A Multimodal Approach for Automatic Depression Analysis

Jyoti J. Dhall

A Thesis Submitted for the Degree of Doctor of Philosophy of the University of Canberra

27 January 2016

Faculty of Education, Science, Technology and Mathematics

UNIVERSITY OF CANBERRA

AUSTRALIA'S CAPITAL UNIVERSITY

Dedicated to my darling son and loving husband.

Acknowledgements

I would like to thank my supervisory panel: Prof. Roland Goecke, Prof. Michael Wagner and Prof. Tom Gedeon for the constant support and guidance throughout my PhD journey. A special gratitude to Prof. Roland Goecke for believing in my abilities and always encouraging me. Thanks to Prof. Michael Breakspear and all the partners of the project at the Black Dog Institute, Sydney and at the Queensland Institute of Medical Research, Brisbane for the elaborative and insightful discussions. Thanks to the administrative staff at the University of Canberra, especially to Ms. Serena Chong, for always being so kind and helpful. I am grateful to all my wonderful friends and colleagues in Canberra for the constant support network.

I thank my husband, Abhinav, who has been very supportive, patient and a true friend to me all along. Despite going through the same PhD journey yourself, you have always motivated and encouraged me to never give up. I will forever be grateful to you for inspiring me to pursue this worthy endeavour. A heartfelt appreciation for the efforts my dear mother, Kiran, has put to make sure that I reach the finish line. It would not have been possible without you. Thanks to my father, Sh. Virender K. Joshi, for always being my strength. Thank you Mama and Papa, for your understanding and your unconditional love. My endless love and thanks to my kind brother, Deepak, for his love and care. My gratitude to my lovely parents-in-law, who have always supported and encouraged me. Deepest love and thanks to dear Abhishek and Monica for always backing me up. Finally, to my little baby, Sid for coping up in my busy days of compiling and writing the thesis.

Abstract

Depression is one of the most common and disabling mental disorders, and has a major impact on society. The landmark WHO 2004 Global Burden of Disease report quantified depression as the leading cause of disability worldwide (an estimated 154 million sufferers). Fortunately, depression can be ameliorated through the provision of suitable objective technology for detecting depression. Disturbances in the expression of affect reflect changes in mood and interpersonal style, and are arguably a key index of a current depressive episode. This leads directly to impaired interpersonal functioning, causing a range of interpersonal disabilities, functioning in the workforce, absenteeism and difficulties with a range of everyday tasks (such as shopping). Whilst these are a constant source of distress in affected subjects, the economic impact of mental health disorders through direct and indirect costs has long been underestimated. Despite its severity and high prevalence, there currently exist no laboratory-based objective measures of illness expression, course and recovery. This compromises optimal patient care, compounding the burden of disability. As healthcare costs increase worldwide, the provision of effective health monitoring systems and diagnostic aids is highly important. With the advancement of affective sensing and machine learning, computer aided diagnosis can and will play a major role in providing an objective assessment.

The research presented in this thesis addresses some of the key issues of automatic depression analysis mentioned as follows: 1) analysing geometrical and appearance descriptors for depression analysis; 2) the role of upper body movements in detecting depression; 3) fusion of audio and video channels; 4) relative body parts movement analysis for depression detection. The central hypothesis of the thesis is that using different modalities and information from body parts, will lead to more accurate detection of depression. To validate the approaches, clinically approved datasets from the Black Dog Institute, Sydney, and the University of Pittsburgh, USA, are used.

First, subject-dependent active appearance model based geometrical descriptors are computed. Subject-independent parts based models are applied in parallel to extract texture descriptors. A thorough compar-

ison is made on the Black Dog Institute clinical data. Furthermore, head movements are computed using the fiducial points and a histogram is constructed. Space Time Interest Points are also computed on the upper body to capture subtle gestures, which can provide discriminative information. The speech signal is also analysed and its feature descriptors are combined with visual information extracted from face and upper body, respectively. Various fusion scenarios are studied in this multimodal framework. The contribution of body expressions are further explored by proposing a relative part movement framework and validating it on the University of Pittsburgh data. To the best of my knowledge, this is the first work in affective computing community to use body expressions for detecting depression. The results presented in this thesis show that, as hypothesised, the multimodal framework outperforms uni-modal approaches in the task of classifying between depressed patients and healthy controls. Moreover, the body expressions, used as an auxiliary modality, provide significant discriminating information for depression recognition.

List of Publications

Publications by the Candidate Relevant to the Thesis

Peer-Reviewed Publications

1. N. Cummins, **J. Joshi**, A. Dhall, V. Sethu, R. Goecke and J. Epps. *Diagnosis of Depression by Behavioural Signals: A Multimodal Approach*. Proceedings of the 3rd International Audio-Visual Emotion Challenge and Workshop (AVEC 2013), 21st ACM International Conference on Multimedia (MM13), Barcelona, Spain, 21-25 Oct 2013.

2. **J. Joshi**, A. Dhall, R. Goecke and J. F. Cohn. *Relative Body Parts Movement for Automatic Depression Analysis*. Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction ACII2013, Geneva, Switzerland, 2-5 Sep 2013. (Oral)

3. **J. Joshi**. *An Automated Framework for Depression Analysis. Doctoral Consortium*. Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction ACII2013, Geneva, Switzerland, 2-5 Sep 2013. **[Fiorella de Rosis Award]**

4. **J. Joshi**, R. Goecke, M. Breakspear and G. Parker. *Can Body Expressions Contribute to Automatic Depression Analysis?* 10th IEEE International Conference on Automatic Faces and Gesture Recognition FG2013, Shanghai, China, 22-26 Apr 2013. (Oral)

5. **J. Joshi**, S. Alghowinem, A. Dhall, R. Goecke, M. Wagner, M. Breakspear, J. Epps and G. Parker. *Multimodal Assistive Technologies for Depression Diagnosis and Monitoring*. Journal on Multimodal User Interfaces, Springer 2013.

6. **J. Joshi**, A. Dhall, R. Goecke, M. Breakspear and G. Parker. *Neural-Net Classification for Spatio-Temporal Descriptor Based Depression Analysis*. Proceedings of the 21st International Conference on Pattern Recognition ICPR2012, Tsukuba, Japan, 11-15 Nov 2012. (Oral)

7. **J. Joshi**. *Depression Analysis: A Multimodal Approach*. Doctoral Consortium at the 14th ACM International Conference on Multimodal Interaction ICMI2012, Santa Monica, CA, USA, 22-27 Oct 2012.

8. I. Radwan, A. Dhall, **J. Joshi** and R. Goecke. *Regression Based Pose Estimation with Automatic Occlusion Detection and Rectification*. Proceedings of the IEEE International Conference on Multimedia & Expo ICME 2012, Melbourne, Australia, 9-13 July 2012. **[Best Paper Award Nomination** (Oral)

Other Publications Relevant to but not Forming Part of the Thesis

1. A. Dhall, **J. Joshi**, K. Sikka, R. Goecke and N. Sebe, *The More the Merrier: Analysing the Affect of a Group of People in Images*. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition FG 2015, Ljubljana, Slovenia, 4-8 May 2015. (Oral)

2. A. Dhall, R. Goecke, **J. Joshi**, K. Sikka and T. Gedeon. *Emotion Recognition in the Wild Challenge 2014: Baseline, Data and Protocol*. Proceedings of the 16th International Conference on Multimodal Interaction ICMI 2014, Istanbul, Turkey, 12-16 November 2014.

3. A. Dhall, R. Goecke, **J. Joshi**, M. Wagner and T. Gedeon. *Emotion Recognition In The Wild Challenge 2013*. Proceedings of the 15th ACM International Conference on Multimodal Interaction ICMI2013, pages 509-516, Sydney, Australia, 9-13 December 2013.

4. A. Dhall, **J. Joshi**, I. Radwan and R. Goecke. *Finding Happiest Moments in a Social Context*. Proceedings of the 11th Asian Conference on Computer Vision ACCV2012, Lecture Notes of Computer Science 7725, pages 613-626, Daejeon, Korea, 5-9 November 2012.

List of Abbreviations

AAM	Active Appearance Model
ACF	Auto Correlation Function
AFER	Automatic Facial Expression Recognition
ANN	Approximate Nearest Neighbour
APA	American Psychiatric Association
ASD	Autism Spectrum Disorder
ASM	Active Shape Model
AU	Action Units
AVEC	Audio Visual Emotion Challenge
BDI	Beck Depression Inventory
BlackDog	The Black Dog Institute Depression dataset
BoA	Bag of Audio features
BoB	Bag of Body expressions
BoF	Bag of Facial dynamics
BoV	Bag of Video features
BoW	Bag of Words
CES-D	The Centre for Epidemiologic Studies Depression Scale
CIDI	Composite International Diagnostic Interview
DIS	Diagnostic Interview Schedule
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders - Fourth Edition
EEG	Electroencephalogram
EMOTIW	Emotion recognition in the Wild
ERP	Event Related Potential

FABO	Face and Body Gesture database
FACS	Facial Action Coding System
FER	Facial Expression Recognition
fMRI	Functional Magnetic Resonance Imaging
FV	Fisher Vector
HHM	Histogram of Head Movement
HOF	Histogram of Optical Flow
HOG	Histogram of Oriented Gradient
HRSD	Hamilton Depression Rating Scale
ICD	International Classification of Disease
IAPS	International Affective Picture System
LBP	Local Binary Patterns
LBP-TOP	Local Binary Pattern in Three Orthogonal Plane
LPQ	Local Phase Quantisation
MADRS	Montgomery-Åsberg Depression Rating Scale
MAP	Mood Assessment Program
MDD	Major Depressive Disorders
MFCC	Mel-scale Frequency Cepstral Coefficients
MINI	Mini-International Neuropsychiatric Interview
MoPS	Mixture of PartS
PCA	Principal Component Analysis
PHQ	Patient Health Questionnaire
PTSD	Post Traumatic Stress Disorders
QIDS	Quick Inventory of Depression Symptomatology
SDM	Supervised Descent Method
SIC	Simultaneous Inverse Compositional
STIP	Space Time Interest Points
SVM	Support Vector Machine
SVR	Support Vector Regression
UPitt	The University of Pittsburgh depression data

VA	Valence-Arousal Model
VJ	Viola-Jones
WHO	World Health Organisation
WMH-CIDI	World Mental Health-Composite International Diagnostic Interview
Zung-SDS	Zung Self-Report Depression Scale

Contents

Declaration v

Acknowledgements vii

Abstract ix

List of Publications xi

List of Abbreviations xiii

List of Figures xxi

List of Tables xxv

1 Introduction 1

1.1 Motivation & Aim 2

1.2 Challenges 4

1.3 Objectives 5

1.4 Structure of the Thesis 5

2 Background 9

2.1 Non-Verbal Affect Analysis 10

2.1.1 Facial Expression analysis 10

2.1.2 Body Expression Analysis 14

2.2 Verbal Affect Analysis 16

2.3 Multimodal Affect Analysis 19

2.4 Depression Sensing Technologies 21

2.4.1 Clinical Depression Assessment . . . 22
2.4.2 Depression Sensing via Non-verbal Behavioural Signals . . . 26
2.4.3 Depression Sensing via Speech Signals . . . 29
2.4.4 Multimodal Depression Sensing . . . 31
2.4.5 Depression Sensing via Social Media . . . 33
2.4.6 Depression Sensing via Physiological Signals . . . 34
2.5 Summary . . . 35

3 Datasets **37**
3.1 Black Dog Institute Dataset . . . 38
3.1.1 Diagnosis and Severity Assessment . . . 38
3.1.2 Data Recording Paradigm . . . 39
3.1.3 Hardware Description . . . 41
3.1.4 Selected Cohort: BlackDog Data . . . 41
3.2 The University of Pittsburgh Dataset . . . 42
3.2.1 Diagnosis and Severity Assessment . . . 42
3.2.2 Hardware Description . . . 42
3.2.3 Selected Cohort: UPitt Data . . . 43
3.3 Audio-Visual Emotion Challenge (AVEC) Depression Dataset . . . 44
3.3.1 Diagnosis and Severity Assessment . . . 44
3.3.2 Data Recording Paradigm . . . 44
3.3.3 Selected Cohort & Limitations . . . 45
3.4 Summary . . . 46

4 Geometric and Texture Facial Features **47**
4.1 Geometric Feature Based Static Analysis . . . 48
4.1.1 Active Appearance Models . . . 48
4.1.2 Supervised Descent Method . . . 51
4.1.3 Displacement Feature . . . 51
4.2 Texture Feature Based Temporal Analysis . . . 52
4.2.1 Local Binary Patterns in Three Orthogonal Plane . . . 52
4.2.2 Bag of Words . . . 53

4.3 Experiments and Results . . . 54
4.4 Summary . . . 57

5 Evaluation of Upper Body Expressions 59

5.1 Approach . . . 60
5.1.1 Upper Body Analysis . . . 61
5.1.2 Face Analysis . . . 62
5.1.3 Key Interest Point Selection . . . 63
5.1.4 Bag of Words . . . 63
5.1.5 Head Movement Analysis . . . 63
5.2 Experiments and Results . . . 65
5.3 Summary . . . 69

6 Fusion of Audio and Visual Channels 71

6.1 Approach . . . 72
6.1.1 Video Features . . . 73
6.1.2 Audio Features . . . 74
6.1.3 Bag of Words . . . 75
6.2 Fusion . . . 76
6.2.1 Feature Fusion . . . 76
6.2.2 Score Fusion . . . 76
6.2.3 Decision Fusion . . . 77
6.3 Experiments and Results . . . 77
6.4 Summary . . . 82

7 Relative Body Parts 85

7.1 Approach . . . 88
7.1.1 Relative Part Movement . . . 88
7.1.2 Holistic Body Movement . . . 89
7.1.3 Fusion . . . 92
7.2 Experiments and Results . . . 92
7.3 Summary . . . 97

8 Conclusion & Future Work **99**

8.1 Summary of Contributions 99

8.2 Future Work 102

A Audio-Visual Emotion Challenge 2013 **105**

A.1 STIP 105

A.2 Time Slice based BoW 106

A.3 Experiment Section 106

A.3.1 Results 107

Bibliography **109**

List of Figures

2.1 Example of some action units from the Cohn and Kanade Database [Kanade 00]. 12

2.2 Plucthik's wheel of emotion [Plutchik 80]. 13

2.3 Valence-Arousal Model for Affect Representation [Russell 80]. 14

3.1 The recording setup at the Black Dog Institute with the subject on the left, the screen showing the stimuli and a camera and microphone recording the face and upper body and the voice, respectively. 40

3.2 The synchronised audio-visual capture of the interviewer and the participant in the University of Pittsburgh data. The images in the left column represent the full view of the recording setup. The images in the right column show the output of four synchronised cameras: Topmost image is the interviewer's face; Second from top is the full-body view of the participant; the last two images represent recording of the face and shoulder 15 degrees to the participant's left and right. [Cohn 09] 43

4.1 Top three shape modes example [Asthana 13] generated using the DeMoLib library . . . 49

4.2 Top three appearance modes example [Asthana 13] generated using the DeMoLib library. 49

4.3 The LBP-TOP feature representation in each block volume (Image Courtesy: [Zhao 07]). (a) Block Volumes (b) LBP feature computed on three orthogonal plane (c) Concatenated features from one block volume with the appearance and motion. 52

4.4 Bag of Words Representation (Image Courtesy : [Li 05a]). In this example image, in BoW sense, the face image is a *document* and the distinct facial parts are the corresponding *words* of the document. 54

5.1 The figure describes the STIP computation on a video from FEEDTUM database [Wallhoff 06]. The blocks represent the gradient information [Vondrick 13] around the generated interest points. The HOG and HOF features represent the local spatio-temporal movements. 61

5.2 Graphs (a) and (b) describe the Histogram of Head Movements (HHM) of patients and controls, respectively, normalised over time. The HHMs are computed over time intervals of 60s and normalised. The angles in the diagram depict the change in a rigid fiducial point over time, measuredin a Cartesian plane. 64

5.3 STIP visualisation. The top two rows show the STIP generated on the upper body visible in the video frames in our database. The bottom two rows show the STIP generated on the aligned facial frames for the same video. The yellow circles indicate the presence and intensity of interest points. It is evident from the comparison of the frames inside the two rectangles that upper body expressions tend to generate many interest points, which provide useful discriminative information. 66

5.4 The four graphs describe the accuracy comparison between depression detection by BoB (Bag of Body expression) and BoF (Bag of Facial dynamics) for different configurations of STIP. Here, STIP1 means STIP with level one cluster size $k_b = k_f = 1000$, STIP2: $k_b = k_f = 1500$, STIP3: $k_b = k_f = 2000$ and STIP4: $k_b = k_f = 2500$ 67

5.5 A comparison of the frequency of occurrence of head movements versus a static head position in depressed patients and healthy controls. 68

6.1 Flow of the proposed system: Audio and video data are processed individually and respective features are computed. All audio features are combined in a Bag of Audio Features (BoA), while video features are combined in a Bag of Visual Features (BoV). Different fusion methods are then experimented on. 72

6.2 Flow of the proposed system. 73

6.3 Flow of the speech processing subsystem to extract audio features: Intensity, Loudness, f0 and MFCC. 75

6.4 The three graphs show the accuracy of the system and the effect of choosing different codebook sizes of BoA while fusing it with a selected BoV codebook combination STIP1_200+LBP1_200 for different fusion methods: a) Feature Fusion, b) Score Fusion, c) Decision Fusion. 79

6.5 The three graphs show the accuracy of the system and the effect of choosing different combinations of C_s and C_l, while fusing with a selected audio feature f0+I+L_750 for different fusion methods: a) Feature Fusion, b) Score Fusion, c) Decision Fusion. 80

7.1 Flow of the proposed framework. Given a video containing a subject, body parts are detected using [Yang 11]. STIP are computed on the body window. Key interest points are chosen and vector quantisation is performed and a histogram is generated. For analysing the relative part movements, parts centres are computed and their relative position is calculated with respect to the torso centre. A polar histogram is computed depicting motion patterns. The two histograms are concatenated and an SVM model is used to infer the label. 87

7.2 RPM histograms for four subjects who were severely depressed at one point in time and have shown improvement over the course of treatment. Each column shows two plots depicting body motion patterns, which belong to the same participant in two different states. It is visually apparent that as the participants' HRSD score decreases, their RPM shows higher activity. 90

7.3 The figure visualises the different body part RPMs of one particular participant and changes in the motion pattern observed over time. The top row represents the motion patterns when the participant was diagnosed as being severely depressed and the bottom row shows the plot for the same participant with lower HRSD score. 91

7.4 The figure demonstrate the variation in the performance of RPM histograms with changes in bin size of distance (r) and orientatoin (θ) for Subset I and Subset II. 93

7.5 Performance comparison of STIP and RPM approaches for part specific movement analysis for Subset I and Subset II. 96

List of Tables

2.1 The table shows an example of relationship between various AUs and six basic emotions. 12

2.2 Comparison of approach of EmotiW 2013 & 2014 challenge participants. Here OS: OpenSmile, MoPS: Mixture of Pictorial Structures, IS: InterSpeech 2013 Challenge, LBP-TOP: Local Binary PatternThree Orthogonal Planes, SVM- Support Vector Machine, PLS- Partial Least Squares, NN-Nearest Neighbor, ANN- Artificial Neural Network, CNN - Convolutional Neural Network, HMM - Hidden Markov Model. 20

2.3 This Table summarises various behavioural differences observed in separate studies, based on [Scherer 14] with some additions. 25

4.1 Performance comparison of AAM, SDM and LBP-TOP on the BlackDog data. 56

5.1 Best classification accuracies and F-score measures for different configurations of STIP for the Bag of Body expressions (BoB) and Bag of Facial dynamics (BoF). 66

6.1 Comparison of classification accuracies for individual video and audio features. Here, STIP1 - Level One clusters C = 2500, STIP2 - Level One clusters C = 5000, LBP1 - LBP-TOP with clip length $t = 6s$, LBP2 - LBP-TOP with clip length $t = 1s$. 78

6.2 Top five classification accuracy for different fusion methods for various parameters of the features. Here, W.Sum - Weighted Sum, W.Product - Weighted Product, Concat. - Concatenated 81

7.1 Best classification accuracies and F-score measures for different configurations of STIP for subset I and subset II for holistic body movement analysis. 94

7.2 This table compares the best classification accuracies and F-score measures from RPM and STIP based approaches for subset I and subset II for full body analysis. The last row shows the increase in the performance on fusion. 95

7.3 This table compares the best classification Part specific accuracies and F-score measures from RPM and STIP based approach for subset I and subset II. 96

A.1 Performance on the Development Set, comparison of the proposed technique is done with the vision baseline 107

A.2 Performance on the Test Set, comparison of the proposed technique is done with the vision baseline. 107