Affect Analysis using Multimodal Cues

Shalini Bhatia
University of Canberra, Human-Centred Technology Research Centre, Canberra, Australia

Abstract— **In this PhD project, I am working on the problem of multimodal affective sensing with focus on developing an objective measure for diagnosing and monitoring depression using multimodal cues, such as facial expressions, body gestures or vocal expressions. As the depression severity of a subject increases, the facial movements become very subtle. In order to quantify depression and its subtypes, we need to reveal these changes. A particular focus of my research is on improving the ability of affective computing approaches, to sense the subtle expressions of affect in face and voice, and to develop approaches that measure the severity of depression. In my work till date, I have investigated the sensitivity and specificity of affective computing approaches and demonstrated that these methods can successfully distinguish subtypes of depression. These results will serve as a baseline for future development towards a more fine grained depression classification and analysis.**

I. Problem and Motivation

This highly inter-disciplinary PhD project will investigate the fundamental affective computing problem of robust non-invasive multimodal approaches for sensing a person's affective state, with a particular focus on measuring the affect intensity. The affect intensity measure gives an indication of how strongly or weakly individuals tend to experience emotions in their everyday life. Affective computing [1] relates to adding emotional intelligence to computers such that the system is capable of recognizing, analysing and synthesizing human emotions. It expands Human-Computer Interaction (HCI) by including affective (emotional) communication together with appropriate means of handling affective information. In this project, multimodal affective sensing approaches will be investigated and developed in general. Novel paradigms to address current open research questions in affective sensing will be explored. The utility of such approaches will be validated on commonly used affective computing benchmark problems, such as the real-world example of quantifying depression severity.

Earlier work has shown that current affective sensing approaches can successfully distinguish distinct affect classes. For example, Joshi *et al.* [2] showed that a multimodal approach can correctly classify healthy controls and severely depressed subjects. However, current affective sensing approaches fail when the expressions of affect are very subtle. The aim of this project is to address the much more difficult problem of quantifying affect intensity on a continuous range as well as dealing with heterogeneous classes such as major depression [3], where the usual categorical approach such as depressed vs. non-depressed does not work, as often found in real-world applications of affective sensing technology [4]. I will exemplify my investigations on a number of commonly used affective sensing datasets [5], [6] around the benchmark problem of quantifying depression and melancholia. A multimodal system to objectively assess the severity of depression would provide a major breakthrough in mental health care and have significant potential for further research and commercialisation.

II. Background

The Black Dog Institute's Clinical Model[1] recognises four broadly different types of depression with different features and causes. Non-melancholic depression has psychological causes, and is linked to stressful events in a person's life, or the individual's personality style. It is the most common of the four types of depression. People with non-melancholic depression experience a depressed mood for more than two weeks, social impairment (for example, difficulty in dealing with work or relationships). In contrast to the other types of depression, it has a high rate of spontaneous remission. This is because it is often linked to stressful events in a person's life, which, when resolved, tend to see the depression also lifting. It responds well to different sorts of treatments (such as psychotherapies, antidepressants and counselling), but the treatment selected should respect the cause (e.g. stress, personality style).

Melancholic depression is the classic form of biological depression. Its defining features are – a more severe depression than is the case with non-melancholic depression, psychomotor disturbance: cognitive processing difficulties, with slowed thoughts and impaired capacity to work or study and an observable motor disorder (slowing and/or agitation of physical movements). Melancholic depression is a relatively uncommon type of depression. It affects only 1-2% of western populations. The numbers affected are roughly the same for men and women. Melancholic depression has a low spontaneous remission rate. It responds best to physical treatments (for example medications) and only minimally to non-physical treatments such as counselling or psychotherapy.

Psychotic depression is a less common type of depression than either melancholic or non-melancholic depression. The defining features of psychotic depression are – an even more severely depressed mood than is the case with either melancholic or non-melancholic depression, more severe psychomotor disturbance (PMD) than is the case with melancholic depression, psychotic symptoms (either delusions or hallucinations, with delusions being more common) and over-valued guilt ruminations. Psychotic depression has a

[1] https://lawsonclinic.com.au/depression/types-of-depression/

very low spontaneous remission rate. It responds only to physical treatments.

There is possibly a fourth type of depression known as atypical depression in contrast with the usual characteristics of non-melancholic depression. Instead of appetite loss, the person experiences appetite increase; and hypersomnia rather than insomnia. Someone with atypical depression is also likely to have a personality style of interpersonal hypersensitivity (that is, expecting that others will not like or approve of them). The individual can be cheered up by pleasant events.

III. Related Work

Inferring emotions from facial analysis has been a popular research topic in the computer vision and affective computing communities. Over the past two decades, various geometric, texture, static and temporal visual descriptors have been proposed for various expression analysis related problems [2]. Facial expression analysis methods can be broadly divided into three categories based on the type of feature descriptor used. Shape feature-based methods are based on facial geometry only. The second class consists of appearance feature-based methods, which analyse the skin texture. The third category is composed of hybrid methods, which use both shape and appearance features.

As described in [7], automated depression analysis can be undertaken from audio or visual features or both. The authors discuss three main types of approaches for automated analysis of depression: *(i)* the mean comparison approach used to compare individual behaviour between groups, *(ii)* the use of classification algorithms for grouping participants into predefined classes, and *(iii)* the use of regression algorithms to determine the severity of depression. All of these approaches use high-dimensional audio-visual features, such as facial expressions, eye gaze, head and body movements, speech, vocal pauses and voice quality. They also discussed some applications of automated depression diagnosis such as identifying behavioural indicators of depression, measuring change over time and response to intervention.

Joshi *et al.* [2] proposed a multimodal framework for depression analysis using audio-visual cues. For video analysis, intra-facial muscle movements and movements of the head and shoulders were analysed by computing Space Time Interest Points (STIP) [8] and appearance features were analysed by using Local Binary Patterns in Three Orthogonal Planes (LBP-TOP) [9]. For audio analysis, frequency, loudness, intensity and mel-frequency cepstral coefficients (MFCC) were used. Three different fusion strategies (feature level, score level and decision level) were explored in combination with a support vector machine [10] for classification. An accuracy of up to 91.7% was achieved in a binary classification task (depressed vs. non-depressed).

While extensive work has been done on facial expression recognition, automated analysis of facial data for depression analysis is still a young field. Recently, a multimodal assessment of depression, using facial, postural, and vocal behavioural measures in participants undergoing treatment for depression and classification as remitted, intermediate, or depressed using logistic regression classifiers and leave-one-out cross-validation was proposed [11]. Face and head movement dynamics outperformed vocal prosody, but a combination of all modalities performed best. The investigation suggests that automatic depression detection is possible from multimodal behavioural signs.

In another interesting work named SimSensei Kiosk [12] that was carried out over a period of two years, a fully automatic virtual human interviewer for differential diagnosis of distress indicators such as depression, anxiety and Post-traumatic stress disorder (PTSD) from verbal and non-verbal behaviours was developed. The virtual human was able to overcome the limitations of a human interviewer such as (i) stress and fear of being judged (ii) fear of disclosing information and (iii) inconsistency of spoken questions and gestures. However, the virtual human also had some limitations, such as lack of automated understanding capabilities and whether the users would feel comfortable sharing personal information with the virtual human and whether it would be sensitive to their non-verbal behaviour. The three stage design of SimSensei led to key functionalities such as dialogue processing, multimodal perception and non-verbal behaviour generation.

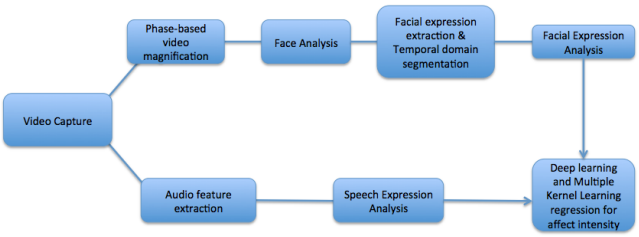

Fig. 1: Block Diagram of the proposed approach

IV. Proposed Work

The block diagram of the approach I propose to investigate in my PhD is shown in Figure 1. The research questions that will address the currently existing challenges in affect analysis are:

1) Can affective computing techniques be improved to reveal the subtle facial expressions which can further help in analysing facial affect?
2) Can fusion of multiple modalities help in estimating depression severity?
3) Can affect intensity be measured over a continuous scale and quantified over longer periods of time?

A. Analysis of Subtle Facial Expressions

It has been observed for the depression datasets, such as Black Dog Institute dataset [5] and University of Pittsburgh depression dataset [6] that facial movements overall are reduced the stronger the level of depression is. This is colloquially also known as the "frozen face", a reference to the underlying psychomotor retardation.[2] Some depressed

[2]https://www.blackdoginstitute.org.au/mental-health-wellbeing/depression

subjects can also show psychomotor agitation, i.e. an increase in facial movements. Any facial expressions and movements that are there are so minute that they are not reliably visible amongst the noise in the data due to low signal to noise ratio (SNR). In order to 'lift out' the relevant facial movements and expressions, I propose to investigate the phase-based [13] video magnification approach, which is designed to amplify minute movements in phase space. Due to the sensitive nature of data, unlike other generic affect datasets, these cannot be shared. While the Black Dog Institute dataset compares depressed patients with healthy controls, University of Pittsburgh dataset monitors depression severity over time.

Phase based video magnification manipulates motion in videos by analysing the signal of local phase over time in different spatial scales and orientations. Complex steerable pyramids are used to decompose the video and separate the amplitude of the local wavelets from their phase. The phases are then temporally filtered independently at each location, orientation and scale. Optionally, amplitude weighted spatial smoothing (phase denoising) is applied to increase the phase SNR. Then, the temporally bandpassed phases are amplified and the video is reconstructed [13].

B. Improved Multimodal Fusion using DCNN

Biologically-inspired perceptual fusion approaches draw models of information fusion largely from psychological and neurological findings pertaining to multi-sensor fusion in the human brain [14]. Convolutional Neural Networks are a class of deep learning models that replace the three stages of the standard video classification pipeline i.e. local feature extraction, fixed size video level description and training a classifier on the resulting feature representation with a single neural network, trained end-to-end from raw pixel values to classifier outputs. I will investigate deep learning approaches, in particular Deep Convolutional Neural Networks (DCNN), over time slices of data in the context of improved fusion as these have shown promise on frame-level and compare and contrast to widely used machine learning techniques (e.g. Support Vector Machine (SVM), Support Vector Regression (SVR)). For the audio features, I will use the spectro-temporal approach of Cummins *et al.* [15] for measuring clinical depression. It builds both two class and five class SVM (with radial basis function (RBF) kernel) 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.

C. Continuous Range Affect Intensity Measurement

I will investigate the utility of Multiple Kernel Learning (MKL) [16] Regression using RBF and Histogram Intersection Kernels (HIK), and then compare against the popular Multiview Learning approaches [17]. This will also help to characterise data as a function of affect intensity and, hence, of depression severity in the presence or absence of melancholic features in major depression to better understand how to characterise disturbances in affect. MKL performs feature selection by learning a convex combination of the kernels.

Preliminary research indicates that recent advances in Long Short Term Memory (LSTM) network frameworks are particularly interesting to investigate for this kind of multimodal analysis. Rajagopalan *et al.* [18] have proposed Multi-View LSTM (MV-LSTM), an extension to LSTM, which partitions memory cells and gates into multiple regions corresponding to different views. This takes advantage of complex view relationships that exist in most real world data. MV-LSTM was successfully applied to the problem of multimodal behaviour recognition and I plan to explore the use of MV-LSTM for the multimodal assessment of depression severity as well.

V. Data

For experimentation, real-world clinician 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 [19]: (a) watching movie clips, (b) watching and rating International Affective Picture System (IAPS) 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 please tell me what types of things make you anxious?" and "Can you tell me about a recent time when you felt sad or low?" The entire video recordings for each subject are 25–30 minutes long. For this study, the interview part of the videos was cropped with durations ranging from $153.6 - 663$s. The full dataset contains 130 subjects (60 patients and 70 healthy controls) carefully selected by clinicians. From the larger 130-subject dataset, 39 subjects were selected: 13 melancholic patients, 13 non-melancholic patients and 13 healthy controls. Note that we could only use 13 samples from each class, because melancholic patients constitute only about 10% of the depressed class. In the dataset available to us, we only had 13 melancholic samples. To ensure balance amongst the different classes, 13 samples were selected from each class.

VI. Current Work

The approach for characterising melancholia from non-melancholia and healthy controls used in this paper comprises of three phases. First, given an input video of facial images, facial regions of interest are detected and aligned (SectionVI-A). Next, a discriminative feature representation of the video is found (Section VI-B), which is then passed to a classifier (Section VI-C).

A. Face Detection and Alignment

Experimentally it has been shown earlier that the motion of a set of landmark points on a face, when exhibiting a particular facial expression, is similar across different people. Hence, for facial image analysis, 49 fiducial points (eyebrows, eyes, nose and mouth) were detected. Face detection and alignment was done using subject-independent regression based on the supervised descent method (SDM) [20], which is trained based on the mean pose deduced from the Viola-Jones face detector [21]. SDM is invariant to changes in pose and illumination and also partially occluded faces (such as eye glasses). For each of the 49 fiducial points, the (x, y) locations in image coordinates were stored for each video frame of the interview part of the dataset.

B. Representation of Affect

After face alignment, the shape and appearance features were extracted and used as the underlying representation and given as input to the classifier. Initially, the classifier was trained on the extracted shape and appearance features, but this led to over-fitting. This could be due to the large dimensionality of the feature vector in comparison to the sparse dataset. To overcome this problem, a Bag of Words (BoW) feature representation was used in this study. Different codebook sizes were experimented on as described below. First, let us look at the chosen geometric and appearance features.

1) Geometric Features: Face alignment or locating semantic facial landmarks, such as eyebrows, eyes, nose and mouth, is essential for intra-subject face analysis. In this study, geometric features were extracted from the video clips across all subjects by calculating the displacement vectors of the 48 face points pertaining to eyebrows, eyes, nose and mouth relative to the tip of the nose as the reference point. To capture the intra-subject facial movements, frame differencing between consecutive frames t and $t + 1$ was performed on the corresponding displacement vectors for all 39 videos.

2) Texture Features: Zhao *et al.* [9] showed that appearance based features are more effective in expression analysis, as they are able to capture subtle facial movements, which are difficult to capture otherwise using shape based features. In this study, texture features are extracted using LBP-TOP [9], which considers patterns in three orthogonal planes: XY, XT and YT, and concatenates the pattern co-occurrences in these three directions. The local binary pattern part of the LBP-TOP descriptor assigns binary labels to pixels by thresholding the neighbourhood pixels with the central value.

The aligned frames were of size 166×178 pixels, but the number of frames were different for all videos due to the varying length of interviews. To overcome this limitation, rather than computing LBP-TOP on the video in a temporally holistic manner, the descriptor was computed temporally piece-wise.

3) Bag of Words: The BoW approach, originally developed in the natural language processing domain, has been successfully applied to many tasks in image and vision analysis [22]. It represents documents based on the unordered word frequency. The power of the BoW framework in image analysis stems from its tolerance to variation in the appearance of objects. In this study, the video frames of a video clip are documents in the BoW sense.

One of the main advantages of using a BoW framework is that it handles different video lengths. The interviews are of different durations, depending on the contents of participants' speech. The use of codebooks makes it simpler to deal with such samples of different length, which is common in the clinical interview data used in this study. Codebooks of sizes ranging from 100 to 500 were computed by clustering the geometric and texture features separately for all videos.

C. Classification

For classifying melancholic, non-melancholic and healthy controls, SVM was used, since it provides good generalisation properties and is widely used as a benchmark. In this implementation, I have used the LibSVM [10]. Due to the limited amount of data and to see how well the approach would generalise to unseen data, leave-one-subject-out cross-validation was performed, without any overlap between training and testing data. A RBF kernel was used and a grid search was performed to find the optimal training parameters.

VII. Results and Discussion

To analyse the effect of codebook on classification, codebook sizes between 100 and 500 were considered. There is variation in accuracy, sensitivity and specificity across different codebook sizes but codebook size of 200 was mostly consistent across accuracy, sensitivity and specificity. It was also be observed that larger values of codebook sizes give higher accuracy but 100% sensitivity and that there is no further change in the performance parameters after increasing the codebook size beyond 200, which means that the classifier's concept of that class has been optimally captured at codebook size 200. Therefore the codebook size of 200 was selected as a standard size and was used for representing the combined geometric and texture features in order to capture any complementary information which would otherwise have been missed out in individual features. The results are given in Table I.

We also performed multi-class classification of melancholic, non-melancholic and healthy controls using the one vs. rest strategy for codebook sizes 200, 300 and 400 using combined features. The results for codebook size 200 are given in Table II. We agree that in real world settings such as hospitals and consulting rooms, the classes are unbalanced with healthy controls outnumbering the depressed patients. While an unbalanced dataset might mimic the real world more, the limitation is that it runs the risk of learning a biased classifier that is highly sensitive to the control class but has low specificity.

More experimental details and results can be found in the paper [23], where we report on a study to investigate facial

TABLE I: Results of combined features for codebook size 200: accuracy, sensitivity, and specificity.

Class ↓	**Acc.**	Sens.	Spec.
Mel vs. Non-Mel	**0.65**	0.75	0.61
Control vs. Mel	**0.69**	0.86	0.63
Control vs. Non-Mel	**0.62**	1	0.57

TABLE II: Results of combined features using one vs. rest strategy for codebook size 200: accuracy, sensitivity, and specificity.

Unbalanced Class ↓	**Acc.**	Sens.	Spec.
Non Mel vs. Mel+Control	**0.67**	0	0.67
Control vs. Mel+Non-Mel	**0.77**	0.83	0.76
Mel vs. Non-Mel+Control	**0.67**	0	0.67

behaviour analysis methods that can characterise melancholia from non-melancholia and healthy controls using direct methods, such as the SDM [20], and indirect methods, such as LBP-TOP [9] and BoW [22]. These methods are designed to capture the facial geometry and appearance from the subject videos that were recorded during an interview with a clinician.

VIII. FUTURE WORK

Both categorical and dimensional approaches are relevant to affective states such as depression. The categorical approach arguably provides a more accurate measure of a current temporary experience of emotion. The dimensional approach is more relevant to the temporal experience of affect such as mood. The next question is whether complex affect, as seen in the case of melancholic depression, can be conceptualised dimensionally and, if it can be, then can measuring this complex affect on a continuous scale help improve the classification accuracy of depressive sub-types? Another future direction is to detect subtle changes or micro expressions, as in the case of melancholic depression, to further improve the classification accuracy. The fusion of multiple modalities such as audio and video in estimating depression severity also need to be explored. Other advantages of using continuous over discrete approach are that analysis can be performed with small sample size and high sensitivity can be achieved.

IX. ACKNOWLEDGMENTS

I would like to thank Prof. Roland Goecke, primary supervisor and panel chair, and Dr. Munawar Hayat, co-supervisor, for their invaluable guidance and support. I would also like to thank the collaborators at the Black Dog Institute, at UNSW, Sydney and QIMR Berghofer, Brisbane.

REFERENCES

[1] R.W. Picard. *Affective Computing*. MIT Press, Cambridge, 1997.

[2] 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*, 7(3):217–228, 2013.

[3] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. American Psychiatric Publishing, Washington, DC, 2013.

[4] E. Kanjo, L. Al-Husain, and A. Chamberlain. Emotions in context: examining pervasive affective sensing systems, applications, and analyses. *Personal and Ubiquitous Computing*, 19(7):1197–1212, 2015.

[5] Jyoti Joshi. *A Multimodal Approach for Automatic Depression Analysis*. PhD thesis, University of Canberra, Canberra, Australia, January 2016.

[6] J.F. Cohn, T. Kruez, I. Matthews, Y. Yang, M. Nguyen, M. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *ACII*, pages 1–7, 2009.

[7] J.M. Girard and J.F. Cohn. Automated audiovisual depression analysis. *Current opinion in psychology*, 4:75–79, 2014.

[8] I. Laptev. On space-time interest points. *IJCV*, 64(2–3):107–123, 2005.

[9] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. PAMI*, 29(6):915–928, 2007.

[10] C. Chang and C. Lin. Libsvm: A library for support vector machines. software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm. *ACM Trans. TIST*, 2(3):27:1–27:27, 2011.

[11] H. Dibeklioglu, Z. Hammal, Y. Yang, and J.F. Cohn. Multimodal detection of depression in clinical interviews. In *ACM ICMI*, pages 307–310, 2015.

[12] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L-P. Morency. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on AAMAS*, pages 1061–1068, 2014.

[13] N. Wadhwa, M. Rubinstein, F. Durand, and W.T. Freeman. Phase-based video motion processing. *ACM Trans. Graphics, (Proc. SIGGRAPH)*, 32(4):80:1–80:9, 2013.

[14] T.S. Andersen, K. Tiippana, and M. Sams. Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research*, 21(3):301–308, 2004.

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

[16] M. Varma and R.B. Bodla. More generality in efficient multiple kernel learning. In *ICML*, pages 1065–1072, 2009.

[17] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[18] S.S. Rajagopalan, L-P. Morency, T. Baltrusaitis, and R. Goecke. Extending long short-term memory for multi-view structured learning. In *ECCV*, pages 338–353, 2016.

[19] G. McIntyre, R. Goecke, M. Hyett, M. Green, and M. Breakspear. An approach for automatically measuring facial activity in depressed subjects. In *ACII*, pages 223–230, 2009.

[20] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE CVPR*, pages 532–539, 2013.

[21] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE CVPR*, volume 1, pages I–511–I–518, 2001.

[22] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE CVPR*, volume 2, pages 524–531, 2005.

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