Heart Rate Estimation From Facial Videos for Depression Analysis

Aamir Mustafa
Department of Electronics and Communication Engineering
National Institute of Technology
Srinagar, India

Shalini Bhatia, Munawar Hayat and Roland Goecke
Human-Centred Technology Research Centre
University of Canberra
Australia

***Abstract*—Automated facial video analysis is useful in numerous health care applications. For example, spatio-temporal analysis of such videos has been previously done for assisting clinicians in the diagnosis of depression. Physiological measures, such as an individual's heart rate, provide very important cues to understand a person's mental health. Unobtrusively estimated heart rate has not been previously used to analyse individuals' mental health. In this paper, we automatically estimate heart rate activity from facial videos. We then study the association of the estimated heart rate activity with the person's mental health, as diagnosed by clinicians. Specifically, from the heart rate activity in response to watching different movies, we classify individuals as either depressed or healthy. The efficacy of the proposed scheme is demonstrated by experimental evaluations on a clinically validated dataset. Our results suggest unobtrusively estimated heart rate to be very effective for depression analysis.**

1. Introduction

According to the World Health Organization (WHO) and World Bank, depression has been identified as the leading cause of disability worldwide [1], only second to heart disease in magnitude of disease burden. WHO has predicted that by 2030, depression will account for the highest level of physical or mental disorder in the world [2]. As depression is becoming an increasingly serious global health problem, automated non-intrusive tools are required for timely detection and intervention as well as the monitoring of treatment progress and ongoing patient well-being. In the existing literature, mostly audio-visual data of a person's face and voice has been analysed to diagnose depression. Heart rate (HR), though being a very important modality to understand mental health of an individual, has not been explored for this purpose in the affective computing community. This is the first paper to unobtrusively measure heart rate activity of subjects and relate it to their mental health.

Joshi *et al.* [3] proposed a multimodal framework for depression diagnosis using audio and video analysis. For the video analysis, intra-facial muscle movements and movements of the head and shoulders were analysed by computing STIP (Space Time Interest Points) [4] and appearance features were analysed by using Local Binary Patterns in Three Orthogonal Planes (LBP-TOP) [5]. Subtle facial movements in videos have been analyzed for other tasks such as expression classification in [6], [7]. For the audio analysis, frequency, loudness, intensity and Mel-frequency cepstral coefficients were used. Cummins *et al.* used a spectro-temporal approach [8] to extract audio features for measuring the severity of clinical depression. They built both two class and five class support vector machine (SVM) [9] classifiers to investigate the potential discriminatory advantages of including long-term spectral information when classifying low or high levels of depression from features derived from the modulation spectrum. Recently, in [10], the authors have performed spatio-temporal analysis of facial videos and have been able to characterise subtypes of depression, i.e. melancholia and non-melancholia from healthy controls.

Correctly distinguishing sub-types of depression is a difficult problem because in depression datasets, specifically for melancholia, the facial activity greatly reduces. Chen *et al.* [11] compared physiological signals such as galvanic skin response (GSR), heart rate variability (HRV) and blood volume pulse (BVP) from depressed patients with those from healthy controls. These signals were found to be lower in depressed patients than in healthy controls. HRV was analysed from electrocardiogram (ECG) signals and GSR and BVP were measured using sensors mounted on the fingers. These techniques are cumbersome, obtrusive and require significant user cooperation. Initial findings in [12] show that the asymmetry of the electro-dermal activity (EDA) signal measured on the wrists could be used to indicate depression but still requires user-worn sensors.

Little work has been done in the field of pattern recognition to passively and unobtrusively determine the psychological state of a person. McIntyre *et al.* [13] have used person specific active appearance models (AAMs) to extract facial shape and appearance features and further SVM to classify facial expressions on a set of depressed patients and healthy controls. The experimental paradigm consisted of watching a set of seven movies as given in Table 1 for emotion elicitation in the subjects. It was found that depressed people respond differently to positive and negative content in movies as compared to healthy controls. In another study,

978-1-5386-0563-9/17/$31.00

Alghowinem *et al.* [14] implemented a framework by using an emotion elicitation paradigm to detect the affective state of the participants. Features such as eye movement, pupil dilation and pupil invisibility were used for the automatic recognition of affective states.

Research has shown that depression is a risk factor for patients with heart disease [15]. Heart rate is an important modality, which can help to understand the mental health of an individual. To the best of our knowledge, this paper is the first attempt to unobtrusively measure heart rate of subjects and relate it to their mental health. For this purpose, given facial videos of a person (captured while they are shown different movie clips), we devise a sliding window based strategy to compute a HR value for each window. The values for all windows are then combined into a vector. Smoothing and filtering operations are performed next to remove outliers and random erroneous spikes. These vectors are finally analysed to classify an individual as being healthy or depressed.

2. Data

For experimentation, real-world clinically validated video data from the Black Dog Institute – a clinical research facility in Sydney, Australia, offering specialist expertise in depression and its subtypes – dataset was used. The experimental paradigm contains several parts similar to [16] (a) watching movie clips, (b) watching and rating International Affective Picture System pictures, (c) reading sentences containing affective content, and (d) an interview between the participants and a clinically trained research assistant. In the interview, the subjects were asked to describe events in response to eight different groups of questions. This was done in order to elicit emotional responses in the participants.

The questions were designed to arouse both positive and negative emotions, e.g. ideographic questions such as "Can you tell me about some recent good news you've had?" and "Can you please tell me about a recent embarrassing experience?" The video recordings for each subject are 25-30 minutes long. For this study, the movie watching part (subject's facial response to different movie clips) of the videos was used with a total duration of 14m 12s as given in Table 1. The full dataset contains 130 subjects (60 patients and 70 healthy controls) carefully selected by clinicians.

The data was captured on an Apple Macbook Pro using the QuickTime Pro software. The video was captured at a resolution of 800×600 pixels at 30 fps using an AVT Pike F-100 FireWire camera. The camera was mounted on a tripod, which was placed behind the monitor, so as to record the face as a frontal pose. The height of the camera was adjusted with respect to the participants height (when seated). The audio was recorded at 44.1 kHz using a Sony microphone attached to participants lapel (at mid-level chest). During the dyadic conversation, the research assistant was standing to the left of the camera, behind the monitor [17].

TABLE 1. PARADIGM MOVIE CLIP LIST

Movie	Emotion	Length (mm:ss)	Colour (in plots)
Bill Cosby	Happiness	02:06	Red
The Champ	Sadness	02:49	Yellow
German Weather	Happiness	00:57	Green
Silence of the Lambs	Fear	03:44	Blue
Cry Freedom	Anger	02:40	Black
The Shining	Fear	01:07	Magenta
Capricorn One	Surprise	00:49	Cyan

3. Methodology

The input data consists of subject video clips of approximately 30 minute duration for each subject. The individuals are classified into two classes labeled as *i)* psychologically depressed patients and *ii)* healthy controls according to their mental health as examined by a clinician. For the purpose of this study, the movie watching part of the videos was used in order to investigate the heart rate (HR) activity in response to different stimuli.

In order to estimate frame-wise heart rate from a video, we devise a sliding window based strategy. For this purpose, we extract a window of frames and estimate the heart rate corresponding to these frames (Sec. 3.1 and 3.2). The window then slides, and the next set of frames is extracted. The process is repeated until the end of the video. The estimated heart rate values are then interpolated and post processed to remove outliers and spikes (Sec. 3.3). The final frame-wise heart rate signal is then used for classification.

3.1. Heart Rate Measurement

The movie watching part (which is approximately the first 14 minutes of each subject's video in the Black Dog dataset) is extracted from the video. Each video consists of clips from different movie genres, such as comedy, psychological horror, and drama, as shown in Table 1. We devise a method to measure HR for a window of minimum number of frames so as to explicitly differentiate and study various trends in the HR values over these genres.

Let $v = [f_1, f_2, f_3, \cdots, f_n]$ be the video clip for an individual where f denotes a video frame and n is the total number of frames. A video is captured at 30 fps, with the total number of frames being 22000 (i.e. $n = 22000$). Consider a sliding window $v_f = [f_1, f_2, \cdots, f_m]$ of m frames at a time, we estimate the HR value for this window. Ideally, a smaller value of m is desirable, since it will better approximate HR value corresponding to each individual frame. From the experiments, we observed that smaller m gives an unstable estimation for HR values and results in random spikes. Larger m yields better values but at the cost of reduced frame-wise specificity. We therefore deem $m = 300$ as a good trade-off between accuracy and specificity.

After estimating the HR value for the first m frames, the sliding window is moved by k frames. The value for k

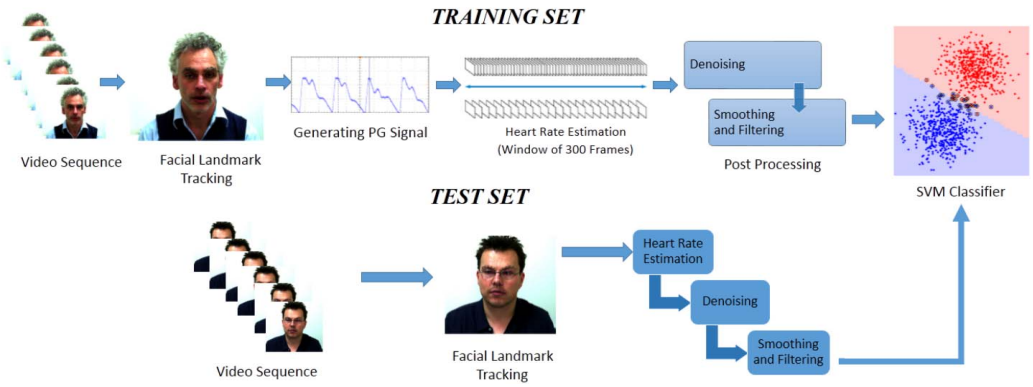

Figure 1. Illustration of the analysis pipeline: Given a video sequence, we extract a 300-frame window at a time. Next, we track facial landmarks for each frame and generate the PG signals. Then, we perform HR estimation from the plethysmograph (PG) signals for the entire sequence by sliding the window by 10 frames, followed by denoising, smoothing and filtering. Finally, we train an SVM classifier.

is empirically set to 10 as it was found to provide adequate accuracy. An even smaller value of k would be desirable, however it makes the experiments computationally expensive. Finally, an estimate of HR value is obtained after every $k = 10$ frames. These values are then concatenated into a vector. In order to estimate frame-wise HR values, we interpolate the measured values and smooth them to remove spikes and outliers. Section 3.2 below describes the technique employed to estimate HR values for the sliding window $v_f = [f_1, f_2, \cdots, f_m]$.

3.2. Heart Rate for Each Time Window

From the movie-watching part, the HR is estimated from each sliding window. The estimated HR values are then concatenated into a vector, interpolated to get frame-wise HR values, which is then considered as a feature vector. This feature vector can further be used for analysis and classification of an unknown person as depressed or healthy. Below, HR estimation from a sliding time window is described.

Given a facial video $v = [f_1, f_2, f_3 \cdots f_n]$ of the person's facial expression response to different movie stimuli, 66 facial landmarks are first tracked by using the method of Yu *et al.* [18]. The locations of the tracked landmarks are then used to register face regions in consecutive frames. This overcomes artifacts caused by rigid head movements. After registration, a pair of points is randomly selected and patches are extracted around these. By applying blind source separation, the plethysmograph (PG) signal is extracted from the green channel of the extracted patches. Temporal filtering is then applied on the extracted PG signal, followed by a frequency domain transformation of the smoothed signal. The peak in the $0.7 - 4$ Hz band of the generated spectrum is considered as the heart rate (HR) value. A confidence measure for the estimated HR value is defined in terms of the difference between the highest and the second highest peaks. If the confidence is greater than a threshold, the estimated HR is considered as final, otherwise, the complete procedure is repeated by selecting another pair of patches. Majority voting is then used to fuse the obtained HR values from different pairs of patches.

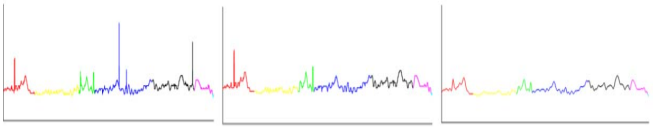

Figure 2. Step-wise post processing techniques. *Left*: Estimated HR values. *Middle*: Results of outlier removal. *Right*: Final HR values after moving average filtering (smoothing). x-axis represents the frames and the y-axis shows the HR values. Different movie clips are represented using colours as mentioned in Table 1.

3.3. Median Denoising

The estimated HR values for all sliding windows are then concatenated into a vector. The first plot of Fig. 2 shows the frame-wise HR values. However, on examining the values, we observe a few outliers, which need to be removed to better understand the variation. These outliers are due to *(i)* failure during tracking of facial landmarks as in the case of prolonged occlusions, *(ii)* excessive non-rigid facial movements e.g. laughing, yawning etc. and *(iii)* data corruption.

To remove these outliers, median filtering is applied. Around each sample, a window size of seven points is chosen, three on either side of the sample and the sample point itself. The median of the window and standard deviation of each sample about the window median are computed. If a sample differs from the median by a large extent, it is replaced by the median value. The frame-wise HR values after median denoising are shown in the second plot of Figure 2.

3.4. Smoothing and Filtering

The sample points obtained after median outlier removal (see previous Section 3.3) are then smoothed using moving

TABLE 2. RESULTS-AVERAGE HR VALUES OF SUBJECTS AND ITS DEVIATION ABOUT MEAN

	Average HR value	Standard Deviation in HR		Average HR Value	Standard Deviation in HR
Control 1	66.647	4.094	Patient 1	60.404	1.810
Control 2	74.591	2.653	Patient 2	79.627	2.115
Control 3	54.626	5.770	Patient 3	70.021	1.606
Control 4	64.407	3.574	Patient 4	81.032	1.510
Control 5	74.412	3.585	Patient 5	71.505	1.749
Control 6	76.748	4.193	Patient 6	82.009	1.640
Control 7	85.509	3.660	Patient 7	71.765	1.301
Control 8	64.670	5.798	Patient 8	81.709	1.229
Control 9	67.769	3.648	Patient 9	89.128	1.917
Control 10	73.562	3.347	Patient 10	61.497	1.324
Average Deviation in HR		**4.032**	**Average Deviation in HR**		**1.620**

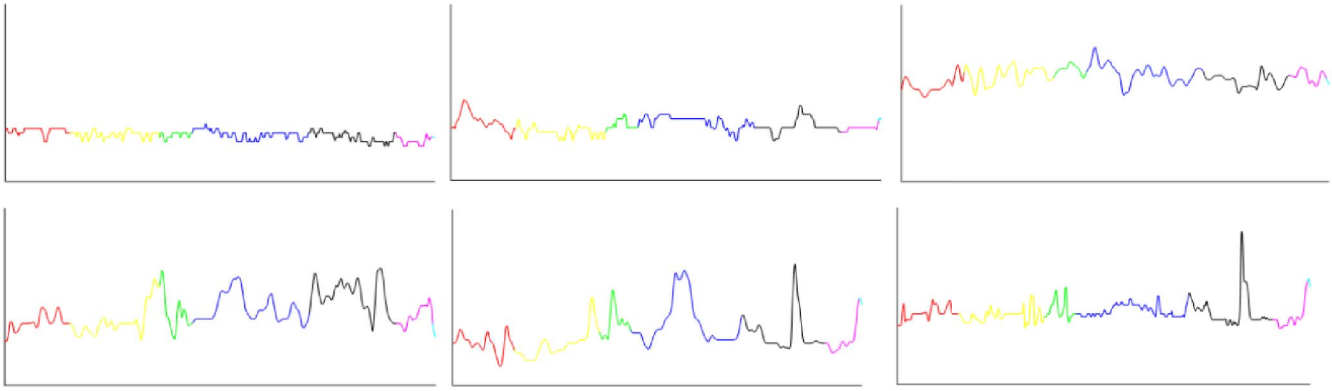

Figure 3. Heart Rate variation for patients (top row) and healthy controls (bottom row) against various genres of movies where the x-axis represents the frames and the y-axis shows the HR values. Note that HR values of patients do not vary much with the stimuli unlike those of healthy controls, which show multiple peaks and dips. A plausible explanation is that the psycho-motor retardation of patients with depression influences the HR response to the stimuli (reduced expression of affect). Different colours refer to the seven movie clips mentioned in Table 1.

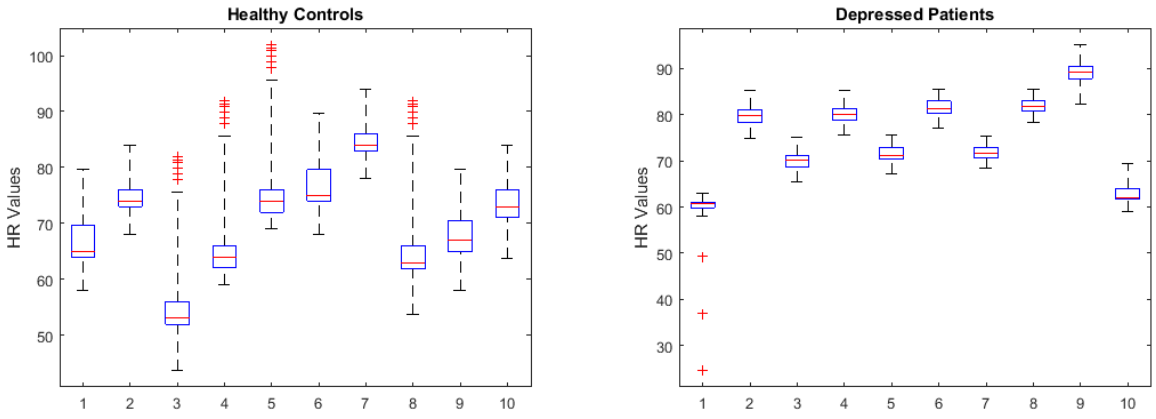

Figure 4. Box-plot shows variation in HR from the median value for healthy controls (*left*) and depressed patients (*right*) where x-axis represents the serial number of individuals (as in Table 2) and y-axis shows the HR values. Note that in case of healthy controls there are considerable variations in HR values, whereas in depressed patients the HR does not deviate much from the median value. This can also be inferred by comparing the standard deviation in HR values of healthy controls and patients from Table 2.

average filtering. Consider a span of five data points, two on either side, with the point to be smoothed at the center. Smoothing is done by replacing each data point with the average of the neighbouring points defined within the span. The final HR values after post-processing the entire video sequence for a depressed patient are shown in the third plot of Figure 2.

4. Results and Discussion

We analyse the heart rate activity of 20 subjects from the rich Black Dog Institute dataset. We use the estimated heart rate activity to classify an individual as depressed or healthy. Our experimental evaluations are discussed below.

1) **Trends in the HR of Healthy Controls and Depressed Patients:** The results reveal that normal heart rate for different individuals varies in a range of 60 to 85 bpm (beats per minute). The heart rate of healthy controls varies a lot with varying stimuli such as the different genres mentioned in the dataset. Gradual peaks and dips can be inferred from the HR plots of mentally healthy people. On the contrary, the heart rate of depressed patients remains more or less constant (flat) irrespective of the type of stimuli they are being subjected to. Fig. 3 shows frame-wise HR values of three patients (top row) and three healthy controls (bottom row). Fig.4 shows the box-plot of 10 controls (left) and 10 depressed patients (right). The inter-quartile range (IQR) in case of healthy controls is much larger as compared to that of depressed patients. Also there is a significant difference in the minimum and maximum HR values in controlled individuals indicating peaks and dips.
   It is worth mentioning that across all psychologically depressed patients, the heart rates show a peak on watching the movie clip 'Cry Freedom'. The movie is a British drama film based on apartheid era in South Africa in the late 1970's. The movie is based on real life events.
2) **Classifying unknown subjects using SVM:** The estimated frame-wise heart rate values for 20 subjects are used as feature vectors to classify them into depressed or healthy controls. For this purspose, we perform a leave-one-out cross-validation experiment, in which one subject is held out for testing, whereas the data for the remaining subjects is used for training. The experiment is repeated for all subjects. Using a linear Support Vector Machine (SVM) as classifier, an accuracy of 100% is achieved. While acknowledging the relatively small sample size, the results clearly indicate the effectiveness of heart rate as a modality for depression analysis. For SVM, we use LibLinear [19], with $C = 100$.

5. Conclusions and Future Work

We have devised a method of effectively dividing a facial video sequence for an individual into windows of frames and estimating HR for each time window. Frame-wise HR values for each individual form a feature vector, which is then used for training a classifier to classify the subjects into healthy controls or depressed patients. In the prior literature, other modalities such as audio-visual data of a person's face, and other invasive techniques (see Section 1) have been used for depression analysis; however, this method of HR estimation performs very strongly. The proposed method can be extensively used in medical (psychological) science to examine the mental health of individuals.

We note two limitations of our method: *(i)* The current Matlab implementation is computationally expensive and time consuming, as we move the sliding window by one-third of a second each time. *(ii)* We have performed our experiments on a smaller dataset of the Black Dog Institute (Section 2). In future, we hope to perform the experiments on larger publicly available datasets and effectively extract the most expressive and information rich video sections to reduce run time.

References

[1] R. C. Kessler, P. Berglund, O. Demler, R. Jin, D. Koretz, K. R. Merikangas, A. J. Rush, E. E. Walters, and P. S. Wang, "The epidemiology of major depressive disorder: results from the national comorbidity survey replication (ncs-r)," *Jama*, vol. 289, no. 23, pp. 3095–3105, 2003.

[2] V. Manicavasagar, "A review of depression diagnosis and management," *InPsych: The Bulletin of the Australian Psychological Society Ltd*, vol. 34, p. 8, 2012.

[3] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *JMUI*, vol. 7, no. 3, pp. 217–228, 2013.

[4] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[5] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on PAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[6] M. Hayat, M. Bennamoun, and A. El-Sallam, "Evaluation of spatiotemporal detectors and descriptors for facial expression recognition," in *Human System Interactions (HSI), 2012 5th International Conference on*. IEEE, 2012, pp. 43–47.

[7] M. Hayat and M. Bennamoun, "An automatic framework for textured 3d video-based facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 301–313, 2014.

[8] N. Cummins, J. Epps, and E. Ambikairajah, "Spectro-temporal analysis of speech affected by depression and psychomotor retardation," in *IEEE ICASSP*, 2013, pp. 7542–7546.

[9] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines. software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm," *ACM Trans. TIST*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[10] S. Bhatia, M. Hayat, M. Breakspear, G. Parker, and R. Goecke, "A video-based facial behaviour analysis approach to melancholia," in *FG 2017, 12th IEEE Conference on Automatic Face and Gesture Recognition (Accepted)*, 2017.

[11] Y.-T. Chen, I.-C. Hung, M.-W. Huang, C.-J. Hou, and K.-S. Cheng, "Physiological signal analysis for patients with depression," in *4th International Conference on Biomedical Engineering and Informatics*, 2011.

[12] S. Fedor, P. Chau, N. Bruno, R. Picard, and J. Camprodon, "Can we predict depression from the asymmetry of electrodermal activity?" in *Connected Health Symposium, Boston, MA*, 2016.

[13] G. McIntyre, R. Goecke, M. Breakspear, and G. Parker, "Facial response to video content in depression," in *MMCogEmS Workshop: Inferring Cognitive and Emotional States from Multimodal Measures, 13th International Conference on Multimodal Interaction ICMI2011, Alicante, Spain*, 2011.

[14] S. Alghowinem, M. AlShehri, R. Goecke, and M. Wagner, "Exploring eye activity as an indication of emotional states using an eye-tracking sensor," *Intelligent systems for science and information*, pp. 261–276, 2014.

[15] L. Culpepper, A. H. Clayton, J. A. Lieberman, and J. L. Susman, "Treating depression and anxiety in primary care," *Primary Care Companion to The Journal of Clinical Psychiatry*, vol. 10, no. 2, 2008.

[16] G. McIntyre, R. Goecke, M. Hyett, M. Green, and M. Breakspear, "An approach for automatically measuring facial activity in depressed subjects," in *International Conference on Affective Computing and Intelligent Interaction ACII*, 2009, pp. 223–230.

[17] J. Joshi, "A Multimodal Approach for Automatic Depression Analysis," Ph.D. dissertation, University of Canberra, Canberra, Australia, January 2016.

[18] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *IEEE International Conference on Computer Vision*, December 2013.

[19] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.