraction of corrupted inputs. Number of pre-training epochs corresponds to the pre-training of the denoising autoencoders. Fixed learning rate corresponds to the fixed error values for pre-training and fine-tuning.

TABLE II

Twenty one features defined in DFS

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

Note: ($^{+}$) refers to features measured form increasing segments and ($^{-}$) to features measured form decreasing segments (see Fig. 5). $\mathcal{D}$ corresponds to the amplitude, $\mathcal{V}$ velocity, and $\mathcal{A}$ acceleration.

TABLE III

Accuracy of depression severity classification using different per-frame and per-video based encoding for facial movement dynamics and head movement dynamics.

Representation		Classification Accuracy (%)				Weighted Kappa
Frame	**Video**	**Remission**	**Mild**	**Moderate to Severe**	**Mean**	
Facial Movement Dynamics						
SDAE Outputs	IFV	**67.57**	**65.71**	**84.48**	**72.59**	**0.62±0.059**
SDAE Outputs	DFS	64.86	60.00	79.31	68.06	0.53±0.062
Raw	IFV	59.46	60.00	81.03	66.83	0.52±0.062
Raw	DFS	56.76	57.14	81.03	64.98	0.50±0.063
Head Movement Dynamics						
SDAE Outputs	IFV	**62.16**	**54.29**	**79.31**	**65.25**	**0.51 ±0.063**
SDAE Outputs	DFS	59.46	48.57	77.59	61.87	0.48±0.064
Raw	IFV	56.76	45.71	75.86	59.44	0.46±0.064
Raw	DFS	54.05	40.00	74.14	56.06	0.40±0.065

Note: Weighted kappa is the proportion of ordinal agreement above what would be expected to occur by chance.

TABLE IV

Accuracy of depression severity classification using the different modalities separately.Voice refers to the prosodic features (see Section III-B).

Modality	Classification Accuracy (%)				Weighted Kappa
	Remission	Mild	Moderate to Severe	Mean	
Current Study					
Facial Movement Dynamics	**67.57**	**65.71**	**84.48**	**72.59**	**0.62±0.059**
Head Movement Dynamics	62.16	54.29	79.31	65.25	0.51±0.063
Voice	54.05	8.57	70.69	44.44	0.23±0.063
Alternative Methods					
Facial Movement Dynamics [41]	56.76	57.14	81.03	64.98	0.50±0.063
Facial Movements [14]	54.05	48.57	75.86	59.50	0.43±0.065
Head Movement Dynamics [41]	54.05	40.00	74.14	56.06	0.40±0.065
Head Movement GMM [19]	48.65	34.29	60.34	47.76	0.27±0.065
Head Movement Functionals [19]	64.86	20.00	74.14	53.00	0.42±0.062
Voice [11]	45.95	00.00	81.03	42.33	0.20±0.057

Note: Alternative methods were re-implemented from their description in the corresponding papers for the measurement of 3-levels of depression severity. For an accurate comparison, for each modality the same feature selection and the same classification architecture were used. Note that the original classification procedure for [14], and [11] does not include a feature selection step, while [19] uses a t-test threshold based feature selection.

TABLE V

Accuracy of depression severity classification with and without feature selection.Voice refers to the vocal features (see Section III-B). See Section V-B for definition of features for alternative methods.

Modality	Mean Accuracy (%)	
	All features	mRMR
	Current Study	
Facial Mov. Dynamics	65.68	**72.59**
Head Mov. Dynamics	59.87	**65.25**
Voice	43.54	**44.44**
	Alternative Methods	
Facial Mov. Dynamics [41]	54.26	**64.98**
Facial Movements [14]	51.46	**59.50**
Head Mov. Dynamics [41]	50.68	**56.06**
Head Mov. GMM [19]	45.63	**47.76**
Head Mov. Functionals [19]	47.29	**53.00**
Voice [11]	42.33	42.33

Note: Alternative methods were re-implemented from their description in the corresponding papers for the measurement of 3-levels of depression severity. “All features” refers to no feature selection and mRMR refers to the mRMR feature selection method. Best performing approach is boldfaced for each modality/feature type.

TABLE VI

Accuracy of depression severity classification by fusing different modalities.Face and head refer to the facial movement dynamics and head movement dynamics, respectively.Voice refers to the vocal features.

Modality	Classification Accuracy (%)				Weighted Kappa
	Remission	Mild	Moderate to Severe	Mean	
Face + Head	75.68	71.43	86.21	77.77	0.71±0.055
Face + Voice	67.57	65.71	86.21	73.16	0.66±0.058
Head + Voice	67.57	51.43	82.76	67.25	0.59±0.061
Head + Face + Voice	**78.38**	**71.43**	**86.21**	**78.67**	**0.73±0.054**