Automatic detection of visual cues associated to depression

Anastasia Pampouchidou

Doctoral Thesis

Automatic Detection of Visual Cues Associated to Depression

Anastasia Pampouchidou

November 2018

Doctoral Thesis

Automatic Detection of Visual Cues Associated to Depression

Anastasia Pampouchidou

Supervised by:
Fabrice Mériaudeau (LE2I/CISIR - UBFC/UTP),
Panagiotis Simos (UoC/FORTH),
Manolis Tsiknakis (BMI/CBML - TEIC/FORTH),
Kostas Marias (BMI/CBML - TEIC/FORTH),
Fan Yang (LE2I - UBFC)

November 2018

Doctoral Program in Affective Computing

Work submitted to the Université de Bourgogne Franche-Comté in partial fulfillment of the requirements for the degree of Doctor of Philosophy

JURY

President of the Jury:

Frédéric Morain-Nicolier, Professor at Université de Reims Champagne-Ardenne

Thesis Director:

Fabrice Mériaudeau, Professor at Université de Bourgogne

Reviewers:

Paul Honeine, Professor at Université de Rouen Normandie

Véronique Eglin, Professor at Institut National des Sciences Appliqué de Lyon

Examiners:

Fan Yang, Professor at Université de Bourgogne

Panagiotis Simos, Professor at University of Crete

Day of the defense: 8 November 2018

Signature from head of PhD committee:

Vultus est index animi

Face is the index of the soul

Publications

Peer-Reviewed Journal Articles - 1st Author

1. **Pampouchidou,A.**, Simos,P., Marias,K., Meriaudeau,F., Yang,F., Pediaditis,M., and Tsiknakis,M. "Automatic Assessment of Depression Based on Visual Cues: A Systematic Review", *IEEE Transactions on Affective Computing*.

2. **Pampouchidou,A.**, Pediaditis,M., Maridaki,A., Awais,M., Vazakopoulou,C.-M., Sfakianakis,S., Tsiknakis,M., Simos,P., Marias,K., Yang,F., and Meriaudeau,F. "Quantitative comparison of motion history image variants for video-based depression assessment", *EURASIP Journal on Image and Video Processing* "Special Issue on Applications of Visual Analysis of Human Behaviour", 2017(1), 64.

Peer-Reviewed International Conferences - 1st Author

1. **Pampouchidou,A.**, Simantiraki,O., Vazakopoulou,C.-M., Marias,K., Simos,P., Yang,F., Meriaudeau,F., Tsiknakis,M. "Détection de la dépression par l'analyse de la géométrie faciale et de la parole", *XXVIème colloque du Groupement de Recherche en Traitement du Signal et des Images* (GRETSI), Juan-Les-Pins, France. September 2017

2. **Pampouchidou,A.**, Simantiraki,O., Vazakopoulou,C.-M., Chatzaki,C., Pediaditis,M., Maridaki,A., Marias,K., Simos,P., Yang,F., Meriaudeau,F., Tsiknakis,M. "Facial Geometry and Speech Analysis for Depression Detection", *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC), IEEE, Jeju, Korea, July 12, 2017

3. **Pampouchidou,A.**, Simantiraki,O., Fazlollahi,A., Pediaditis,M., Manousos,D., Roniotis,A., Giannakakis,G., Meriaudeau,F., Simos,P., Marias,K., Yang,F., and Tsiknakis,M. "Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text", *Audio/Visual Emotion Challenge and Workshop* (AVEC 2016) "Depression, Mood and Emotion", ACM Multimedia, Amsterdam, The Netherlands, October 16, 2016

4. **Pampouchidou,A.**, Pediaditis,M., Chiarugi,F., Marias,K., Simos,P., Yang,F., Meriaudeau,F., Tsiknakis,M. "Automated Characterization of Mouth Activity for Stress and Anxiety Assessment". *IEEE International Conference on Imaging Systems and Techniques*, Chania, Crete Island, Greece, October 4-6, 2016, October 5, 2016

5. **Pampouchidou,A.**, Marias,K. Tsiknakis,M., Simos,P., Yang,F., Lemaître,G., Meriaudeau,F. "Video-Based Depression Detection Using Local Curvelet Binary Patterns in Pairwise Orthogonal Planes", *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC), Orlando, Florida USA, 16-20 August, 2016.

6. **Pampouchidou,A.**, Kazantzaki,E., Karatzanis,I., Marias,K., Tsiknakis,M., Meriaudeau,F., Yang,F., and Simos,P. "Preliminary Evaluation of a Web-Oriented Assessment Tool for Emotion Recognition". *13th International Conference on Wearable, Micro & Nano Technologies for Personalized Health*, FORTH, Heraklion, Crete, Greece, 29-31 May, 2016. Also appears in the journal Studies in Health Technology and Informatics 224 (2016): 95.
7. **Pampouchidou,A.**, Marias,K., Tsiknakis,M., Simos,P., Yang,F., and Meriaudeau,F. "Designing a Framework for Assisting Depression Severity Assessment from Facial Image Analysis". *International Conference on Signal and Image Processing Applications*, Kuala Lumpur, Malaysia, 19-21 October 2015.

Peer-Reviewed International Conferences - Contributing

1. Megat S'adan,M.A.H., **Pampouchidou,A.**, Meriaudeau,F. "Deep Learning Techniques for Depression Assessment", *International Conference on Intelligent & Advanced System*, Kuala Lumpur, Malaysia, 13-15 August, 2018
2. Maridaki,A., **Pampouchidou,A.**, Marias,K., Tsiknakis,M. "Machine Learning Techniques for Automatic Depression Assessment". *41st IEEE International Conference on Telecommunications and Signal Processing*, Athens, Greece, 4-6 July 2018
3. Bourou,D., **Pampouchidou,A.**, Tsiknakis,M., Marias,K., and Simos,P. "Video-based Pain Level Assessment: Feature Selection and Inter-Subject Variability Modelling", *41st IEEE International Conference on Telecommunications and Signal Processing*, Athens, Greece, 4-6 July 2018
4. Vazakopoulou,C-M., **Pampouchidou,A.**, Yang,F., Meriaudeau,F., Marias,K., and Tsiknakis,M. "Détection de la dépression par l'analyse de la géométrie faciale et apprentissage automatique", *Congrés National de la Recherche des IUT "CNRIUT'2018"*, Aix-en-Provence, 7-8 June 2018
5. Simantiraki,O., Charonyktakis,P., **Pampouchidou,A.**, Tsiknakis,M., and Cooke,M. "Glottal Source Features for Automatic Speech-based Depression Assessment", *INTERSPEECH*, Stockholm, Sweden, August 20-24 2017.
6. Simantiraki,O., Giannakakis,G., **Pampouchidou,A.**, and Tsiknakis,M. "Stress Detection from Speech Using Spectral Slope Measurements", *6th International Symposium on Pervasive Computing Paradigms for Mental Health*, EAI, Barcelona, Spain, 28-29 November 2016
7. Pediaditis,M., Giannakakis,G., Chiarugi,F., Manousos,D., **Pampouchidou,A.**, Christinaki,E., Iatraki,G., Kazantzaki,E., Simos,P.G., Marias,K., and Tsiknakis,M. "Extraction of Facial Features as Indicators of Stress and Anxiety". *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBS), Milan, Italy, August 25-29 2015.
8. Chiarugi,F., Iatraki,G., Christinaki,E., Manousos,D., Giannakakis,G., Pediaditis,M., **Pampouchidou,A.**, Marias,K., and Tsiknakis,M.N. "Facial signs and psycho-physical status estimation for well-being assessment". *7th IEEE International Conference on Health Informatics*, Angers, Loire Valley, France, 3-6 March 2014.

Acknowledgements

This has been a long project, with many challenges, but with much of support too. Beginning the acknowledgments I would like to thank the sources of my funding during the last five years. First, I would like to thank Prof. Tsiknakis and Prof. Marias, who supported me financially during the first two years of my PhD through partial funding from research and development projects executed at that time by the Computational BioMedicine Laboratory (CBML) in Foundation for Research and Technology - Hellas (FORTH). Next, I would like to thank the Greek State Scholarship Foundation, which funded me through the bequest "in the memory of Maria Zaousi", may God rest her soul.

Continuing I would like to thank my supervisors each one and separately. I will forever be grateful to Prof. Marias who matched me to this topic I came to adore, it was an insightful matching that kept my motivation high during these years. My sincere gratitude goes also to Prof. Tsiknakis, it was inspiring to see you working so hard and being active all the time. Prof. Simos I would like to thank you for teaching me so much and helping me gain some background in terms of the clinical aspects of my work, and for always being close to me and supporting me. Prof. Meriaudeau also for being supportive and close to me, as well as for giving me the opportunity to enter the PhD programme, and for guiding me and offering insights in terms of the machine learning and image processing fields. My sincere thank you to Prof. Yang, for always being supportive too.

Besides my supervisors I had support from fellow researchers, my postdoc-friends! Dr M.Pediaditis, Mat I will never say thank you enough, for the

discussions, for helping me in every step, for the brainstormings, for always being there willing and available any time I dropped by your office, once again thank you! My gratitude also goes to Dr S.Sfakianakis, who was also willing to discuss with me about issues I had doubt, especially regarding the field of statistics. Thank you both for your mentoring.

Continuing I would like to thank the people who supported me during the long term endeavor of the data collection project, it wouldn't have been possible, or at least would have been even more difficult without your help. Thus I would like to thank D.Manousos, G.Karatzanis, O.Simantiraki, G. Zacharioudakis, and C.Chatzaki for their technical support. My sincere gratitude goes also to the psychologists that executed the protocol and interviewed the participants during the recordings, I.Apostolaki, E.Kazantzaki, E.Papastefanakis, along with the psychologists who carried the annotation of the video recordings K.Argiraki and M.Daskalaki, thank you all so much!

Next I would like to thank the master students I co-advised through these years for their co-operation. C.-M.Vazakopoulou, A.Maridaki, and A.-D. Mpourou. Thank you all for the co-operation, I also learned from you too. I am also thankful for an unexpected co-operation, this with my old friend Dr A.Fazlollahi, it was great to have the chance to work together again!

V.Maniadi and A.Vergetaki thank you for your support in any administration related issues, but also for your valuable friendship. My friends K.Spanakis, G.Christodoulakis, and K.Nikiforaki, with whom we were the "early" birds in the office, thank you for sharing mornings and thoughts! C.Hernadez Matas, I would like to thank you for sharing the same concerns during the PhD struggle! My dearest friend Koula thank you for always being there for me!

My academic journey has never been a straight line, and there were many people whom I feel the need to thank, as I believe they all had some contribution to me reaching to this point. My highschool teachers Argyris Nikolaidis, who saw my talent in programming and encouraged me to follow computer science when I was going for English literature, but also R.Tsaknaki for giving me strong programming skills. I would also like to thank Prof. Papadourakis for recruiting me as his student assistant during my undergrad studies, and

introducing me to the academic world. My gratitude to all my professors in the Master Computer Vision that taught me how to extend my limits, accelerate, and improve. More importantly I want to thank Prof. Fougerolle for always believing in me against all odds, and encouraging me to continue!

Closing this section I would also like to thank my mother for teaching me how to be hard working and a fighter in this life! My sincere gratitude goes also to our babysitter Athina, thank you so much for taking great care of our children during the endless hours of work! Last but not least, I would like to thank my family, my loving husband Giannis, who stood by me all the way since I started my Master degree, standing by me until now I reached to the PhD, supporting me in every possible way! Finally, I would like to thank our beautiful children Galatea and Ellie, for their patience and understanding for missing from home a lot. This beautiful family helped me keep my sanity along this way!

Abstract

Depression is the most prevalent mood disorder worldwide having a significant impact on well-being and functionality, and important personal, family and societal effects. The early and accurate detection of signs related to depression could have many benefits for both clinicians and affected individuals. The present work aimed at developing and clinically testing a methodology able to detect visual signs of depression and support clinician decisions.

Several analysis pipelines were implemented, focusing on motion representation algorithms, including Local Curvelet Binary Patterns-Three Orthogonal Planes (LCBP-TOP), Local Curvelet Binary Patterns- Pairwise Orthogonal Planes (LCBP-POP), Landmark Motion History Images (LMHI), and Gabor Motion History Image (GMHI). These motion representation methods were combined with different appearance-based feature extraction algorithms, namely Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Local Phase Quantization (LPQ), as well as Visual Graphic Geometry (VGG) features based on transfer learning from deep learning networks. The proposed methods were tested on two benchmark datasets, the AVEC and the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ), which were recorded from non-diagnosed individuals and annotated based on self-report depression assessment instruments. A novel dataset was also developed to include patients with a clinical diagnosis of depression (n=20) as well as healthy volunteers (n=45).

Two different types of depression assessment were tested on the available datasets, categorical (classification) and continuous (regression). The MHI with VGG for the AVEC'14 benchmark dataset outperformed the state-of-the-art with 87.4% F1-Score for binary categorical assessment. For continuous assessment of self-reported depression symptoms, MHI combined with

HOG and VGG performed at state-of-the-art levels on both the AVEC'14 dataset and our dataset, with Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) of 10.59/7.46 and 10.15/8.48, respectively. The best performance of the proposed methodology was achieved in predicting self-reported anxiety symptoms in our dataset, with RMSE/MAE of 9.94/7.88.

Results are discussed in relation to clinical and technical limitations and potential improvements in future work.

Résumé

La dépression est le trouble de l'humeur le plus répandu dans le monde avec des répercussions sur le bien-être personnel, familial et sociétal. La détection précoce et précise des signes liés à la dépression pourrait présenter de nombreux avantages pour les cliniciens et les personnes touchées. Le présent travail visait à développer et à tester cliniquement une méthodologie capable de détecter les signes visuels de la dépression afin d'aider les cliniciens dans leur décision.

Plusieurs pipelines d'analyse ont été mis en œuvre, axés sur les algorithmes de représentation du mouvement, via des changements de textures ou des évolutions de points caractéristiques du visage, avec des algorithmes basés sur les motifs binaires locaux et leurs variantes incluant ainsi la dimension temporelle (Local Curvelet Binary Patterns-Three Orthogonal Planes (LCBP-TOP), Local Curvelet Binary Patterns- Pairwise Orthogonal Planes (LCBP-POP), Landmark Motion History Images (LMHI), and Gabor Motion History Image (GMHI)). Ces méthodes de représentation ont été combinées avec différents algorithmes d'extraction de caractéristiques basés sur l'apparence, à savoir les modèles binaires locaux (LBP), l'histogramme des gradients orientés (HOG), la quantification de phase locale (LPQ) et les caractéristiques visuelles obtenues après transfert de modèle issu des apprentissage profonds (VGG). Les méthodes proposées ont été testées sur deux ensembles de données de référence, AVEC et le Wizard of Oz (DAIC-WOZ), enregistrés à partir d'individus non diagnostiqués et annotés à l'aide d'instruments d'évaluation de la dépression. Un nouvel ensemble de données a également été développé pour inclure les patients présentant un diagnostic clinique de dépression (n = 20) ainsi que les volontaires sains (n = 45).

Deux types différents d'évaluation de la dépression ont été testés sur les ensembles de données disponibles, catégorique (classification) et continue (régression). Le MHI avec VGG pour l'ensemble de données de référence AVEC'14 a surpassé l'état de l'art avec un F1-Score de 87,4% pour l'évaluation catégorielle binaire. Pour l'évaluation continue des symptômes de dépression "autodéclarés", LMHI combinée aux caractéristiques issues des HOG et à celles issues du modèle VGG ont conduit à des résultats comparatifs aux meilleures techniques de l'état de l'art sur le jeu de données AVEC'14 et sur notre ensemble de données, avec une erreur quadratique moyenne (RMSE) et une erreur absolue moyenne (MAE) de 10,59 / 7,46 et 10,15 / 8,48 respectivement. La meilleure performance de la méthodologie proposée a été obtenue dans la prédiction des symptômes d'anxiété auto-déclarés sur notre ensemble de données, avec une RMSE/MAE de 9,94 / 7,88.

Les résultats sont discutés en relation avec les limitations cliniques et techniques et des améliorations potentielles pour des travaux futurs sont proposées.

List of Abbreviations

AAM Active Appearance Models

AFCL After Fully Connected Layer

APA American Psychiatric Association

ASM Active Shape Models

AU Action Units

AVEC Audio/Visual Emotion Challenge

BDI Beck's Depression Inventory

BDI-II Beck's Depression Inventory II

BFCL Before Fully Connected Layer

BoW Bag-of-Words

BW-LBP-TOP Block-Wise LBP-TOP

CCA Canonical Correlation Analysis

CFS Correlation-based Feature Selection

CHMM Coupled Hidden Markov Model

CLM Constraint Local Models method

CNN Convolutional Neural Networks

CRF Cascade Random Forest

DAIC-WOZ Distress Analysis Interview Corpus - Wizard of Oz

DBS-SCC Deep Brain Stimulation of the Subcallosal Cingulate Cortex

DCNN Deep Convolutional Neural Networks

DCNN-DNN Deep Convolutional Neural Netowork - Deep Neural Network

DCS Divergence-Curl-Shear

DM Discriminative Mapping

DNN Deep Neural Networks

DoG Difference of Gaussians

DSC Depression Recognition Sub-challenge

DSM-5 Diagnostic and Statistical Manual of Mental Disorders of the (APA) fifth edition

DSSS Depression and Somatic Symptoms Scale

DTL Deep Transformation Learning

ELM Extreme Learning Machines

FC Fully Connected

FDHH Feature Dynamics History Histogram

fps frames per second

GLM Generalized Linear Models

GMHI Gabor Motion History Image

GMM Gaussian Mixture Models

GSR Gaussian Staircase Regression

HAM-D Hamilton Depression Rating Scale

HCRF Hidden Conditional Random Fields

HDR Histogram of Displacement Range

HMM Hidden Markov Model

HOG Histogram of Oriented Gradients

IAPS International Affective Picture System

ILSVRC ImageNet competition

k**NN** k-Nearest Neighbour

LBP Local Binary Patterns

LBP-TOP Local Binary Patterns in Three Orthogonal Planes

LCBP-POP Local Curvelet Binary Patterns- Pairwise Orthogonal Planes

LCBP-TOP Local Curvelet Binary Patterns-Three Orthogonal Planes

LDA Linear Discriminant Analysis

LGBP-TOP Local Gabor Binary Patterns in Three Orthogonal Planes

LMHI Landmark Motion History Images

LMM Landmark Motion Magnitude

LogR Logistic Regression

LOO Leave-One-Out

LOSO Leave-One-Subject-Out

LPQ Local Phase Quantization

LPQ-TOP Local Phase Quantization-Three Orthogonal Planes

LR Linear Regression

LSTM-RNN Long Short Term Memory Recurrent Neural Networks

MAE Mean Absolute Error

MaxEnt Maximum Entropy Model

MBH Motion Boundary Histogram

MDD Major Depressive Disorder

MHH Motion History Histograms

MHI Motion History Images

MIM Mutual Information Maximization

MINI Mini International Neuropsychiatric Interview

MPGI Moore-Penrose Generalized Inverse

mRMR minimum Redundancy Maximum Relevance

NB Naïve Bayes

NNet Neural Networks

OLS Ordinary Least Squares

ORI Oregon Research Institute

PCA Principal Components Analysis

PHQ Patient Health Questionnaire

PLS Partial Least Square Regression

PTSD Post Traumatic Stress Disorder

QIDS-SR Quick Inventory of Depressive Symptomatology-Self Report

ReLU Rectified Linear Unit

RF Random Forest

RFR Random Forest Regressor

RMSE Root Mean Squared Error

RVM Relevance Vector Machines

S.D. Standard Deviation

SDA Stacked Denoising Autoencoders

Self-RIs self-report scales and inventories

SGD Stochastic Gradient Descent

STAI State Trait Anxiety Inventory

STFT short-term Fourier transform

STIP Space-Time Interest Points

SVM Support Vector Machines

SVR Support Vector Regression

VFE Variability of Facial Expression

VGG Visual Graphic Geometry

VJ Viola & Jones

VLBP Volume Local Binary Pattern

VM Virtual Machine

WHO World Health Organization

List of Figures

1.1 Prevalence of depressive disorders as reported by WHO 2

2.1 Number of relevant studies per year 11
2.2 Typical workflow for automatic depression assessment 15
2.3 CLM fitted on a facial image 19
2.4 Taxonomy of visual features for depression assessment 21
2.5 Ranking of classification algorithms employed in relevant studies 24
2.6 Ranking of regression algorithms employed in relevant studies 25

3.1 VJ face detection: Haar features applied on a facial image 41
3.2 Examples of landmarks on a facial image 42
3.3 LBP from Three Orthogonal Planes 44
3.4 Ridgelets' example 45
3.5 Curvelet tiling of space and frequency 45
3.6 Example of curvelet pseudo-images wrapped 48
3.7 Curvelet transform on facial video and X-Z and Y-Z planes 49
3.8 Visual comparison of motion history image variants 50
3.9 Example for Gabor inhibited filtering vs Gabor filtering 54
3.10 Geometric features considered in time-series 56
3.11 Example of LBP pipeline 58
3.12 Illustration of LBP patterns 58
3.13 Example of the LPQ pipeline 59
3.14 Example of HOG pipeline 60
3.15 Visualization of appearance-based features 61
3.16 Architecture of VGG16 62

3.17 Visualization of Pool Relu1_1 activations 63
3.18 Visualization of Pool 5 activations 64
3.19 Example of time-series from eye region distances 65
3.20 Example of kNN 67
3.21 Example of RF 68
3.22 Example of SVM 69

4.1 Experiment 1: LCBP-TOP on eye region 74
4.2 Experiment 2: Pipeline for the LCBP-POP 76
4.3 Experiment 3: LMHI 78
4.4 Experiment 5: Flow of the motion history variants methodology 82

5.1 Average participant ratings of Sad video clips 93
5.2 Average pilot study participant ratings of Fear and Disgust video 94
5.3 Data collection experimental setup 97
5.4 Dispersion of BDI-II, STAI, and expert judgment values per group 99
5.5 Bivariate scatter plots of STAI, BDI-II and expert annotations across participant groups. 100
5.6 Average performance of the categorical assessment 103
5.6 Average performance of the categorical assessment **(cont.)** 104
5.7 Average normalized RMSE 107
5.8 Continuous assessment: Prediction of BDI-II (gender-based mode) 108
5.9 Continuous assessment: Prediction of BDI-II (gender-independent mode) 109
5.10 Continuous assessment: Prediction of STAI (gender-based mode) 110
5.11 Continuous assessment: Prediction of STAI (gender-independent mode) . 111
5.12 Continuous assessment: Prediction of expert judgment of depression (gender-based mode) 112
5.13 Continuous assessment: Prediction of expert judgment of depression (gender-independent mode) 113
5.14 Comparison of our approach (Proposed) with previously reported results on the AVEC'2014 115

6.1 Posterior probability classification model for fusion 125

List of Tables

1.1 Non-verbal manifestations of depression 5

2.1 Search terms and web-resources employed for literature review 10
2.2 Datasets reported in relevant studies 14
2.3 Classification algorithms employed in relevant studies 23
2.4 Regression algorithms employed in relevant studies 25
2.5 Comparison of approaches for categorical assessment of depression 30
2.5 Comparison of approaches for categorical assessment of depression **(cont.)** 31
2.5 Comparison of approaches for categorical assessment of depression **(cont.)** 32
2.6 Other approaches for categorical assessment of depression 34
2.7 Summary of studies employing continuous depression assessment 35
2.7 Summary of studies employing continuous depression assessment **(cont.)** 36
2.7 Summary of studies employing continuous depression assessment **(cont.)** 37

4.1 Results of Experiment 1 for four levels of depression 74
4.2 Experiment 1: Confusion matrix for best result (4 classes) 75
4.3 Experiment 2 results (%) 77
4.4 Experiment 2: Multi-class 77
4.5 Experiment 3 results: gender-dependency 79
4.6 Experiment 3: Comparison of the proposed to the challenge baseline . . . 80
4.7 Experiment 4: Experimental results(F1-Score) 80
4.8 Experiment 5: Experimental results of appearance-based descriptors . . . 81
4.9 Experiment 5: Experimental results of VGG features 83
4.10 Experiment 5: Confusion matrix for the best result (2 classes) 83
4.11 Experiment 5: Different performance metrics for the best result 83

4.12 Experiment 5: Performance comparison with the literature 84
4.13 Summary of achievements during preliminary experiments 85

5.1 Summary of work toward data collection 89
5.2 Socio-demographics and clinical data for both groups, control (n=45) and patients (n=20) 90
5.3 Data collection protocol in the main study 96
5.4 Camera Specifications 96
5.5 Best-performing classification schemes 104
5.6 Best-performing continuous assessment schemes 106
5.7 Best-performing continuous assessment schemes for the AVEC'14 dataset 114
5.8 Comparison of specifications between AVEC and our dataset 116

A.1 Complete socio-demographics participants profiles 131
A.1 Complete socio-demographics participants profiles **(cont.)** 132
A.1 Complete socio-demographics participants profiles **(cont.)** 133

Contents

List of Abbreviations **xix**

List of Figures **xxv**

List of Tables **xxvii**

1 Introduction **1**

1.1 Depression Facts 1

1.2 Depression Assessment 3

1.2.1 Clinical diagnosis & Self-Reports 3

1.2.2 Nonverbal Signs of Depression 4

1.3 Research Motivation 6

1.4 Thesis Outline 8

2 Literature Review **9**

2.1 Review Design 9

2.2 Depression Datasets 10

2.2.1 Data Collection Procedure 11

2.2.2 Reported Datasets 12

2.3 State-of-the-Art Methods 15

2.3.1 Preprocessing 16

2.3.2 Feature Extraction / Manipulation 16

2.3.2.1 Feature Extraction 16

2.3.2.2 Dimensionality Reduction 20

2.3.3 Machine Learning 22

2.3.3.1 Cross Validation Methods 22

2.3.3.2 Classification . . . 23
2.3.3.3 Regression . . . 24
2.4 Selected Approaches and Meta-Analysis . . . 24
2.4.1 Categorical Depression Assessment . . . 26
2.4.1.1 Pittsburgh . . . 28
2.4.1.2 BlackDog . . . 28
2.4.1.3 DAIC-WOZ . . . 29
2.4.1.4 AVEC . . . 29
2.4.1.5 Other Datasets . . . 29
2.4.2 Continuous Depression Assessment . . . 33
2.5 Scope of the Thesis . . . 33

3 Methodology **39**
3.1 Preprocessing . . . 39
3.1.1 Illumination Normalization . . . 39
3.1.2 Face Detection . . . 40
3.1.3 Facial Landmarks Detection . . . 41
3.2 Motion Representation . . . 43
3.2.1 Local Curvelet Binary Patterns . . . 43
3.2.1.1 Curvelet Transform . . . 44
3.2.1.2 Three Orthogonal Planes vs Pairwise Orthogonal Planes 47
3.2.2 Motion History Image . . . 50
3.2.2.1 Original MHI . . . 51
3.2.2.2 Landmark Motion History Image . . . 52
3.2.2.3 Gabor Motion History Image . . . 53
3.2.3 Time-series of Geometrical Features . . . 55
3.3 Feature Extraction . . . 57
3.3.1 Appearance-based Descriptors . . . 57
3.3.1.1 Local Binary Patterns . . . 57
3.3.1.2 Local Phase Quantization . . . 59
3.3.1.3 Histogram of Oriented Gradients . . . 60
3.3.1.4 Image Histogram . . . 60
3.3.2 Transfer Learning from Pretrained Networks . . . 61

3.3.3 Statistical Descriptors . . . 65
3.4 Machine Learning . . . 66
3.4.1 Dimensionality Reduction with PCA . . . 66
3.4.2 Cross Validation . . . 66
3.4.3 Gender Dependency . . . 67
3.4.4 kNN . . . 67
3.4.5 Random Forest . . . 68
3.4.6 SVM . . . 69
3.5 Fusion . . . 70
3.6 Summary . . . 70

4 Preliminary Experimental Evaluation **71**
4.1 Employed Datasets . . . 71
4.1.1 AVEC . . . 72
4.1.2 DAIC . . . 73
4.2 Experiment 1 . . . 73
4.3 Experiment 2 . . . 75
4.4 Experiment 3 . . . 78
4.5 Experiment 4 . . . 79
4.6 Experiment 5 . . . 80
4.7 Summary . . . 84

5 Main Experiment **87**
5.1 Data Collection . . . 87
5.1.1 Ethics and Data Protection . . . 88
5.1.2 Participants . . . 88
5.1.3 Psychological measurements and experimental procedures . . . 90
5.1.3.1 Tools assessing depression-related symptoms . . . 91
5.1.3.2 Emotion elicitation paradigm . . . 91
5.1.4 Technical Setup . . . 95
5.1.5 Blinded Expert Annotation . . . 98
5.1.6 Statistical Results . . . 101
5.2 Experimental Tests . . . 101
5.2.1 Experimental Setup . . . 102

5.2.2 Categorical Assessment 102
5.2.3 Continuous Assessment 105

6 Discussion **117**
6.1 Algorithm Development and Performance 117
6.2 Data Related Issues 120
6.3 Future Plans 124

7 Conclusions **127**
7.1 Research Questions Addressed 127
7.2 Major Contributions 129

Appendices **131**

A Complete List socio-demographics participants profiles **131**

References **135**

Chapter 1

Introduction

The present chapter presents concepts and facts related to depression, in order to acquaint the reader with the disorder. Methods for depression assessment are summarized, including diagnostic procedures and self-reports. Non-verbal signs of depression, along with the motivation of the work described herein, are also presented. Parts of the contents of this chapter have been published in [165].

1.1 Depression Facts

Depression is the most common mood disorder characterized by persistent negative affect [17] (c.f. Fig. 1.1). Clinically distinct depressive disorders encompass a wide range of manifestations. According to the Diagnostic and Statistical Manual of Mental Disorders of the American Psychiatric Association (APA) [28], now in its fifth edition (DSM-5), subtypes of depressive disorders include: Major Depressive Disorder (MDD), Persistent Depressive Disorder, Disruptive Mood Dysregulation Disorder, Premenstrual Dysphoric Disorder, Substance/Medication-Induced Depressive Disorder, Depressive Disorder Due to Another Medical Condition, and Other Specified Depressive Disorder or Unspecified Depressive Disorder.

MDD, also commonly referred to as Clinical Depression, is considered as the most typical form of the disease [217]. According to DSM-5 MDD can be diagnosed by the presence of a) depressed mood most of the day, and/or b) markedly diminished interest or pleasure, combined with at least four of the following symptoms for a period exceeding two weeks: significant weight change of over 5% in a month, sleeping disturbances (in-

somnia or hypersomnia), psychomotor agitation or retardation almost every day, fatigue or loss of energy almost every day, feelings of worthlessness or excessive guilt, diminished ability to concentrate or indecisiveness almost every day, recurrent thoughts of death or suicidal ideation. An additional common feature of all depressive disorders is "(...) *clinically significant distress or impairment in social, occupational, or other important areas of functioning* (...)" [28].

MDD is reported to be the fourth most prominent cause of disability and is expected to become the second by 2020 due to its increasing prevalence [127]. The "Survey of Health, Ageing and Retirement in Europe" [14] documents a consistent rise of depression among adults with increasing age, which is associated with significantly elevated risk for suicidal behavior [213]. The ongoing economic crisis in Europe resulting in high unemployment is implicated as a trigger, since 70-76% of unemployed people have been reported to display significant depressive symptomatology [82]. Further studies have shown that the economic burden of MDD has increased during the 2005-2010 period by 21.5% in the US, while in Europe the cost is estimated at 1% of Gross Domestic Product [190]. The total cost of MDD in 2010 in 30 European countries was estimated at 91.9 billion euros [155].

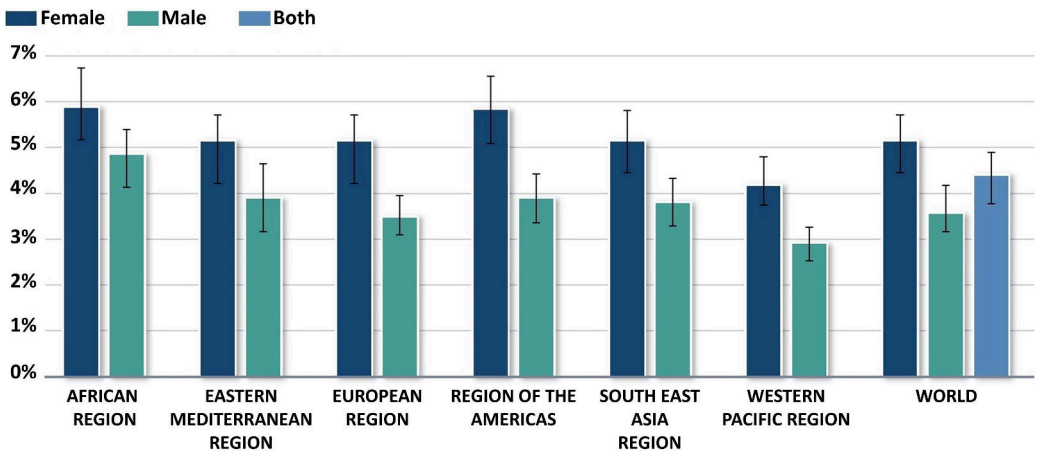


Figure 1.1: Prevalence of depressive disorders (% of population), by WHO Region (Taken from [229])

1.2 Depression Assessment

Various procedures for depression assessment are presented in this section, while non-verbal signs of depression are also described, emphasizing visual signs.

1.2.1 Clinical diagnosis & Self-Reports

A structured clinical interview assessing the presence of DSM-5 criteria is the standard procedure for depression diagnosis [80]. Quantification of the presence and severity of depressive symptomatology is often aided by rating scales completed by a specially trained mental health professional in the context of the clinical interview. The Hamilton Depression Rating Scale (HAM-D) is one of the most popular scales in clinical settings. HAM-D assesses the severity of 17 symptoms, such as depressed mood, suicidal ideation, insomnia, work and interests, psychomotor retardation, agitation, anxiety, and somatic symptoms [97]. Both HAM-D and DSM-5 clinical criteria have been criticized regarding their reliability [30] [49], as diagnosis of MDD is not as consistent as other common medical conditions [129]. In general, "*there is no blood test*" for depression [202], as the disorder lacks biological gold standards [124].

Clinical diagnosis of depression may also be supported by scores on self-report scales and inventories (Self-RIs). Most often used Self-RIs in affective computing research are the various forms of Patient Health Questionnaire (PHQ)-2/8/9, comprised of 2, 8, or 9 items, respectively and Beck's Depression Inventory (BDI). The Depression and Somatic Symptoms Scale (DSSS) was also used in one study. Self-RIs are convenient and economical, with reported sensitivity and specificity approaching 80-90% (e.g., PHQ-9 [224]), but bear certain disadvantages. Importantly, they do not take into account the clinical significance of reported symptoms, and do not permit adjustments for individual trait characteristics, other psychiatric and medical comorbidities, and potentially important life events, as opposed to a clinical interview [168]. Additionally Self-RIs are vulnerable to intentional (such as norm defiance) or unintentional reporting bias (e.g., subjective, central tendency [i.e., avoiding extreme responses], social desirability, and acquiescence) [35]. In summary, although Self-RIs alone are insufficient to support the diagnosis of depression [171] [193], they are widely used for screening purposes in various settings, including primary health care. While the cost-effectiveness of widespread screening practices for improving the quality of depression care is debated [149], practical issues related

to the aforementioned limitations of Self-RIs raise questions regarding the overall utility and effectiveness of this practice for population-based mental health.

1.2.2 Nonverbal Signs of Depression

It is well known that depression manifests through a variety of nonverbal signs [76] [222]. Involuntary changes in the tonic activity of facial muscles, as well as changes in peripheral blood pressure and skin electrodermal response, often mirror the frequent and persistent negative thoughts and feelings of sadness that characterize depression. Preliminary findings suggest that electroencephalographic recordings may contain features related to depression [104]. Functional Near-Infrared Spectroscopy has also attracted interest [19] [199]. Additionally, speech conveys non-verbal information on the mental state of the speaker; prosodic, source, and acoustic features, as well as vocal tract dynamics, are speech-related features affected by depression [60]. Furthermore, depression as a mood disorder, is portrayed on the individual's appearance, in terms of facial expression, as well as body posture [76] [222]. Face as a whole, and individual facial features, such as eyes, eyebrows or mouth, are of particular interest when it comes to depression assessment. Some of the visual signs identified in the literature are briefly described in the paragraphs that follow.

Specific facial expressions, have been widely examined for depression assessment, in terms of frequency of occurrence, variability, and intensity of a specific expression. Typically, the facial expression classification system proposed by Ekman [75] is employed, which includes a set of six basic emotions (joy, surprise, anger, sadness, disgust, fear). Measuring the frequency of occurrence of each of the six emotional expressions [237] [84] [195] [176] relies on the premise that depressed individuals tend to show reduced expressivity [76]. Other studies focused on specific facial features, such as the eyes and mouth. These include gaze direction [177] [178], reduced eye contact [135], eyelid activity [22], eye openings/blinking [237] [96] and iris movement [24]. Smile intensity [177] [178], smile duration [177] [178], mouth animation [96], listening smiles (smiles while not speaking) [93], and lack of smiles [135] also constitute potentially useful facial signs for assessing depression.

Head pose, orientation, and movement have been used extensively for depression assessment [237] [84] [195] [90] [176] [239] [96] [23] [118], along with motor variability and general facial animation [195] [176] [96]. Another visual sign that has drawn considerable

attention by clinicians in relation to depression assessment is pupil dilation. Siegle et al. [185] reported faster pupillary responses in non-depressed individuals to positive as compared to negative stimuli. More recently Price et al. [169] investigated attentional bias, including pupil bias and diameter, to predict depression symptoms over a two year follow up period in a sample of adolescents displaying high ratings of anxiety. Additionally, body gestures [114] [116] involving the entire body, upper body, or separate body parts can also contribute to the assessment. Finally, shaking and/or fidgeting behavior, self-adaptors, and foot tapping [93] have also been considered as signs of depression. Table 1.1, taken from Pampouchidou et al. [165], is summarizing signs and signals related to depression assessment as found in the literature to date.

Table 1.1: Non-verbal manifestations of depression

Eyelid activity (openings, blinking)
Eye gaze (limited & shorter eye contact)
Visual fixation
Low frequency & duration of glances
Eyebrow activity
"Veraguth fold"*
Frowns
Fewer smiles
More frequent lip presses
Smile intensity & duration
Mouth corners angled down
Mouth animation
Listening smiles (smiling while not speaking)
Facial expression occurrence (variability & intensity)
Sad / negative / neutral expression occurrence
Head pose (orientation, movement)
Body gestures (full or upper body, or body parts)
Slumped posture
Limp & uniform body posture
Reduced & slowed arm and hand movements
Shaking and/or fidgeting behavior
Self-adaptors
Foot tapping
Motor variability

* "veraguth fold" is a fold (wrinkle) of skin on the upper eyelid, between the eyebrows

1.3 Research Motivation

Recent classification schemes (e.g., DSM-5) run the risk of confusing normal sadness (e.g., bereavement) with depression, raising the likelihood of false positive diagnoses [218]. Depression assessment is a complex process and diagnosis is associated with a significant degree of uncertainty, given the lack of objective boundaries, and the need to evaluate symptoms within the person's current psychosocial context and past history [181]. Diagnostic accuracy typically improves when results from successive clinical assessments, performed over several months, are taken into account [150]. Importantly, a simple "*symptom checklist*" approach is severely limited and diagnosis requires considerable time investment in order to develop rapport with the patient [202]. The validity and clinical significance of strict classification schemes has also been questioned [142]. For instance, MDD has been questioned as a "*homogeneous categorical entity*" [91] and the notion of a "*continuum of depressive disorders*" is often advocated [138]. These reasonable concerns go beyond the scope of the present work, given that currently affective computing research relies heavily upon established clinical practice tools and procedures.

Objective measures of psychoemotional state, which are implicitly desirable in clinical and research applications alike [103] [88], could complement Self-RIs and help overcome some of their shortcomings. Certain Self-RIs are sufficiently brief and can be completed on a regular basis (e.g., monthly or weekly) as part of electronic platforms designed to support long-term monitoring of persons at risk. As suggested by Girard and Cohn [87], technological advances in the field have paved the way for viable automated methods for measuring signs of depression with multiple potential clinical applications. Thus, decision support systems capable of capturing and interpreting nonverbal depression-related cues, combined with verbal reports (Self-RIs), could be valuable in both clinical and research applications. In principle, such measures may reduce or even eliminate report bias. In addition, such measures are minimally invasive and do not require extra effort on the part of the respondent, thus likely to increase long-term compliance.

Apart from the high prevalence of MDD, and the complexity of the diagnosis, an additional motivation to pursue the development of a methodology for such a decision support system, is the high number of underdiagnosed depressive episodes. In the overall 85% of depressed individuals are underdiagnosed [77]. Relevant research [29] also showed that about 30% of patients suffering from an episode of major depression do not

seek treatment, with eventually only 10% of them being adequately treated. In addition, applying technological innovations could enhance accessibility to mental healthcare by overcoming traditional barriers [54]. Current technological means can provide the infrastructure for monitoring psychoemotional state in high-risk individuals as part of early detection and/or relapse prevention programs. A system devoted to the assessment of depressive symptomatology based on visual cues could provide reliable indices, partly based on facial expression analysis, in an unobtrusive manner.

Furthermore, the widespread and relatively low-cost accessibility to computer and internet technologies, webcams, and smart phones, renders an efficient system for depression assessment viable. Practical issues involved in developing such a system, like the storage of sensitive data, could become an issue if not handled properly, but there are ways to tackle such challenges; encryption, protection by password, or even an authorization procedure could be implemented to regulate access to sensitive personal data.

In parallel, a great number of Web-based tools for depression management have been developed and used clinically displaying a high degree of acceptance and patient adherence [41]. Early attempts for internet-based interventions for prevention and treatment of depression have shown promising outcomes [144], as well as mobile-based interventions that help in reducing relevant symptomatology [112]. In the overall telepsychiatry promotes patient-centered care [102].

Currently, video-based systems for depression assessment have only been found in research-related projects, and have not been applied in the general population to evaluate their feasibility. Although currently limited to research applications, the field has been very popular, with a dedicated section within the Audio/Visual Emotion Challenge (AVEC). AVEC'13 had three papers accepted for the Depression Recognition Subchallenge (DSC) [209] [208], while AVEC'14 [210] respectively attracted 44 submissions by 13 teams worldwide; the AVEC'16 attracted submissions from 7 teams for the DSC [211], while the latest DSC in terms of AVEC'17 attracted submissions by 10 teams. Besides from being an active field drawing broad interest, AVEC submissions document the sheer number of research groups working towards the development of such methods. This fact implies that the idea of developing automated depression assessment methods is not only promising, but is continuously progressing towards more robust and reliable measures.

However, given the current state of the art, video-based systems for depression assessment are not intended as standalone tools, but mainly to function as a decision support systems assisting mental health professionals in the monitoring of persons at risk. "Behaviormedics" is the term Valstar [206] introduced for applications designed for the automatic analysis of affective and social signals, among others, to support objective diagnosis. Finally, the development of tools to assist clinicians in the diagnosis and monitoring response to treatment and illness progression is gradually being supported by clinical studies [50]. For the purpose of the present thesis, the term "*depression assessment*" will refer to the process of detecting and assessing the severity of signs of and/or presence of depression.

1.4 Thesis Outline

The current thesis is organized in six Chapters. Chapter 2 exhaustively reviews the state-of-the-art, presenting relevant methods employed in the literature, as well as a meta-analysis of existing approaches. Methodology developed in terms of this PhD is described in Chapter 3, with preliminary evaluation of the different methods being presented in Chapter 4. Chapter 5 presents the methods used for collecting the clinical data, in terms of technical setup, protocol, and participants, along with the main experiment conducted as part of the current study. Discussion of experimental results and future work takes place in Chapter 6. Finally, Chapter 7 presents the conclusions of this work.

Chapter 2

Literature Review

The present chapter is a systematic review of existing methods for automatic detection and/or severity assessment of depression. In the current exhaustive review of more than eighty studies, technical details, potential limitations of each approach and classification accuracy achieved are evaluated, focusing on image processing and machine learning algorithms applied to depression detection. State of the art methods are presented highlighting their advantages and limitations, based on a quantitative meta-analysis of their results. Datasets created to serve the various studies and the corresponding data acquisition protocols are also described and discussed. Ultimately, gaps in the literature are identified to be addressed in terms of the proposed thesis. Parts of this Chapter have been published in IEEE Transactions on Affective Computing (Pampouchidou et al. 2017) [165].

2.1 Review Design

The technical report entitled "*Procedures for Performing Systematic Reviews*" by Kitchenham [128] was used as a guide for the present review. The keywords used to search electronic databases and related resources are listed in Table 2.1. Inclusion criteria for the review involved a) adequate description of an algorithm for automatic depression assessment utilizing visual cues and b) presentation of systematically derived data, producing concrete results. Fig.2.1 illustrates the rapid increase of relevant studies during the past few years proving that automatic depression assessment based on visual cues is a rapidly growing research domain. The small drop in 2015 can be attributed to the

Table 2.1: Search terms and web-resources employed in the current literature review

Keywords	Web-resources
• Depression • Facial Expression • Non-verbal communication • Image Processing • machine learning • Biomedical Imaging • Face • Emotion • Computer Vision	• ACM Digital Library [1] • IEEE Xplore Digital Library [4] • Elsevier [2] • Springer [10] • Wiley Online Library [12] • NASA [6] • Oxford University Press [7] • US National Library of Medicine [8] • Scopus [9] • Google Scholar [3] • Medpilot [5]
• Depression • Definition • Types • Frequency or rate • Diagnostic tests • Etiology and risk factors • Predictability	• World Health Organization [17] • Survey of Health, Ageing and Retirement in Europe [14] • Mayo Clinic [13] • National Comorbidity Survey [16] • World Health Organization - Regional Office for Europe [15]

Note: The first row contains keywords and web-resources that were canvassed to identify relevant approaches. Elements pertaining to the clinical relevance of studies are listed in the bottom row.

fact that there was no Depression sub-challenge in the 2015 AVEC; similarly, the sharp rise of interest in 2013, 2014, 2016, and 2017 can be attributed to the respective AVEC challenges.

2.2 Depression Datasets

The availability of empirical data is of paramount importance for the evaluation of methods for automatic assessment of depression. Such data are critical during algorithm development and testing. Due to the sensitive nature of clinical data, availability is neither wide nor free, and this is the reason that most research groups resort to generating their own data sets. This section describes procedures used for data collection and derived datasets found in the reviewed studies (c.f. Table 2.2).

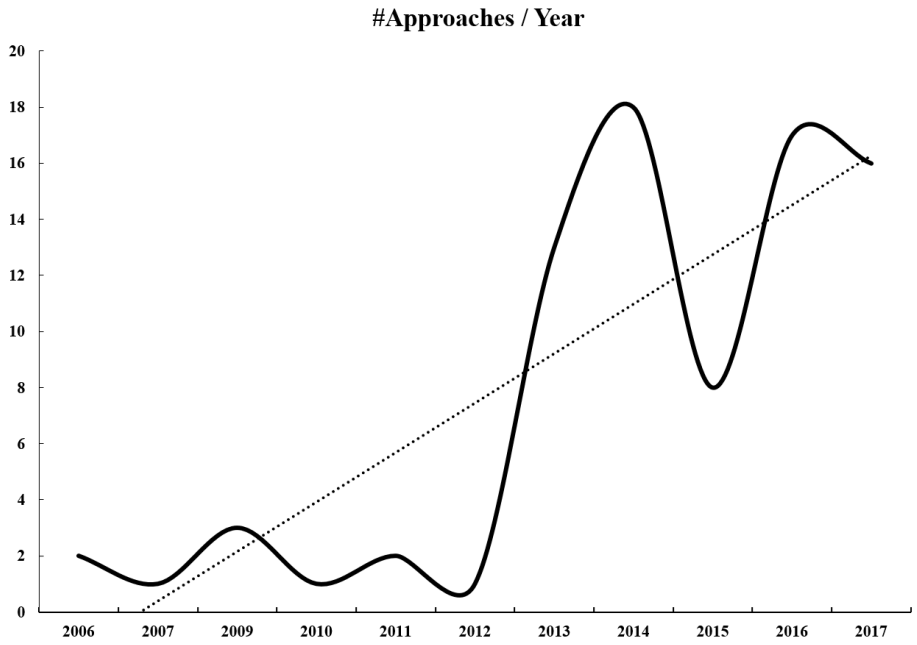


Figure 2.1: Number of studies in the field of depression assessment per year of publication

2.2.1 Data Collection Procedure

Recruitment of participants is perhaps the most challenging step in this line of research. Patients with MDD can be recruited from the community, in many cases by clinical psychologists or social workers, assessed using DSM-IV [27] criteria [185] [186] [219] [53] [90] [118] and/or HAM-D scores [53] [89] [118] [116]; patients may be medicated, unmedicated, or in remission. The Mini International Neuropsychiatric Interview (MINI) was employed in the data collection for the dataset reported in [25] in order to obtain diagnosis, and the Quick Inventory of Depressive Symptomatology-Self Report (QIDS-SR) for defining the symptom severity. BDI has also been used in [185] to establish whether a given patient was in remission. Comparison data were obtained from individuals who had never been diagnosed with depression or other mood disorder. Data collection from non-clinical samples, employed Self-RIs such as PHQ-9 [194] [84] [195] [176] [177] [93], and BDI [209] [210], assessing the severity of (sub-clinical) depression-related symptomatology. Recruitment methods further included flyers, posters, institutional mailing lists, social networks, and personal contacts.

Establishing conditions which enable the collection of signs related to depression is by far the most important step, as also discussed in [60]. Emotion elicitation is

used to measure reactions to emotionally charged stimuli, given that such reactions significantly differ between healthy and patient groups. The Handbook of Emotion Elicitation and Assessment [51] describes several methods for eliciting emotion in the laboratory including: emotional film clips used in [145] [113], images selected from the International Affective Picture System (IAPS) used in [145] [113], social psychological methods, and dyadic interaction. Emotionally charged images and clips are in principle capable of eliciting observable responses, although ethical considerations set limits to the shocking nature of the content. In this regard it is imperative that patients with depression are not subjected to unnecessary and unwanted stress or anxiety.

Structured Interviews are usually employed for gathering depressive symptoms, but have also been used for eliciting specific emotions by asking participants to describe personal, emotionally charged events [145]. Interviews can take place over one or more sessions, conducted by a therapist or a virtual character (VC), or guided by instructions on a computer screen. Typically the interview topic changes smoothly from casual to more intimate topics [53] [145] [194] [195] [147] [90] [89] [176] [239] [177] [178] [113] [157].

The amount of visual data that is necessary for a reliable assessment depends heavily upon the temporal nature of MDD. The specificity of the assessment method may benefit from multiple recording sessions, such as that of the data reported in [53] [90] [89]. Recording length depends on the elicitation method, with structured interviews being considerably longer in comparison with recordings based on emotion elicitation through films.

Studies vary widely depending on the types of equipment utilized and particular signs monitored. For instance, studies focusing on the pupillary response may only use a pupilometer [185] [186] [111] [219] and pay special attention to ambient illumination in order to optimize sensitivity. Again, depending on the approach, one or more cameras, typically color, are simultaneously used to cover more than one viewing angles and fields of view (e.g., both face and body separately [53]). Depth sensors (e.g., Microsoft Kinect) have also been utilized in some cases [194] [177].

2.2.2 Reported Datasets

The various datasets reported in relevant work are summarized in Table 2.2. Participant demographics, stimuli, ground truth, selection criteria, research question, as well as technical specifications, are some of the features that vary across studies. Most of

the studies employed adult participants, while two recruited adolescents. Methods for collecting depressive symptomatology included: a) interpersonal, i.e. interview with a clinical psychologist or interaction with family members, b) non-social, where participants were presented with stimuli on a computer, and c) combination of (a) and (b). The ground truth for the presence of depression varied accordingly, relying on clinical assessment in the majority of the cases, and on self-reports in two of the studies.

The selection criteria used depended greatly on the research question. DSM and HAM-D criteria were used for detection of depression [53] [90] [114] [118] [116] [139] or differentiation from Bipolar Disorder [234]. Others had more specific criteria, i.e. patients recovering from Deep Brain Stimulation of the Subcallosal Cingulate Cortex (DBS-SCC) [99], in order to monitor recovery progress. Studies assessing the predictive value of the method for future emergence of clinical depression in adolescents involved 9-12 year old participants at the initial data collection, with clinical reassessment after a two year interval [157] [156].

Technical specifications of the video recording equipment varied to some extent, but not significantly, as the setup typically involved a single camera monitoring participants' face/upper body. A notable exception was the setup employed for the Pittsburgh dataset, utilizing four hardware-synchronized analogue cameras; two for monitoring the participant's head and shoulders, one for full body monitoring, and the fourth monitoring the interviewer, together with two microphones for speech recording.

Regarding dataset availability, AVEC is the only fully available dataset for free download[1], while the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset is also partly available[2]. Both datasets require a signed End User License Agreement (EULA) in order to provide download access. The remaining reported datasets are proprietary, while in some cases they have been made available to visiting researchers. The number of participants listed in Table 2.2 is that reported in the latest publication employing the related dataset. However, different published papers report results obtained from different subsets; accordingly, sample size used in each published report is specified in Section 2.4 (Tables 2.5 and 2.6).

[1] http://avec2013-db.sspnet.eu/
[2] http://dcapswoz.ict.usc.edu/

Table 2.2: Datasets employed by the reviewed studies for depression assessment

Corpus	Population / Total (Control / Study)	Symptomatology Collection Methods	Ground Truth	Selection Criteria	Research Question	Image Resolution / Frame Rate	Availability to Third Parties
Pittsburgh	Adults / 49 (-/-)	Inter	Clinical Assessment	DSM-IV, HAM-D>15	Detection	640x480 / 29.97	Visual & Audio Features (Expected)
BlackDog	Adults / 130 (70/60)	Comb	Clinical Assessment	Control: No history of mental illness, Study: DSM-IV, HAM-D>15	Detection	800x600 / 24.94	-
DAIC-WOZ	Adults / 189 (-/-)	Comb	Self-report	Age, language, eye-sight	Detection, Severity	-	Visual & Audio Features, Audio Recordings, Transcripts
AVEC	Adults 58 (-/-)	Non-social	Self-report	-	Severity	-	Full Video Recordings, Visual & Audio Features
ORI	Adolescents / 8 (4/4)	Inter	-	-	Detection	-	-
ORYGEN	Adolescents / 30 (15/15)	Inter	Clinical Assessment	Stage1: No depression, age 9-12 years Stage2: Depression, 2 years after	Prediction	-	-
CHI-MEI	Adults / 26 (13/13)	Comb	Clinical Assessment	DSSS, HAM-D	Unipolar Depression / Bipolar Disorder	640x480 / 30	-
EMORY	Adults / 7 (-/7)	Inter	Clinical Assessment	DBS-SCC Treatment	Recovery	-/30	-

Inter: Interpersonal, Comb: Combination, Non: Non-social

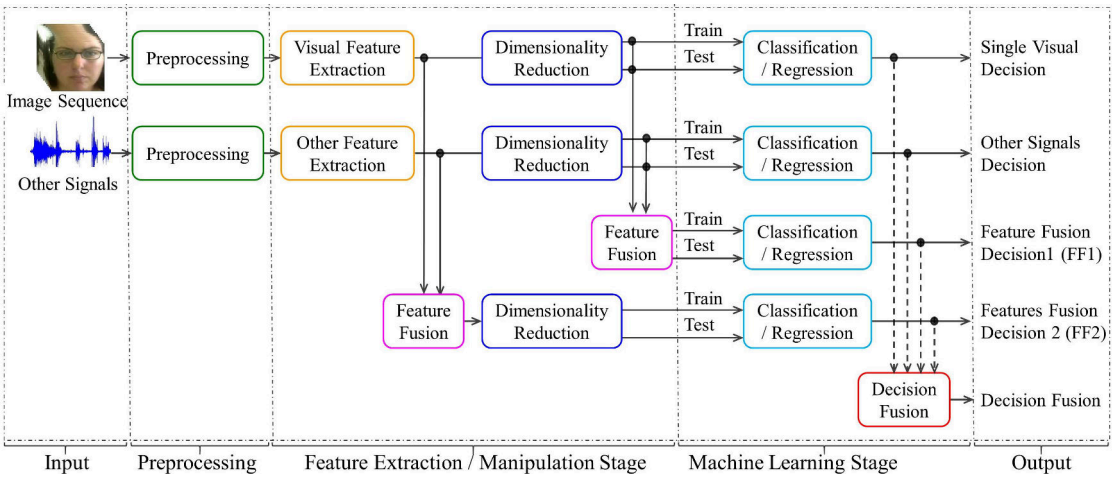


Figure 2.2: Typical workflow for automatic depression assessment. The output can be derived from a) single feature sets/modalities, b) feature fusion, with dimensionality reduction before (FF1) or after fusion (FF2), and c) decision fusion with any possible combination of outputs from single feature sets/modalities and feature fusion

2.3 State-of-the-Art Methods

The generic processing flow of an automatic system for depression assessment, combining the standard structure for automated audiovisual behavior analysis proposed by Girard and Cohn [87], with fusion methods presented in Alghowinem's thesis [21], is illustrated in Fig.2.2. Given a visual input (image sequence), along with other types of signals, such as audio, text from transcripts, and physiological signals, the prerequisite step is that of preprocessing. Feature extraction algorithms are subsequently applied to all visual signals, as described in Subsection 2.3.2.1 and illustrated in Fig. 2.4, while methods for dimensionality reduction are reported in Subsection 2.3.2.2. Machine learning algorithms are employed, depending on the research question, i.e., presence of depression or severity assessment. Classification approaches are suitable for categorical assessment, such as discriminating between a given number of classes (i.e., depressed vs. not depressed or low vs. high depression severity). For continuous assessment of depression (e.g., depression severity according to BDI scores ranging between 0-63) regression analyses are more appropriate. Methods used in the reviewed studies (i.e., algorithms for feature extraction from visual signs, feature selection algorithms, decision methods) are described in turn in the following section. In terms of better readability, references of the algorithms are inserted in relevant tables and figures, and not within the text.

2.3.1 Preprocessing

Given a visual input (video), illumination normalization, registration and alignment between the image sequences, and face detection are typical required preprocessing steps. Other types of signals, such as speech or physiological recordings, may also need preprocessing, such as segmentation. The most popular algorithm for face detection has been proposed by Viola and Jones [215]. Some off-the-shelf facial expression analysis applications have also been used widely as preprocessing tools, enabling researchers to focus on deriving high level information. An example of such a tool is the OpenFace freeware application[4] [31]. The Computer Expression Recognition Toolbox (CERT) [134] has been quite popular, but has now become commercialized. Z-Face [110] has also been employed for alignment and landmark detection.

2.3.2 Feature Extraction / Manipulation

This Subsection describes processes involved in feature extraction, dimensionality reduction, and fusion. The output of this processing stage generates the input to the machine learning stage, where no further manipulation of features is taking place.

2.3.2.1 Feature Extraction

Feature extraction is an important step in the processing workflow, since subsequent steps entirely depend on it. The approaches reviewed employ a wide range of feature extraction algorithms which, according to the well-established taxonomy in [55], can be classified as a) geometry-based, or b) appearance-based. In the field of depression assessment, several features are derived from the time-series of both (a) and (b) in the form of dynamic features. Close inspection of depression manifestations, listed in Table 1.1, reveals that the majority of signs involve muscle activity, which accounts for the temporal nature of the features. Features can be further categorized as high or low level; high level features directly translate to human common sense, while low level features are based on ”traditional” image processing descriptors. Depending on the approach, the software packages mentioned in Preprocessing could also serve as feature extraction methods (e.g., OpenFace, CERT, etc). In the present work feature extraction algorithms are grouped into those focusing on the face region and those relying on the

[4]https://www.cl.cam.ac.uk/ tb346/res/openface.html

body region. A pictorial taxonomy of the various algorithms, including the region of interest on which they are applied, the features computed, and references to respective studies, is presented in Fig. 2.4. The various features, retrieved from relevant studies, are described below in detail.

Face

Features related to the face are classified here into features from full face, AUs, facial landmarks, and mouth/eyes.

Full Face As it becomes apparent from Fig. 2.4, approaches employing feature extraction from the entire face region comprise the most popular category. Certain high level features extracted from the face as a whole concern basic emotional expressions displayed, given that depression is associated with reduced variability of emotional expression and greater persistence of a neutral expression. As expected, geometrical features, such as edges, corners, coordinates, and orientation, are often used to represent facial expressions. Functionals derived from the time series of geometric features are quite popular. Some examples are average, minimum, and maximum values of displacements, velocities, or accelerations of the coordinates that define the face region as a whole.

Appearance-based algorithms are also very popular for full-face based features. Among the most prevalent texture descriptors are Local Binary Patterns (LBP). Several variants of LBP have been created for automatic depression assessment, such as an extension of LBP that considers patterns on Three Orthogonal Planes (LBP-TOP). Local Gabor Binary Patterns in Three Orthogonal Planes (LGBP-TOP) extends LBP-TOP by computing patterns on the output of Gabor-filtered data, rather than on the original intensity image. Along the same lines, Local Curvelet Binary Patterns-Three Orthogonal Planes (LCBP-TOP) was introduced in some studies, which entails computing the patterns on the curvelet transform of the original image. Local Curvelet Binary Patterns- Pairwise Orthogonal Planes (LCBP-POP) is yet another variation of the algorithm operating on pairs of orthogonal planes. Additionally, the Block-Wise LBP-TOP (BW-LBP-TOP) method which computes the LBP-TOP for a specific number of non-overlapping blocks, has also been employed. Local Phase Quantization (LPQ) is another texture descriptor with a similar extension of Local Phase Quantization-Three Orthogonal Planes (LPQ-TOP). Another popular algorithm for motion-based approaches is

the histogram of optical flow, which estimates the motion within visual representations. Divergence-Curl-Shear (DCS) descriptors are also based on optical flow.

The Motion History Histograms (MHH) algorithm, which extends Motion History Images (MHI), has also been found in the related literature. A further extension of MHH is the 1-D MHH, which is computed on the feature vector sequence, instead of the intensity image. In a similar manner, Feature Dynamics History Histogram (FDHH), captures the dynamic variation which occurs among the extracted descriptors of image sequences. Gabor Motion History Image (GMHI) is another variant, which considers the Gabor filtered image, instead of the original frames. The Difference Image is a simplified process, which considers intensity differences between the first and the last frame. Motion Boundary Histogram (MBH) is another motion based algorithm, which considers the gradient in order to suppress the constant motion. Finally, the Space-Time Interest Points (STIP) algorithm, which detects local structures characterized by significant intensity variability in both space and time.

Deep learning, a subfield in machine learning, has become quite popular during recent years, presenting winning approaches in many contests [179]. Deep learning methods can be employed all the way from the beginning of the pipeline, incorporating feature extraction and machine learning level, or they can be utilized in a transfer-learning manner as pre-trained networks just for feature extraction, combined with other machine learning algorithms. DCNN, DNN, SDA, DTL, ResNet, VGG, and Alex-Net are some of the networks that can be found in the relevant literature.

Facial Landmarks Facial landmarks have been very popular in addressing problems related to facial expression analysis, and have been applied to depression assessment. Such algorithms localize fiducial points of the face and facial features, which are very useful in extracting high level traits directly associated with signs of depression, e.g. smiling. The Constraint Local Models method (CLM) introduced by Saragih et al. [175] is displayed in Fig. 2.3 to illustrate its application to the modeling of facial geometry. Active Shape Models (ASM) as well as Active Appearance Models (AAM) have also been utilized for depression assessment methods. Facial landmark data have also been analyzed as time series. Displacement, velocity, acceleration, as well as the landmark coordinates alone, have been used as features. In addition, polynomial fitting, as well as statistics derived from shape eigenvector velocities have been utilized. Landmark Motion History Images (LMHI) is a low level feature which, instead of the actual intensities,

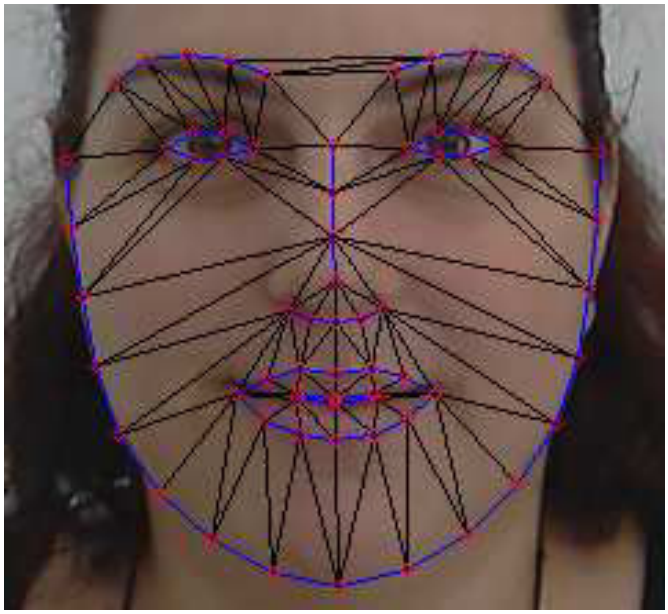


Figure 2.3: CLM landmarks fitted on a facial image (created using algorithm from [175])

computes the MHI on the motion of the facial landmarks. Finally, the Landmark Motion Magnitude (LMM) algorithm has also been applied to the vectors which displace each landmark from one frame to the next. Histogram of Displacement Range (HDR) in horizontal and vertical manner has also been proposed.

Action Units Action Units (AU) encode the coordinated activity of groups of facial muscles in correspondence to specific actions, including specific emotional expressions. They can be employed for measuring the Variability of Facial Expression (VFE), as depressed individuals tend to be less expressive. Other approaches apply AU as high-level features. AU occurrence by itself is meaningful as there are specific facial actions that are directly linked to the presence of depression (reduced smiling, mouth corners angled down, etc.). Additionally, although several approaches implement AU dynamically (e.g. duration, base rate, ratio of onset/offset), AU are essentially static signs.

Mouth & Eyes Apart from the face as a whole, features extracted separately from the mouth and eyes have also been found in the reviewed literature. Smile intensity and duration is a mouth-based feature which has been employed for automatic depression assessment, consistent with the clinical literature, as depressed individuals tend to smile less often. Mouth deformations, velocity and acceleration of horizontal and vertical movements have also been proposed. For the eye region, average vertical gaze, blinking

rate, and pupil dilation have been reported. Additionally, functionals from velocity and acceleration of horizontal and vertical eyelid movement have been used. Additional features include saccade latency, peak velocity of initial saccade, saccade duration, mean and standard deviation (SD) of intersaccadic intervals.

Body

Although body signs in general have been shown to convey manifestations of depression, few approaches have exploited their utility. Existing applications can be classified as relying on either upper body or relative body part movements. Features for upper body movements have been extracted through the STIP and DCS algorithms. Relative body part movements, on the other hand, have been exploited via the parts algorithm that represents orientation and distance from the torso center expressed in polar coordinates.

2.3.2.2 Dimensionality Reduction

Many feature extraction algorithms produce vectors of high dimensionality. The goal of dimensionality reduction algorithms is to reduce the number of features in a meaningful manner, in order to avoid corrupting the classifier. Dimensionality reduction algorithms can be classified in two groups: (a) Feature Transformation, and (b) Feature Selection [132]. In the first group features are transformed/combined by being projected from a high-dimensional space to low-dimensional space, to increase separability. On the other hand, in the second group, as the name implies, a selection procedure takes place, and the most discriminative/significant features are selected. Below, examples from both groups, as retrieved in reported approaches, are being described.

Feature Transformation

Principal Components Analysis (PCA) is the most popular algorithm in this category [53] [225] [108] [182] [189] [107] [126] [100] [221] [162] [163] [141] [119] and has been used to generate new features based on a linear transformation of the original features. Another set of approaches for reducing dimensionality involves codebooks. Bag-of-Words (BoW) [114] [118] [115] [59] [117] [121] [36] [37] [100] [63], initially intended for document classification, has been applied to image processing problems. A histogram-based approach was presented in [108] which entails maintaining the highest-scoring bins based on a predefined threshold adjusted to the total samples size.

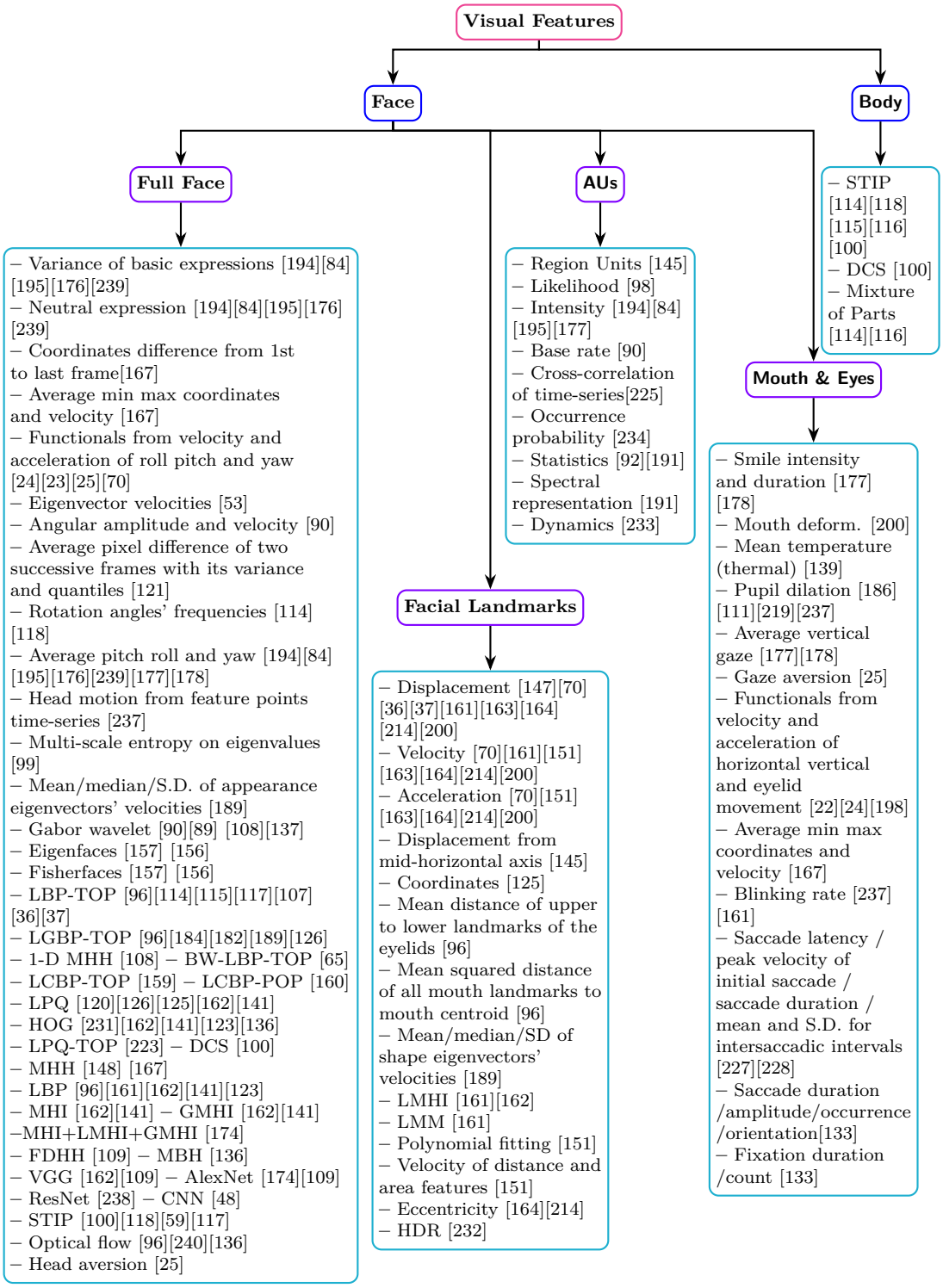


Figure 2.4: Taxonomy of visual features utilized in the reviewed studies for depression assessment

Feature Selection

In [24], [25], and [84] distributional statistics (e.g. t-tests) were employed to select only those features that met a predefined statistical threshold. In [70] [71] [119] the authors implemented the minimum Redundancy Maximum Relevance (mRMR) feature selection, which considers statistical dependency between features. Several feature selection approaches were evaluated in [96]: supervised feature selection, brute-force selection, and a backward selection scheme using bivariate correlations. Greedy forward feature selection [194] [195], relief from WEKA[1] [167], and Mutual Information Maximization (MIM) [151], are additional algorithms used for feature selection in the reviewed studies. Correlation-based Feature Selection (CFS) has also been employed in some studies [133] [92].

2.3.3 Machine Learning

The next step following feature extraction and manipulation, in all methods reviewed, is the machine learning stage. Depending on the particular research goals, different types of decision methods may be applied. Classification methods are appropriate to address categorical questions (e.g., "*depressed*" versus "*non-depressed*" and low versus high depression severity). When the research question concerns the concurrent prediction of depression severity through video-derived indices in a continuous manner, regression approaches are predominantly employed. Cross validation methods are typically applied before classification / regression step.

2.3.3.1 Cross Validation Methods

Cross validation methods are employed to establish algorithm reliability, namely its capacity to generalize well with newly introduced data. To establish reliability, a given data set is divided in two parts, one used to train the proposed algorithm, and another (left-out) to test its performance. Specific procedures used for dataset splitting include the leave-one-out [53] [84] [90] [89] [178] [22] [23] [117] [70] [161] and the k-fold method [239] [25] [160] [71]. In the leave-one-out procedure, for a dataset of N samples, N training sets are created of size N-1, each time consisting of all but one sample. The algorithm is then tested N times on its capacity to classify the "*left-out*" cases for each

[1] http://www.cs.waikato.ac.nz/ml/weka/

Table 2.3: Classification algorithms employed in relevant studies

SVM [237] [53] [147] [89] [176] [239] [22] [24] [23] [114] [118] [115] [116] [25] [117] [184] [107] [162] [141] [174] [133] [119]
k-Nearest Neighbour (kNN) [159] [157] [156] [160] [163] [133]
Gaussian Mixture Models (GMM) [22] [23] [137]
Naïve Bayes (NB) [194] [195] [133]
Random Forest (RF) [231] [161] [133]
Logistic Regression (LogR) [70] [133] [136]
Neural Networks (NNet) [118] [115]
Maximum Entropy Model (MaxEnt) [84] [239]
Relevance Vector Machines (RVM) [182] [63]
Hidden Conditional Random Fields (HCRF) [239]
Hidden Markov Model (HMM) [234]
Coupled Hidden Markov Model (CHMM) [234]
Stacked Denoising Autoencoders (SDA) [71]

set. Samples could be several for one subject, and therefore the leave-one-out could also be implemented in a leave-one-subject-out manner, where all samples from a specific subject are excluded each time. In the k-fold procedure the dataset is randomly split into k partitions, with one partition kept each time for testing and the remaining used for training the algorithm. This procedure is repeated for k times. In the context of the AVEC challenges, partitioning of the dataset into training and test sections was performed by the organizers, to permit direct comparisons between the algorithms used by participating groups.

2.3.3.2 Classification

An exhaustive list of classifiers employed in the reviewed studies can be found in Table 2.3 and were ranked in Fig.2.5. Support Vector Machines (SVM) is by far the most popular method for categorical assessment of depression. This can be justified by the fact that SVMs are well suited for binary problems of high dimensionality [32], such as the distinction of low symptom severity/absent depression from high symptom severity/present depression.

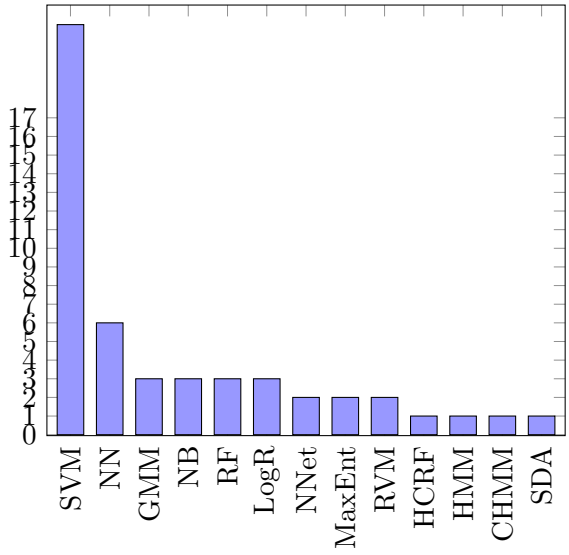


Figure 2.5: Ranking of classification algorithms employed in relevant studies

2.3.3.3 Regression

The continuous nature of depressive symptomatology is well supported by the clinical literature, as discussed in Subsection 1.2.2. As a result, relevant approaches have recently been gaining momentum, including the AVEC challenges aimed at predicting scores on self-report depression scales as a continuous variable using speech and video cues. As it can be observed in Table 2.4, the most popular regression algorithm, similarly to the classification-based approaches, is Support Vector Regression. Again the ranking of algorithms is illustrated in Fig.2.6.

2.4 Selected Approaches and Meta-Analysis

In this section different approaches, either classification- or regression-based, are compared in a quantitative manner. To be included in the analysis, studies must have reported results on automatic assessment of depression using visual features. Direct comparison is not the case for the various approaches included in this analysis and summarized in Tables 2.5, 2.6, and 2.7, since they were typically evaluated on different datasets (or subsets from the same corpus of data). Even in the case of AVEC participations, direct comparisons across the three challenges are not possible given that different data sets were used in each. The specific objective of our quantitative meta-analysis is

Table 2.4: Regression algorithms employed in relevant studies

Support Vector Regression (SVR) [96] [59] [120] [121] [182] [231] [223] [226] [92] [200] [63] [69]
Linear Regression (LR) [225] [108] [231] [109]
Partial Least Square Regression (PLS) [148] [108] [200] [109]
Gaussian Staircase Regression (GSR) [226] [105] [63]
Canonical Correlation Analysis (CCA) [126] [125]
Relevant Vector Machines (RVM) [182] [105]
Random Forest Regressor (RFR) [212] [172]
Deep Convolutional Neural Networks (DCNN) [231] [191]
Deep Convolutional Neural Netowork - Deep Neural Network (DCNN-DNN) [233] [232]
Deep Transformation Learning (DTL) [123] [192]
Cascade Random Forest (CRF) [198]
Ordinary Least Squares (OLS) [136]
Discriminative Mapping (DM) [223]
Moore-Penrose Generalized Inverse (MPGI) [126]
Extreme Learning Machines (ELM) [225]
Long Short Term Memory Recurrent Neural Networks (LSTM-RNN) [48]
Stochastic Gradient Descent (SGD) [92]

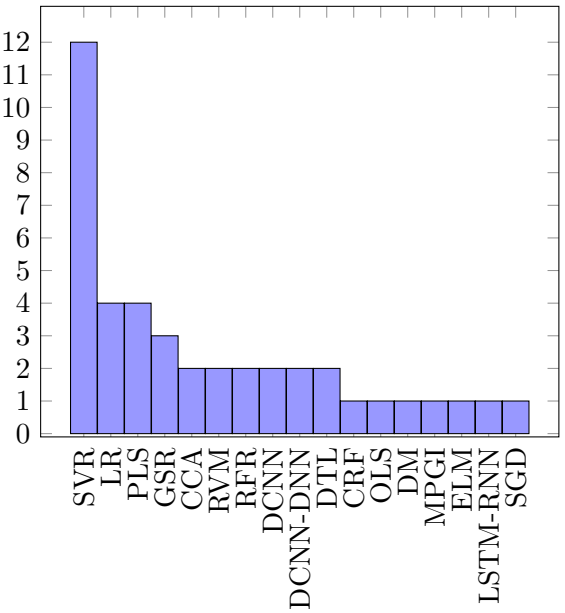


Figure 2.6: Ranking of regression algorithms employed in relevant studies

to identify general trends, key and strong points, to be considered in future studies of automatic depression assessment, given that a direct comparison of results is not viable.

2.4.1 Categorical Depression Assessment

Approaches for categorical depression assessment presented in this subsection are grouped and compared in terms of the employed dataset, in accordance with Table 2.2. Further, the results are considered with regard to the evaluated features, in reference to the taxonomy presented in Fig. 2.4. The various approaches, apart from reporting different performance metrics, were tested on datasets or particular subsets of varying sizes. Performance metrics in each report are explained next.

Accuracy, which was reported in the majority of studies, is computed according to Equation (2.2) based on the following confusion matrix:

$$C = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \tag{2.1}$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives. Certain studies report "*depressed accuracy*", which implies sensitivity (or recall) given by (2.4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

Where *precision* is given by:

$$precision = \frac{TP}{TP + FP} \tag{2.3}$$

and *recall* by:

$$recall = \frac{TP}{TP + FN} \tag{2.4}$$

The F1 score, also reported in several studies, is given by:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.5}$$

The aforementioned performance metrics, however, fail to take chance agreement into consideration, which varies across different studies. To address this limitation and to permit direct comparisons between classification approaches, Cohen's Kappa statistic

[52], a chance and skew robust metric, was computed whenever possible. It is based on the confusion matrix (2.1) and given by:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{2.6}$$

where p_0 is the proportion of accurately predicted decisions given by the accuracy formula as defined in (2.2), and p_e the proportion of expected chance agreement, given by:

$$p_e = \frac{M_a + M_b}{TP + FN + FP + TN} \tag{2.7}$$

where M_a and M_b are defined as follows:

$$M_a = \frac{(TP + FN) * (TP + FP)}{TP + FN + FP + TN} \tag{2.8}$$

$$M_b = \frac{(TN + FP) * (TN + FN)}{TP + FN + FP + TN} \tag{2.9}$$

Whenever confusion matrices for user-/gender-independent depression assessment (based on one or more visual cues) were not included in the original publication, they were requested from the study authors. For the cases that the requested information was not provided, and if at least two performance metrics where reported in the original publication, along with the total number of subjects per class (depressed/non-depressed, as defined in (2.10) and (2.11)), a 4×4 linear system of equations was solved in order to derive the confusion matrix.

$$\#depressed = TP + FN \tag{2.10}$$

$$\#not - depressed = TN + FP \tag{2.11}$$

In the case of [176] a quadratic system was solved, since the reported metrics were averaged for the two classes (depressed/non-depressed). Finally, the computed confusion matrices were cross-checked to reproduce the originally reported performance metrics. If the estimated confusion matrices for a given study could not be verified by the reported performance metrics, the relevant study was not considered any further. It should be noted that Dibeklioğlu et al. [71], Pampouchidou et al. [162], and Kacem et al. [119]

are the only reviewed publications which originally included Kappa statistics in their published report. Table 2.5 groups the reviewed studies according to the dataset used. Studies are ranked by decreasing κ value within each dataset-specific group. Table 2.6 presents similar information on studies using datasets or dataset combinations reported in single studies, precluding direct across-study comparisons.

2.4.1.1 Pittsburgh

The first report involving the Pittsburgh dataset is that of Cohn et al. [53]. Girard et al. [90] extended this work by investigating the correlation of changes in patient clinical status with corresponding changes in facial expression and head motion patterns [89]. The cross-cultural study of Alghowinem et al. [24] was also tested on the Pittsburgh dataset among others, reporting an average recall of 94.7%. Dibeklioğlu et al. [70] tested several feature settings on the Pittsburgh dataset. More recently Dibeklioğlu et al. [71] presented a deep learning approach for detecting three levels of depression. Finally, Joshi et al. [116], and Joshi [114], reported 97.2% accuracy for assessing depression severity based on the Pittsburgh data.

2.4.1.2 BlackDog

McIntyre et al. [145] were the first to report on the BlackDog dataset. They reported identifying two clusters of patients, those who showed psychomotor agitation and those who showed motor slowing [147] [146]. In a subsequent study, Joshi et al. [115] achieved depression detection accuracy as high as 88.3%. Higher performance, up to 91.7%, was achieved when additional modalities were included (speech, independent and relative movement of body parts) in Joshi et al. [118] and Joshi et al. [117]. Alghowinem et al. studied depression detection based on the analysis of either eye movements in [22], or head pose and head movements in [23]. The maximum reported recall rate was 80% for the eye-based approach, and 82.6% for the head-based approach among women. In the same cross-cultural study mentioned in subsection 2.4.1.1, Alghowinem et al. [24] combined the two approaches (eye-based and head-based) achieving an average recall of 85%. Recently, Alghowinem et al. [25] reported improved recall performance of 88.3%, by extending their approach in terms of feature extraction, as well as machine learning methods.

2.4.1.3 DAIC-WOZ

Several approaches have been tested on this dataset, mainly from the primary research group, but also as part of the AVEC'16 depression sub-challenge. Scherer et al. [177] examined the value of the audiovisual approach, achieving 89.74% accuracy. Stratou et al. [194] corroborated Alghowinem's findings [22] [23] on gender differences in classification accuracy, reporting F1 scores of 0.858 for women and 0.808 for men in detecting presence of depression and PTSD. Finally, DAIC-WOZ served as the benchmark dataset in AVEC'17 and AVEC'16. Raw data were provided for audio recordings and transcripts, while for videos only features extracted with OpenFace were provided. Despite this limitation several interesting approaches were presented.

2.4.1.4 AVEC

The AVEC dataset, although intended for continuous assessment of depression, was also used for categorical assessment using case groupings based on BDI scores. For instance, Senoussaoui et al. [182] achieved classification accuracy of 82% for categorical assessment of depression by using a cutoff score of 13/14 points on BDI. In their cross-cultural study, Alghowinem et al. [24] tested their algorithm on a subset of the AVEC data and reported an average recall of 68.8% for the fixed set of features across the three datasets. Pampouchidou et al. [162] achieved 89% accuracy, which is the highest reported performance for categorical assessment on the AVEC dataset.

2.4.1.5 Other Datasets

Datasets or dataset combinations used in a single study are summarized in Table 2.6. Alghowinem et al. [24] attempted to merge several datasets (e.g., Pittsburgh and AVEC). This resulted in an improvement of classification performance as compared to relying solely on the AVEC dataset, reporting an average recall of 85.7%. In one of the earliest studies, the corpus constructed by the Oregon Research Institute (ORI), was used to test the approach of Maddage et al. [137] for video-based depression detection in adolescents. The ORYGEN dataset was employed for a more challenging endeavor undertaken by Ooi et al. [157], in order to predict whether initially non-depressed adolescents would develop depression at the end of a two-year follow-up period. Zhou et al. [237] at the University of Rochester departed from the traditional laboratory settings and obtained

data in realistic conditions. They reported 0.817 precision and 0.739 recall for classifying patients versus healthy volunteers reporting high levels of negative mood. Finally, recent results from the CHI-MEI dataset [234], attempting to distinguish unipolar depression (MDD) from bipolar disorder, reached 65.38% classification accuracy when combining AUs and audio features.

Table 2.5: Comparison of approaches for categorical assessment of depression grouped according to the dataset used, ranked within group based on Kappa

Paper	Population (Study/-Control) / Male rate	Features	Classification Algorithm	Reported Accuracy (or as otherwise noted)	Kappa
		Pittsburgh			
Joshi (2013) [116] [114]	36 (18/18) / 36.1%	Body, Full Face	SVM	97.2%	0.94
Alghowinem et al. (2015) [24]	38 (19/19) / 36.8%	Eyes, Full Face	SVM	mean recall =94.7%	0.89
Dibeklioğlu et al., (2015) [70]	95 (58/37)*1 /40.4%	Facial Landmarks, Full Face, Audio	LR	91.38%	0.78
Dibeklioğlu et al., (2017) [71]	130 ([58/35]/37)*1*2 /40.4%	Facial Landmarks, Full Face, Audio	SDA	78.67%	0.73
Kacem et al., (2018) [119]	126 ([56/35]/35)*1*2	Full Face	SVM	70.8%	0.65
Dibeklioğlu et al., (2015) [70]	130 ([58/35]/37)*1*2 /40.4%	Facial Landmarks	Logistic Regression	84.49%	0.63
Dibeklioğlu et al., (2017) [71]	130 ([58/35]/37)*1*2 /40.4%	Facial Landmarks	SDA	72.59%	0.62
Dibeklioğlu et al., (2015) [70]	-//-	Full Face	LR	86.21%	0.60

*1 Reported sample size corresponds to number of sessions

*2 Three classes were considered, corresponding to the annotation in the table as follows: ([Moderate to Severe/Mild]/Remission)

*3 The confusion matrix was computed based on performance metrics provided by the authors of the original report

Continue on the next page

Table 2.5: Comparison of approaches for categorical assessment of depression **(cont.)**

Paper	Population (Study/-Control) / Male rate	Features	Classification Algorithm	Reported Accuracy (or as otherwise noted)	Kappa
Cohn et al. (2009) [53]	107 (66/41)*1 /35%	Facial Landmarks	SVM	79%	0.53
		BlackDog			
Joshi (2013)[114] [117]	60 (30/30) / 50%	Upper Body, Full Face, Audio	SVM	91.7%	0.83
Alghowinem et al. (2016) [25]	-//-	Full Face, Audio	SVM	mean recall =88.3%	0.77
Alghowinem et al. (2016) [25]	-//-	Eyes, Full Face	SVM	mean recall =78.3%	0.57
Alghowinem et al. (2013) [23]	-//-	Full Face	SVM	mean recall =76.8%	0.53
Joshi 2013, Joshi et al. (2013) [114] [118]	-//-	Upper Body	SVM	76.7%	0.53
Alghowinem et al. (2015) [24]	-//-	Eyes, Full Face	SVM	mean recall =76.7%	0.53
Alghowinem et al. (2013) [22]	-//-	Eyes	SVM	mean recall =75%	0.50
		DAIC-WOZ			
Yang et al. (2016) [231]	35 (7/28) / 45.7%	Facial Landmarks, AU, Full Face, Audio, Text	Random Forest	F1 = 0.86	0.82
Scherer et al. (2013) [176]	39 (14/25) / 62%	Full Face, Audio	SVM	89.74%	0.76*3
Yu et al. (2013) [239]	130 (30/100) / 53%	Full Face, Audio	HCRF	F1=0.644	0.58*3

*1 Reported sample size corresponds to number of sessions

*2 Three classes were considered, corresponding to the annotation in the table as follows: ([Moderate to Severe/Mild]/Remission)

*3 The confusion matrix was computed based on performance metrics provided by the authors of the original report

Continue on the next page

Table 2.5: Comparison of approaches for categorical assessment of depression **(cont.)**

Paper	Population (Study/-Control) / Male rate	Features	Classification Algorithm	Reported Accuracy (or as otherwise noted)	Kappa
Nasir et al. (2016) [151]	35 (7/28) / 45.7%	Facial Landmarks	SVM	F1=0.63	0.55*3
Pampouchidou et al. (2016) [161]	-//-	Facial Landmarks, Audio	RF	F1=0.62	0.53
Valstar et al. (2016) [212]	-//-	Facial Landmarks, AU, Audio	SVM	F1=0.5	0.45*3
Valstar et al. (2016) [212]	-//-	Facial Landmarks, AU	SVM	F1=0.5	0.45*3
Pampouchidou et al. (2016) [161]	-//-	Facial Landmarks	RF	F1=0.5	0.39
Scherer et al. (2013) [176]	39 (14/25) / 62%	Full Face	SVM	64.1%	0.20*3
		AVEC			
Pampouchidou et al. (2017) [162]	200 (34/166)	Full Face	SVM	89%	0.78
Senoussaoui et al. (2014) [182]	50 (25/25)	Full Face	SVM	82%	0.64
Alghowinem et al. (2015) [24]	32 (16/16)/ 28.1%	Eye, Full Face	SVM	mean recall =68.8%	0.38
Pampouchidou et al. (2016) [160]	200 (34/166) *1 / 33%	Full Face	kNN	74.5%	0.13

*1 Reported sample size corresponds to number of sessions

*2 Three classes were considered, corresponding to the annotation in the table as follows: ([Moderate to Severe/Mild]/Remission)

*3 The confusion matrix was computed based on performance metrics provided by the authors of the original report

2.4.2 Continuous Depression Assessment

The majority of the approaches reviewed here are based on AVEC datasets, either as participations to the actual challenge or as independently published studies. The depression challenge of AVEC 2013 and AVEC 2014 required prediction of individual BDI scores based on corresponding video recordings (both visual and speech data are available). Video recordings were divided into three subsets (training, development and testing), with labels being released only for the first two. AVEC 2013 included the complete recordings of the participants executing 12 tasks, while for AVEC 2014 only two tasks were kept: the Northwind (reading a novel passage) and Freeform (answering a series of questions-both neutral and potentially emotionally challenging). AVEC 2016, although focused on categorical assessment, encouraged participants to address prediction of self-reported scores on the PHQ-8 scale, employed by DAIC-WOZ. Table 2.7 summarizes relevant proposed approaches. In all cases, performance metrics were the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) given by equations (2.12) and (2.13), where n is the number of samples, p the predicted value, and a the actual value.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - a_i)^2} \tag{2.12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - a_i| \tag{2.13}$$

2.5 Scope of the Thesis

The scope of this thesis is based on the motivation to deliver an unobtrusive depression assessment methodology based on visual cues, as described in Section 1.3, as well as key issues that emerged from reviewed literature.

The most important issue concerning the data is availability. As mentioned before, obscuring participant identity is practically impossible in raw video data compared to other modalities (e.g. speech or physiological signals). Consequently, open access is strictly prohibited, while licensing to a third party is seriously restricted.

Regarding the categorical approaches tested on the AVEC and DAIC-WOZ datasets, there are additional reasons which could have affected performance, as revealed through comparison of κ values cross studies. Both datasets were collected in a way that limited

Table 2.6: Approaches for categorical assessment of depression employing datasets, or combination of datasets, which have not been reported elsewhere

Data	Paper	Population (Study/-Control) / Male rate	Features	Classification Algorithm	Reported Accuracy (or precision)	Kappa
Pittsburgh + AVEC'14	Alghowinem et al. (2015) [24]	70 (35/35) / 32.9%	Eyes, Full Face	SVM	mean recall =85.7%	0.57
Rochester	Zhou et al. (2015) [237]	10 (5/5) /-	Full Face, Eyes	LogR	precision= 0.82	0.57
ORI	Maddage et al. (2009) [137]	8 (4/4) / 50%	Full Face	GMM	75.6%	0.45 *
ORYGEN	Ooi et al. (2011) [157] [156]	30 (15/15)/ 51%	Full Face	kNN	51%	0.03 *
CHI-MEI	Yang et al. (2016) [234]	26 (13/13) / -	AU, Audio	CHMM	65.38%	0.26
	Yang et al. (2016) [234]	26 (13/13) / -	AU	HMM	53.85%	0.08

* Confusion matrix was computed, and not provided by the authors of the original research report

the audience effect (human-computer interaction setup), eliminating cues that could only occur in a social context.

Relying on self-reported symptoms for data annotation, as in AVEC and DAIC-WOZ, is far more complex in comparison to the other datasets. Self-RIs scores depend on a variety of biasing factors (subjective, social, etc.), and although depression may be accurately portrayed by facial expressions, this type of ground-truth may not accurately measure depression severity. It appears that an optimal dataset should be comprised of a number of patients diagnosed by experienced psychiatrists, using largely uniform diagnostic criteria. Consequently, the development of an algorithmic tool for depression assessment, would most definitely require direct supervision by clinicians.

With respect to video acquisition parameters, average image resolution and frame rate are typically reported. However, it is not clear whether image acquisition with higher-level specifications could improve assessment accuracy. A quantitative comparison of different resolutions and frame rates employing the same algorithm(s) may be required to conclusively address this question.

Given the temporal variability of depressive manifestations, the majority of facial signs utilized in the reviewed studies are dynamic, while the occurrence of static signs is typically considered over time. Therefore it is clear that video recordings are of interest, as opposed to static images.

Treating depression as a continuous variable (i.e., reflecting depression severity) is gaining ground over categorical decision systems. This may be due to the fact that, despite the apparent simplicity of categorical assessment, it does not represent neither properly nor reliably the complex nature of mood disorders.

Consequently, the scope of this thesis can be summarized as:

- Constructing a clinically valid dataset, comprising diagnosed patients as well as healthy control individuals
- Test different types of stimuli, in order to compare the effect of interpersonal vs non-social context
- Evaluate several annotation methods to test their accuracy; specifically:
    1. Diagnosis
    2. Self-report
    3. Expert annotation
- Experiment with different video acquisition parameters
- Develop video-based methodology to extract features correlating with signs of depression
- Investigate categorical vs continuous depression assessment

Table 2.7: Summary of methods and results of studies employing continuous depression assessment based on the dataset provided by AVEC'13, AVEC'14 and AVEC'16. Performance metrics are given in the form of MAE / RMSE and approaches have been ranked according to reported performance primarily on Test-RMSE, and secondly on Development-RMSE

Paper	Regression	Features	Development	Test
AVEC 2013 - BDI prediction - Complete recordings				
Zhou et al. (2018) [238]	CNN	Full Face, Eyes, Mouth	–	6.20 / 8.28

Continue on the next page

Table 2.7: Summary of studies employing continuous depression assessment **(cont.)**

Paper	Regression	Features	Development	Test
Ma et al. (2017) [136]	LDA+OLS	Full Face	–	7.26 / 8.91
Zhu et al. (2017) [240]	DCNN	–	–	7.58 / 9.82
Kaya et al. (2014) [126]	MPGI		–	8.254 / 10.315
Cummins et al. (2013) [59]	SVR		– / 12.08	– / 10.45
Meng et al. (2013) [148]	PLS		7.09 / 8.82	9.14 / 11.19
Valstar et al. (2013) [209]	SVR		8.74 / 10.72	10.88 / 13.61
AVEC 2014 - BDI prediction - Northwind / Freeform tasks				
Jan et al. (2017) [109]	PLS	Full Face	7.25 / 9.52	6.68 / 8.01
Zhou et al. (2018) [238]	CNN	Full Face, Eyes, Mouth	–	6.21 / 8.39
Kang et al. (2017) [123]	DTL	Full Face	–	7.74 / 9.43
Zhu et al. (2017) [240]	DCNN		–	7.47 / 9.55
Kaya & Salah (2014) [125]	CCA		–	7.86 / 9.72
Jain et al. (2014) [107]	SVR		6.969 / 8.167	8.399 / 10.249
Jan et al. (2014) [108]	PLS+acslr		7.36 / 9.49	8.44 / 10.50
Kächele et al. (2014) [120]	SVR		7.03 / 8.82	8.97 / 10.82
Valstar et al. (2014) [210]	SVR		7.577 / 9.314	8.857 / 10.859
Senoussaoui et al. (2014) [182]	GLM+acsrvm + SVR		6.95 / 8.52	-
Smailis et al. (2016) [189]	SVR		- / 9.07	–
He et al. (2015) [100]	SVR		7.99 / 9.63	-

Continue on the next page

Table 2.7: Summary of studies employing continuous depression assessment **(cont.)**

Paper	Regression	Features	Development	Test
Sidorov & Minker (2014) [184]	SVR		14.843 / 17.667	-
AVEC 2016 - PHQ-8 prediction - DAIC-WOZ				
Sun et al. (2017) [198]	CRF	Facial Landmarks, Eyes, AU	4.6 / 5.9	4.89 / 6.23
Stepanov et al. (2017) [192]	LSTM-RNN	Facial Landmarks	4.66 / 6.09	5.36 / 6.72
Valstar et al. (2016) [212] [172]	RFR		5.88 / 7.13	6.12 / 6.97
Song et al. (2018) [191]	CNN	Full Face, AU, Eyes	4.37 / 5.84	–
Williamson et al. (2016) [226]	GSR		5.33 / 6.45	–
Dham et al. (2017) [69]	SVR	Facial Landmarks	4.91 / 6.46	–
Dang et al. (2017) [63]	GSR	AU	5.34 / 6.67	–

Chapter 3

Methodology

The work-flow for the proposed methodology was based on the one presented in Fig. 2.2, which was synthesized based on the literature review. Different methods corresponding to the preprocessing, feature extraction, dimensionality reduction, and machine learning stages are described in the respective sections. The algorithms involved in the different stages are described in some level of detail, while not all of them were implemented together. Later on in the thesis, and more specifically in Ch.4 Preliminary Experimental Evaluation different experiments are described, while the main study is described in Ch.5. Thus the experiment that involves each algorithm described in this Chapter is mentioned accordingly.

3.1 Preprocessing

Conditions during data collection are not always conveniently established, and therefore preprocessing can help in obtaining meaningful information for further analysis. The relevant algorithms employed during different experimental tests of this thesis included: a) illumination normalization, b) face detection, c) facial landmarks detection, and d) face alignment. The specific algorithms for each preprocessing step are explained within the following subsections.

3.1.1 Illumination Normalization

Illumination normalization is employed in order to establish the appropriate contrast, brightness and other attributes of the images related to light. It is important in order

to make sure that the illumination level is adjusted to a level which can contribute to further analysis. Illumination normalization was employed for the methodology and experimental tests carried in terms of Pampouchidou et al. [159]. The algorithm utilized for the illumination normalization was the one provided by the INFace Matlab Toolbox[1] [197] [196]. During this procedure the original images (video frames) were first passed through homomorphic filtering to obtain an illumination normalized image, followed by contrast enhancement. Homomorphic filtering is an algorithm employed to improve the image quality by compressing the intensity range while enhancing contrast.

3.1.2 Face Detection

Although visual signs of depression may involve upper or full body, the face portrays the majority of the manifestations, and thus the proposed methodology focused on facial expression analysis; therefore one of the most crucial steps involved is the accurate face detection. Viola & Jones (VJ) [215] is perhaps the most popular face detection algorithm, based on Haar features, AdaBoost learning, and cascade classifiers. Haar features use specified rectangle patterns related to the face structure. Based on intensity differences, such as the ones displayed in Figure 3.1, the image is scanned for a given pattern and the features corresponding to each pattern are computed. In general, intensities corresponding to the white and black regions are summed separately, and the difference of the two sums is then calculated. More specifically, for the two-rectangle feature, VJ computes sums of the pixel values individually in black and white regions, and then finds the difference between them. For the three-rectangle feature, the two outer rectangles are summed together, and their sum is subtracted by the sum that results from the central region. In the four-rectangle case the algorithm sums the diagonal pairs, and computes the difference of the resulting sums.

To reduce the time needed for these summations, an intermediate step is used, that of the integral image. The image is scanned for several scales, starting at the size of 24×24, and every next run grows 1.25 times higher, until it reaches the total image size. Following, AdaBoost takes the obtained features for a training set containing positive and negative examples of faces. A collection of weak (i.e., easy to construct classification

[1] https://www.mathworks.com/matlabcentral/fileexchange/26523-the-inface-toolbox-v2-0-for-illumination-invariant-face-recognition

(a) Examples of Haar features (b) On a face 1 (c) On a face 2

Figure 3.1: VJ face detection: Haar features applied on a facial image

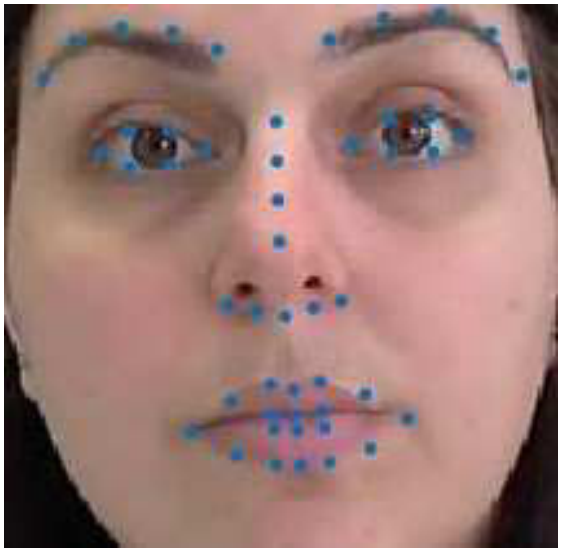
functions) are combined into a strong classifier. For a series of hypothesis, the classification system is trained, each time keeping the training weights associated with the lowest classification error. For example, in the case of a classifier based on perceptrons, the AdaBoost will keep that perceptron associated with the highest accuracy after each testing. Cascade classifiers are used to reduce computational time and enhance performance. They succeed by starting classifying with simpler classifiers, in order to reject subwindows that are definitely not faces, then moving forward they minimize the false positive rate by using more complex classifiers.

For the needs of the first experiment (see section 4.2), face parts were detected using the algorithm proposed by Tanaka [201], who extended Viola & Jones object detector by exploiting facial features hierarchy to reduce false positive rate. Left and right eyes were extracted separately, while the eye-pair region was eventually selected and created, by combining the two separate eye regions as implemented in Pampouchidou et al. [159].

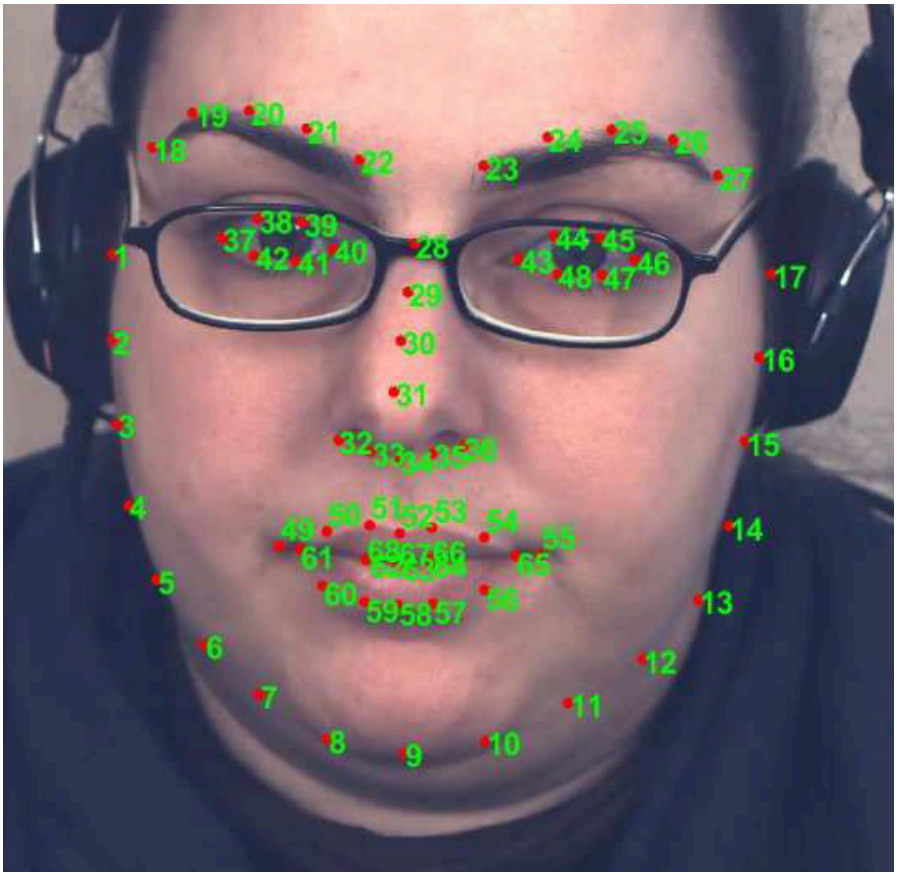
3.1.3 Facial Landmarks Detection

The accurate detection of the facial region was a rather complex albeit important first step in the present work. Another option, in order to obtain the face region was based on the use of facial landmarks. In this respect, two different algorithms were assessed at two different stages of the thesis. Initially, and for the works published in Pampouchidou et al. [159] and Pampouchidou et al. [160] (experiments described in subsections 4.2 and 4.3) the GNDPM [1] [205] [230] was applied on the image sequences (see Fig.3.2.a).

[1] https://ibug.doc.ic.ac.uk/resources/gauss-newton-deformable-part-models-face-alignment/



(a) Gauss-Newton facial landmarks



(b) OpenFace facial landmarks

Figure 3.2: Examples of landmarks on a facial image based a) on Tanaka face-parts detection combined with GNDPM and b) OpenFace facial landmarks

The GNDPM model requires accurate face detection, which was implemented based on the Tanaka face parts detection [201], and the detected landmarks were further used for aligning the face regions based on the eyes. GNDPM algorithm is pretrained and minimizes the reconstruction error between the original image and a Generative Parts Model iteratively, with the use of Gauss-Newton optimization.

The OpenFace[1] by Baltrušaitis et al. [31] (see Fig. 3.2.b) served in the remaining published approaches developed during this thesis [161] [163] [162] [164] (described in sections 4.4, 4.5, 4.6, and 5.2.1). OpenFace, which was applied directly on video recordings, is an open source tool that provides a variety of options apart from detecting landmarks; it also detects head pose, eye gaze, HOG features, AU, and basic emotions. In some experiments (e.g. experiment 4.6) features provided by OpenFace were tested for classification in addition to the ones proposed by the thesis work. OpenFace is based on Conditional Local Neural Fields, which is an instance of CLM, advanced in the use of patch experts and optimization function. The Point Distribution Model implemented within OpenFace captures landmark shape variations, while the patch experts capture local appearance variations of each landmark.

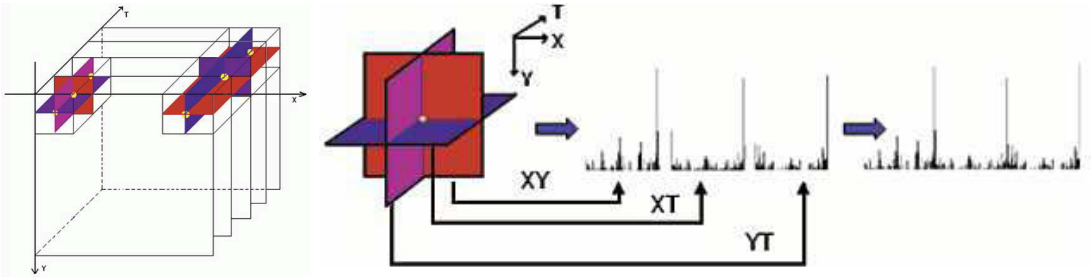
3.2 Motion Representation

The ultimate goal of the proposed thesis was to detect signs of depression based on visual input. Based on the clinical background and the signs reviewed in Table 1.1, it is evident that most of the signs are of temporal nature, thus video based analysis is of focus rather than mere static image processing. This constitutes the motion representation as the most important step of the pipeline, and therefore most of the work has been done within this context.

3.2.1 Local Curvelet Binary Patterns

LBP was initially proposed for extracting texture-based information from a static image, however it was later extended in order to extract spatiotemporal information. Work presented in Zhao and Pietikainen [236] proposed the Volume Local Binary Pattern (VLBP) in order to provide descriptors for dynamic texture analysis. VLBP examines a sequence of images (e.g. video frames) in the {X,Y,Z} space, where X and Y denote

[1] https://github.com/TadasBaltrusaitis/OpenFace/



(a) (b)

Figure 3.3: LBP from Three Orthogonal Planes (Taken from:[236])

the spatial coordinates, and Z the frame index (time). Volume textons are interpreted in histograms, in the same manner as in LBP, for all three planes, as illustrated in Fig. 3.3. Another notable extension of this approach, was from Almaev and Valstar [26], who considered Gabor filtered image frames rather than the original intensity-based images. Furthermore, inspired by Pampouchidou [158], where Curvelet transform was used for facial expression recognition, another extension was proposed in the context of this thesis, which considered the Curvelet transformed frames for applying the Three Orthogonal Planes process. Next the Curvelet transform is being explained in short.

3.2.1.1 Curvelet Transform

Curvelets belong to the family of geometrical wavelets, and were proposed by Candès et al. [43]. Their origin can be found in ridgelets [42], and thus it is preliminary to begin with the ridgelet definition. A characteristic ridgelet waveform can be seen in Figure 3.4, while ridgelet transform for an image function $f(x, y)$ is given by:

$$\Re_f(a, b, \theta) = \int \int \psi(x, y) f(x, y) dx dy$$

where $a > 0$, $b \in \mathbb{R}$, and θ are respectively the scale, the translation, and the orientation. ψ is the ridgelet function given by:

$$\psi_{a,b,\theta}(x, y) = a^{-\frac{1}{2}} \psi(\frac{x \cos\theta + y \sin\theta - b}{a}) \tag{3.1}$$

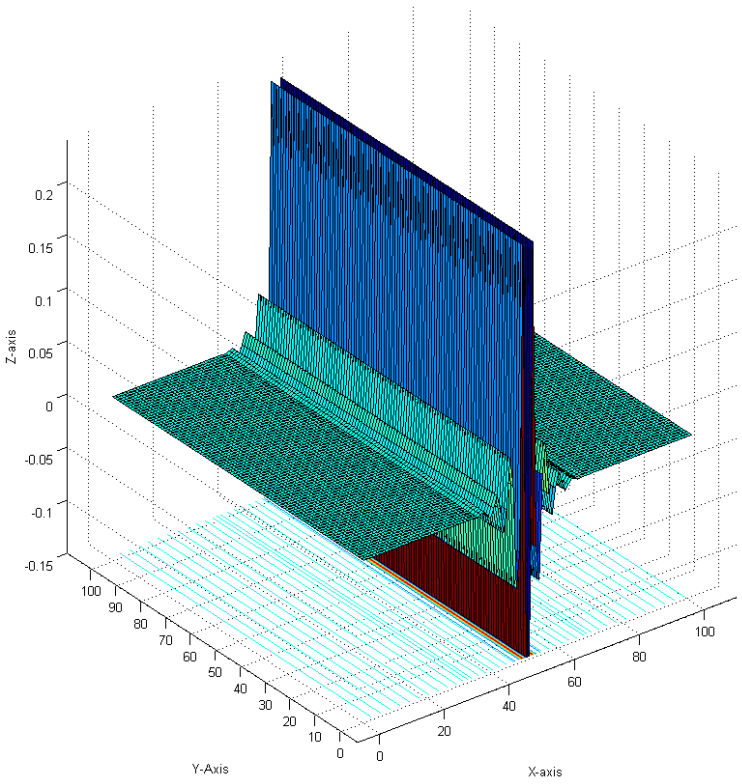


Figure 3.4: Ridgelets' example

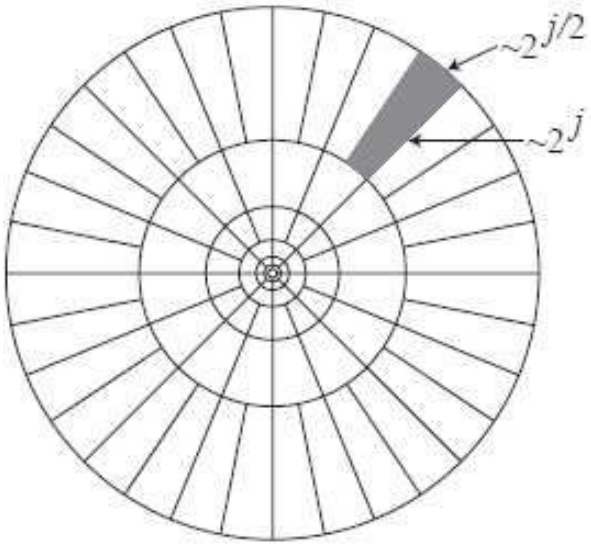


Figure 3.5: Curvelet tiling of space and frequency. Source:[45]

Just like Gabor, ridgelets have the property of being tuned in different scales and orientations [73]. Accordingly, the digital curvelets transform for an image $f(n, m)$ of dimensions M by N is given by [106]:

$$\mathfrak{C}_f(a, b, \theta) = \sum_{0 \leq m < M} \sum_{0 \leq n < N} f(m, n) \psi_{a,b,\theta}(m, n) \tag{3.2}$$

Curvelets, that have been characterized as an *"optimal representation of objects with piecewise C2 singularities"* [44], can be expressed in frequency domain by [106]:

$$\mathfrak{C}_{a,b,\theta} = \hat{\mathfrak{F}}(\mathfrak{F}(f(m, n) \times \mathfrak{F}(\psi_{a,b,\theta}(m, n))) \tag{3.3}$$

Zhang et al. in [235] introduced the term *"curvefaces"* in an effort to reduce the dimensionality of the facial image [140]. Motivation for further exploration of curvelets lies into the fact that the facial expression is characterized by the geometry of the facial features. Numerous examples demonstrate how the geometry of the face gives information about the expression. For instance something that people learn in a young age when drawing: on a happy face the mouth curve is facing up, while on a sad face the opposite happens.

The method introduced by Candès et al. in [45] is being followed, according to the available implementation in the Curvelab toolbox[1]. As already explained previously, curvelets originate in ridgelets, and therefore inherit and extend their properties. Extensions are such as that any arbitrary function can be expanded as series of curvelets, but also the parabolic scaling of the geometrical wavelet. Parabolic scaling means that the curvelet is elongated more in certain directions, and does not keep the same profile as ridgelets.

Curvelets are also discrete in terms of scale, location, and angle. Scale is doubled with each discretization level; that is at each scale, resolution is doubled as it moves further from the center in frequency domain, as shown in Figure 3.5. Curvelet algorithm in simple steps is given by first computing the 2D Fast Fourier Transform (FFT) of the image, then fixing the scales to be considered by the parabolic scaling in a resulting $\hat{f}$. Finally the curvelet coefficients are obtained by the inverse FFT applied on the product of $\hat{f}$ and ψ function as given in Equation 3.1. In Fig. 3.6 curvelet coefficients are wrapped

[1] http://www.curvelet.org/software.html

for display purposes; the coarse level for scale one and original orientation is placed in the center, then moving towards outside, coefficient for different scales and rotations are collocated. Computational complexity of curvelets for an $n \times n$ image is $O(n^2 \log n)$.

The curveface image computed in this step consists of complex values, therefore before further processing it should be converted to real. The justification applied in order to obtain real values, is to take the Euclidean distance between the real part and the imaginary, given by taking the magnitude:

$$\|x - y\| = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + ... + (y_n - x_n)^2} = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \tag{3.4}$$

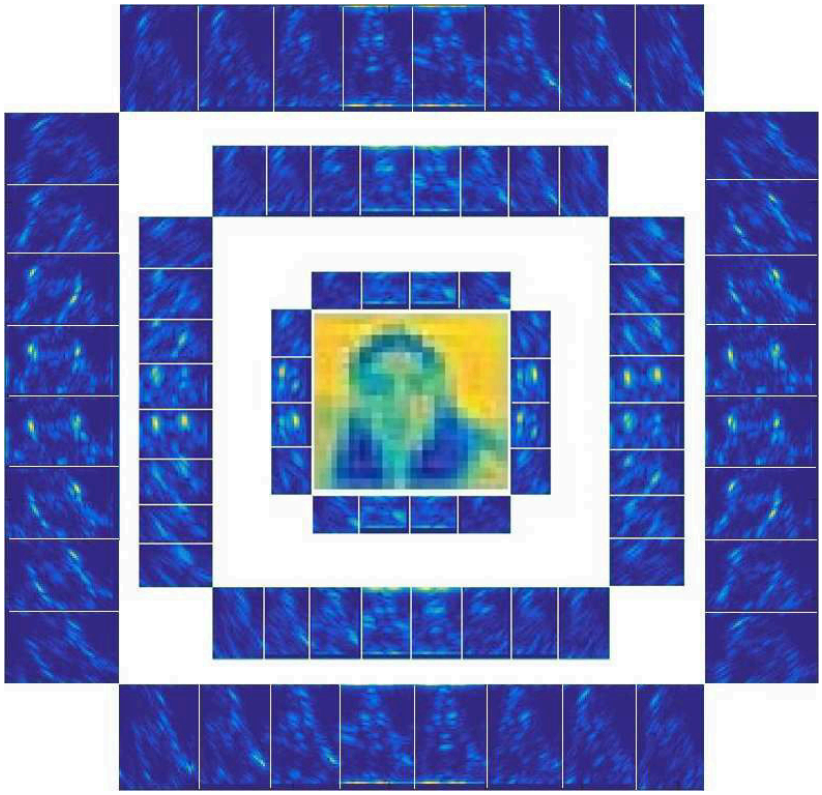
3.2.1.2 Three Orthogonal Planes vs Pairwise Orthogonal Planes

The first integrated methodology implemented in terms of this PhD was the one published in Pampouchidou et al. [159] and involved the LCBP-TOP. As mentioned previously, the idea of this preliminary version derived from the LGBP-TOP proposed by [26], with the difference that in the present Gabor wavelets are replaced with Curvelet transform [46].

The norm for approaches based on Orthogonal Planes is to take all descriptors from each of the three planes (XY, XZ, YZ), where Z refers to the time, and concatenate them all together, producing a vast vector of thousands of features. In the work published in Pampouchidou et al. [160] an effort was made to reduce this high dimensionality, by introducing the Pairwise Orthogonal Planes. Thus, in [160] two approaches were tested: a) *Frame-based Classification*, which considered only XY planes for each frame separately, and b) *Video-based Classification* for which XZ and YZ planes are considered in pairs of two. This way, for the Video-based approach the plane corresponding to the first row is combined with the plane corresponding to the first column, second row with second column, (...), and the last row with the last column.

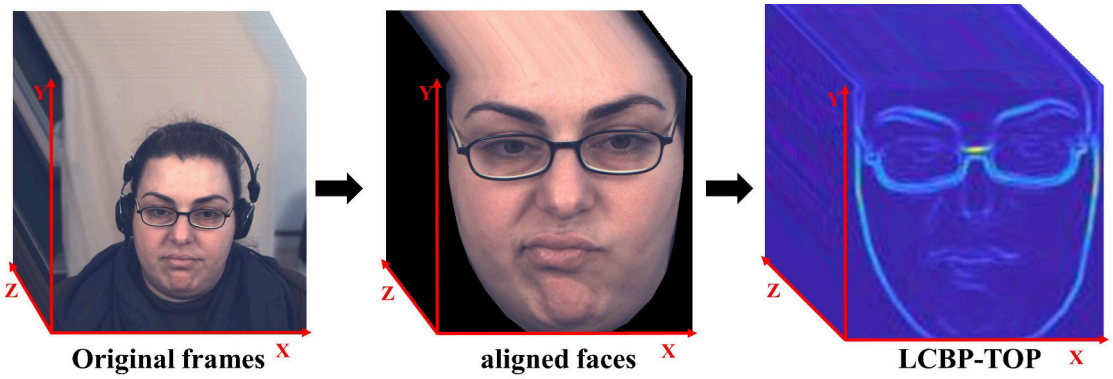
The LCBP-POP modification has the advantage of preserving the motion information in both axis, with a considerably shorter feature vector. A further improvement in compare to the previous approach was that overlapping window was used instead of the previous that employed sequential windowing. In addition, apart from facial expression, curvature contains information on person-specific biometrics (e.g. different shapes of facial features, varying symmetry), as well as occlusions (e.g. facial hair, eye-glasses).

(a) Original Image

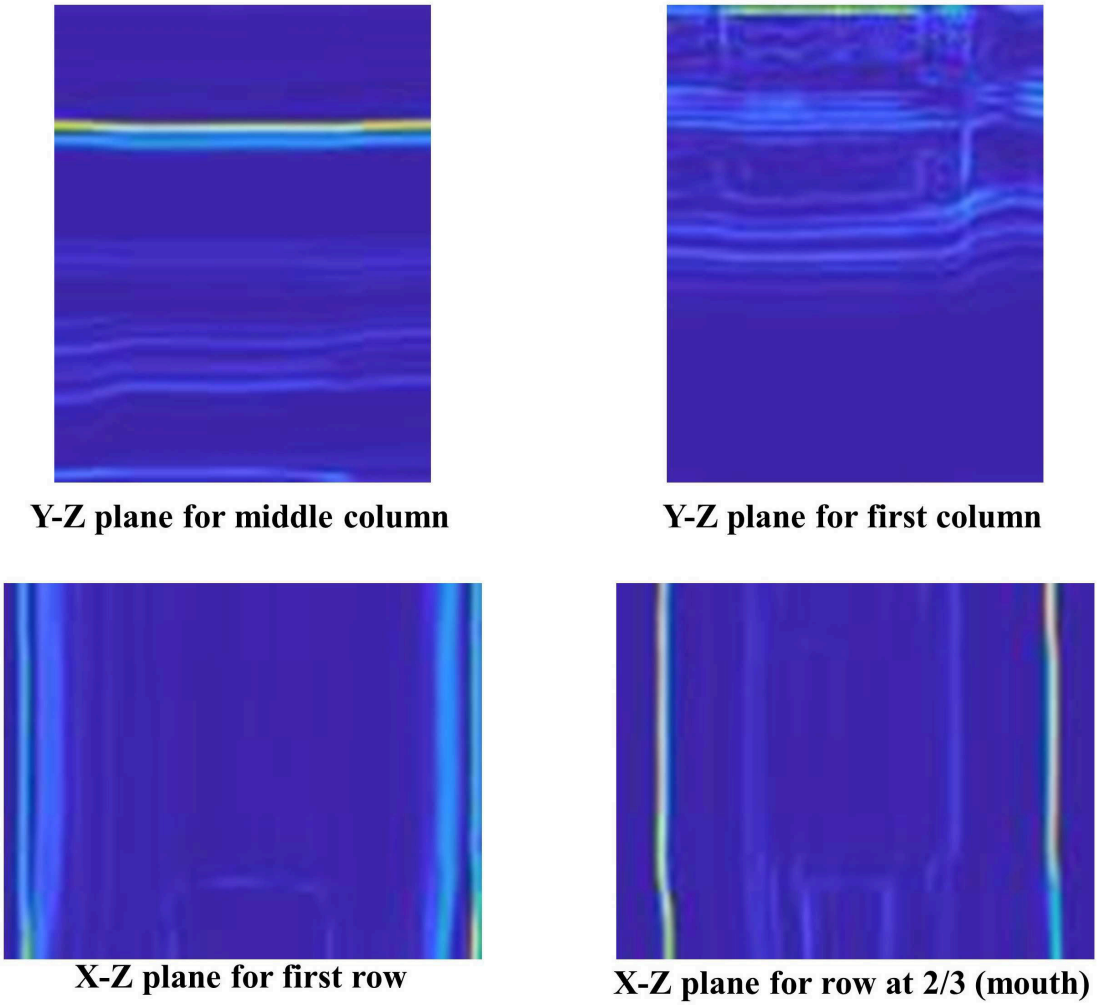


(b) The coarsest level (scale=1,orientation=1) is in the center, while the number of orientations doubles for every second scale

Figure 3.6: Example of curvelet pseudo-images wrapped by CurveLab toolbox, for scale=4, and orientation=4

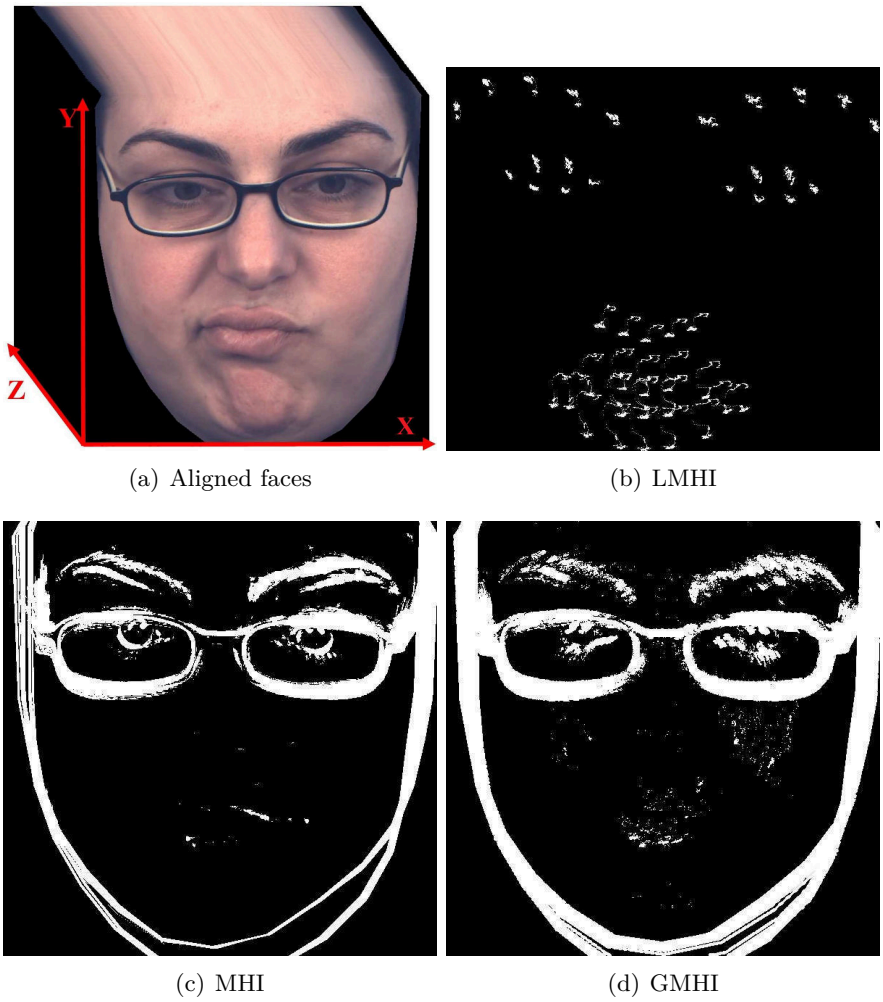


(a) Flow from image frames to Curvelet transform



(b) Examples for X-Z and Y-Z planes

Figure 3.7: Flow Curvelet transform on facial video frames, and examples for X-Z and Y-Z planes



(a) Aligned faces (b) LMHI

(c) MHI (d) GMHI

Figure 3.8: Visual comparison of motion history image variants, as extracted from the specific aligned images sequence

Video-based Classification was implemented in order to overcome this limitation. The LBP descriptor was again computed on each plane.

3.2.2 Motion History Image

It is already established that most of the non-verbal signs of depression are dynamic by nature [76, 222]. Therefore, the use of video-based methods (dynamic), as opposed to frame-based (static) is preferable. In addition to the previously described motion representation methods, three different motion history images were also implemented: a) *MHI* as derived from the basic algorithm, b) *LMHI* which relies on facial landmarks,

and c) *GMHI*.

The LMHI algorithm was introduced as part of our DSC-AVEC'16 participation with Pampouchidou et al. [161], which instead of considering intensities from image sequences, considers sequences of facial landmarks. In addition a quantitative comparison of different motion history variants was published in Pampouchidou et al. [162] (GMHI). More details regarding the specific motion representation algorithms are presented below with implementation examples illustrated in Fig. 3.8.

MHI is a robust, yet relatively straightforward, algorithm developed to represent the motion that occurs in the course of a complete video recording with a single image [20]. The algorithm produces a grayscale image, in which the white pixels correspond to the most recent movements and the darkest gray correspond to the earliest motion elements. Black pixels indicate absence of movement. It is a popular algorithm for motion analysis [20], and has been extensively used in the field of human action recognition [39].

An early approach of MHI to facial image analysis was that of Valstar et al. [207], who employed MHI in facial action recognition from videos. Meng et al. [148] published a continuous depression assessment approach in their participation to the DSC of AVEC'13; they proposed an extension of MHI, the MHH, which considers patterns of movement. In the DSC of AVEC'14 Pérez Espinoza et al. [167] employed MHI, and for the same challenge Jan et al. [108] proposed the 1-D MHH, an extension of MHI, which is computed on the feature vector sequence instead of the intensity image.

3.2.2.1 Original MHI

The MHI is a gray scale image, where white pixels correspond to the most recent movement in the video, intermediate gray scale values to corresponding less recent movements, and black pixels to the absence of movement. The MHI algorithm, with slight variations as explained next, is applied on the aligned face image sequences derived from the preprocessed data using OpenFace.

The MHI H, with a resolution equal to the one of the aligned faces, is computed based on an update function $\Psi(x, y)$ as follows:

$$H_i(x,y) = \begin{cases} 0 & i = 1 \\ i \cdot s & \Psi_i(x,y) = 1 \\ H_{(i-1)}(x,y) & otherwise \end{cases} \tag{3.5}$$

where $s = 255/N$, N the total number of video frames, (x, y) the position of the corresponding pixel, and i the frame number. $\Psi_i(x, y)$ represents the presence of movement, derived from the comparison of consecutive frames, using a threshold ξ:

$$\Psi_i(x,y) = \begin{cases} 1 & D_i(x,y) \geq \xi \\ 0 & otherwise \end{cases} \tag{3.6}$$

where $D_i(x, y)$ is defined as a difference distance:

$$D_i(x,y) = \left| I_i(x,y) - I_{(i-1)}(x,y) \right| \tag{3.7}$$

$I_i(x, y)$ is the pixel intensity value in (x, y) at the ith frame. The final MHI is the $H_N(x, y)$.

3.2.2.2 Landmark Motion History Image

LMHI encodes the motion of the facial landmarks into a grayscale image, with the most recent movement corresponding to white pixels, the earliest corresponding to the darkest gray, and temporally intermediate movements indexed by corresponding gray values. The extension of the proposed work in comparison to that of Ptucha and Savakis [170] is that the in-between motion is also preserved with the use of respective gray-scales, which is important for the descriptors applied later on the LMHI.

The landmarks considered are the ones that correspond to the facial features (eyes, eyebrows, nose-tip, and mouth), while the face outline is excluded. This step was taken in order to emphasize inner-facial movements, and ignore the overall head movements. This is achieved by co-registering the involved landmarks using affine transformation before computing the LMHI, through alignment of the points corresponding to the temples, chin, inner and outer corners of the eyes (landmarks {1, 9, 17, 37, 40, 43, 46}).

LMHI differs from the conventional MHI in that image intensities are not considered, but only the facial landmarks, which are detected in each frame. The adopted LMHI algorithm is similar to MHI, by maintaining the same H_i as in (3.5), and modifying Ψ_i as follows:

$$\Psi_i(x,y) = \begin{cases} 1 & (x,y) \in L_i \\ 0 & otherwise \end{cases} \tag{3.8}$$

where L_i corresponds to the selected landmarks as detected in the ith frame.

3.2.2.3 Gabor Motion History Image

Gabor filters have been frequently used in both facial expression analysis and emotion recognition [78, 203]. In relevant approaches the feature vector is extracted from the convolution of the original image with a 2D Gabor wavelet function at different orientations and wavelengths. This describes the spatial frequency structure around each pixel. In [58] the Gabor energy was used for facial emotion recognition, which gives a smoother response to an edge or a line of appropriate width with a local maximum exactly at the edge or in the center of the line. The authors also applied background texture suppression on the response of the filter, by removing an image filtered by the difference of Gaussians (DoG) from the original response for each orientation. This approach, also known as anisotropic inhibition [94], removes noise and provides a sharper representation of facial features.

GMHI is another variant of MHI, where Gabor inhibited images substitute original image intensities. The motivation for implementing this variant is that it focuses on the important details of the facial features, and thus extracts the most relevant information. The motion representation algorithm is identical to the one described in subsection 3.2.2.1, but the input image I is the result of the Gabor inhibition. The process of obtaining the Gabor inhibited image is explained in detail below.

The Gabor wavelet at position (x, y) is given by:

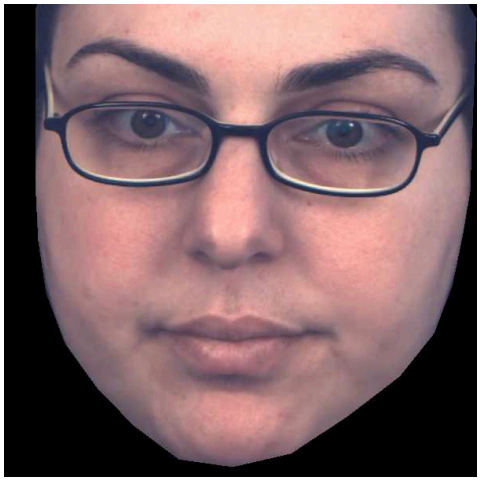
$$\Psi_{\lambda,\theta,\phi,\sigma,\gamma}(x,y) = exp(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2})cos(2\pi\frac{x'}{\lambda}+\phi) \tag{3.9}$$

with

$$\begin{aligned} x' &= xcos\theta + ysin\theta \\ y' &= -xsin\theta + ycos\theta \end{aligned} \tag{3.10}$$

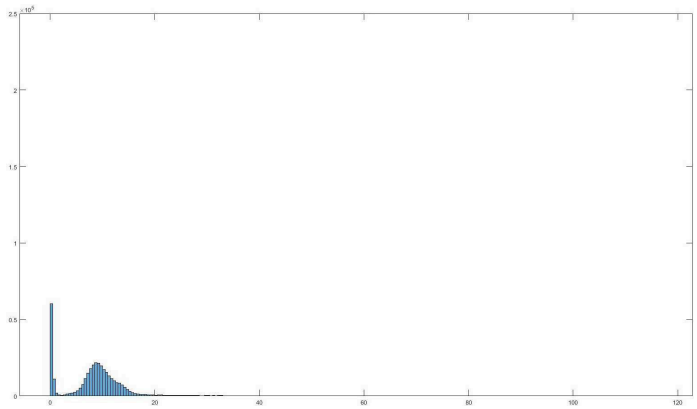
where λ stands for the wavelength, θ for the orientation, ϕ for the phase offset, σ for the standard deviation of the Gaussian, and γ for the spatial aspect ratio [61].

The input image is usually filtered with many wavelets for multiple orientations and wavelengths. The energy filter response is obtained by combining the convolutions obtained from two different phase offsets ($\phi_0 = 0$ and $\phi_1 = \pi/2$) using the $L2$-norm. Background texture suppression is applied on the filter response, by removing a Difference of Gaussians (DoG) filtered image from the original response for each orientation [94]. Finally, the mean response of Gabor filtering is used to combine the responses across



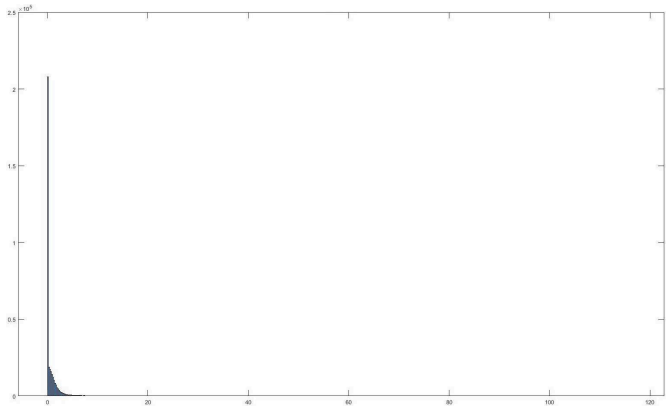
(a) Facial image

(b) Gabor filtered image



(c) Histogram of the Gabor filtered image

(d) Inhibited image



(e) Histogram of the inhibited image

Figure 3.9: Example for the computation of Gabor inhibited filtering on a facial image, along with the corresponding histogram to illustrate the noise removal. As it is obvious in the second histogram (after the inhibition) the values are concentrated and the contrast is enhanced in order to provide a more meaningful outcome.

the different orientations, resulting in the pseudo-image used to compute the GMHI. An example of applying the common Gabor and the Gabor inhibited algorithms to an aligned face image is illustrated in Fig. 3.9, where the Gabor inhibited image appears to be sharper and with less texture in uniform regions than the original Gabor response.

3.2.3 Time-series of Geometrical Features

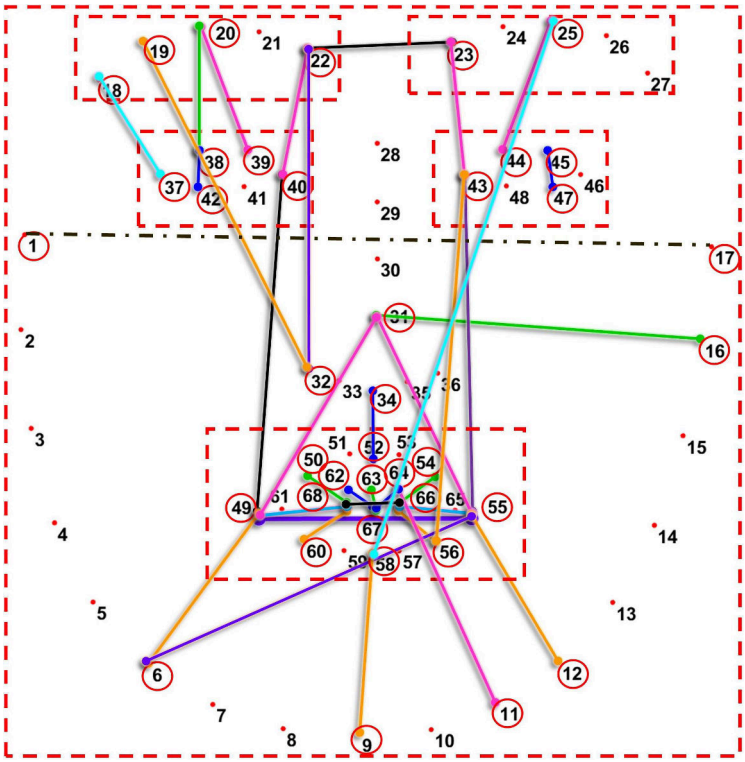
Geometric features are quite popular in the field of facial expression recognition, while relevant approaches using these features in depression assessment have also been found. Given that smiling tends to be reduced in individuals suffering from depression [76], we focused on certain distances between facial landmarks, which are affected by smiling [216]. Landmarks that correspond to the Veraguth fold were also considered, as well as blinks. Right and left eye width, as well as mouth width and height shown in figure 3.10, were also considered in time-series manner for further feature extraction. Taking all these into account we defined the final set of distances, illustrated in Fig.3.10.a, considered as time-varying parameters. In order to account for changes of the distances due to external movements (not intra-facial), all distances were normalized based on the distance between the side temples (see Fig.3.10.a landmarks 1 and 17).

An additional video based feature set was subsequently constructed. This feature set is based on landmarks' activity, in terms of displacement, velocity, and acceleration, that were estimated for all 68 landmarks for a specified window of frames. The landmarks were grouped according to facial features, as illustrated in Figure 3.10.b: right eyebrow, left eyebrow, right eye, left eye, mouth, and face as a whole. The values of displacement, velocity, and acceleration, were averaged, resulting in three corresponding time-series for each region. Statistical measures extracted are the same as with the distance time-series.

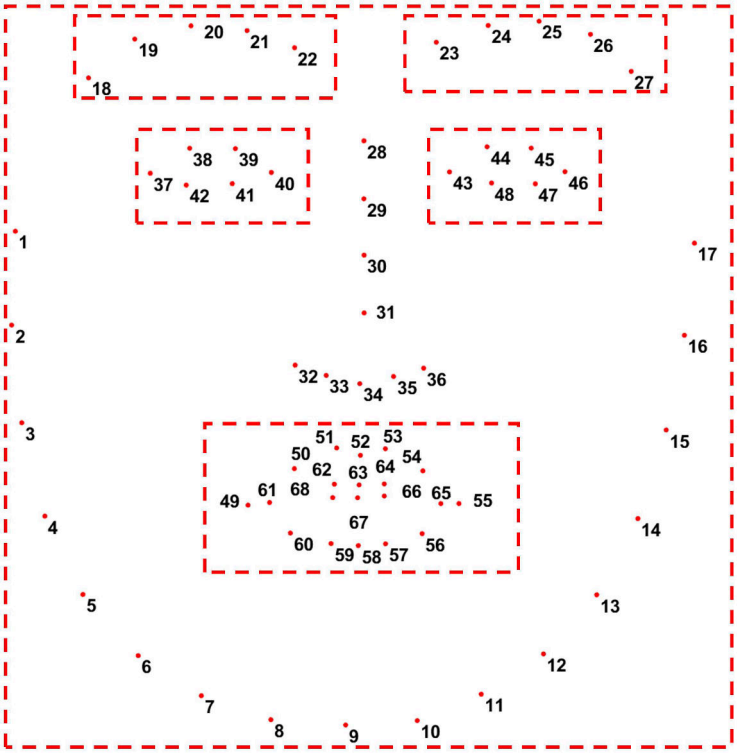
Distances, both in space and time, were computed based on Euclidean distance:

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 + y_2)^2} \tag{3.11}$$

Distance in time implies considering the coordinates of one landmark at a given time frame and the coordinates of the same landmark for a next time frame; in this manner the displacement of a specific landmark is estimated. While for distance in space the coordinates of two different landmarks are considered in the same time frame.



(a) Distances considered in time-series for geometrical features



(b) Regions of facial features considered for landmark motion features

Figure 3.10: Geometric features considered in time-series

3.3 Feature Extraction

Following motion representation, the need to construct meaningful descriptors comes next. In the cases of LCBP-TOP and LCBP-POP the LBP was used (experiments 4.2 and 4.3). In the case of the different variants of motion history image several appearance based descriptors were employed (experiments 4.4 and Experimental:Exp5). Pre-trained deep neural networks were also used through transfer learning for motion history images for both the preliminary experiment in subsection 4.6 and the main study in subsection 5.2.1. Finally, for time-series representation statistical descriptors were used (experiment 4.5). In the following subsections the different feature extraction algorithms are described in some level of detail.

3.3.1 Appearance-based Descriptors

Appearance descriptors employed hereby exploit intensity and more specifically texture based attributes. The appearance-based descriptors employed in terms of the present work include the LBP, Local Phase Quantization (LPQ), HOG, and the conventional image histogram. The descriptors are explained within the next subsections and are illustrated in Fig. 3.15 for the example of MHI.

3.3.1.1 Local Binary Patterns

LBP were introduced by Ojala et al. in [152], who extended Wang and He's work [220] in texture classification. At their first attempt [152] they obtained gray-scale invariance, while later [153] they also achieved rotation invariance, yet keeping the algorithm computationally simple and efficient.

LBP [153] entails dividing the image into partially overlapping cells. Each pixel of the cell is compared to its neighbors to produce a binary value (pattern). The resulting descriptor is a histogram which represents the occurrence of different patterns. LBP for two sets of {radius, neighbourhood} results to feature vectors of size 1×59 for {1,8} and size 1×243 for {2,16}. The example of the LBP pipeline is illustrated in Fig.3.11.

When LBP is applied the image is divided into cells of a given size, e.g. 16×16. For each pixel in the cell LBP descriptor is computed according to the example in Fig.3.12.a as follows. The central Pixel is compared to all of its neighbours, depending on the neighbourhood and radius defined, and each of the neighbours are assigned a binary

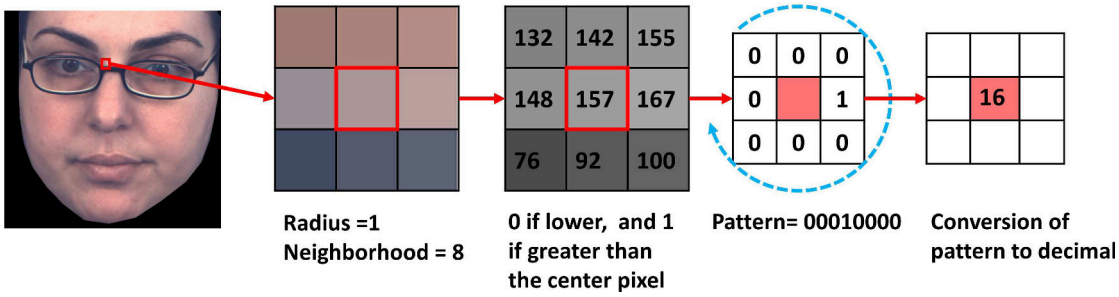


Figure 3.11: Example of LBP pipeline

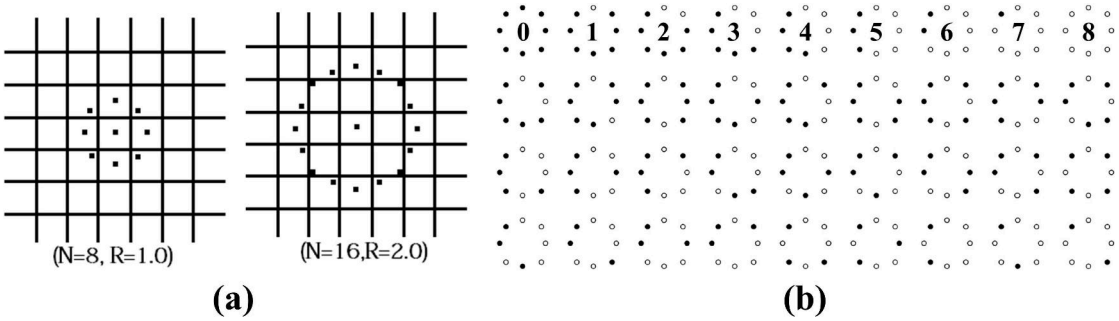


Figure 3.12: Illustration of LBP patterns (Taken from:[153]): a) LBP pattern for different values of neighbourhood and radius b)Rotation invariant LBP patterns

value: a) one if their value is greater than the central, and b) zero if it is lower. In other words, the central pixel behaves as a threshold. The resultant digits are taken in the same order (either clockwise or counter-clockwise), providing with a binary number converted to decimal for convenience.

Ojala et al. in [153] proved that for neighbourhood 8 and radius 1 there exist 36 unique different orientations for the LBP descriptor as illustrated in Fig.3.12.b, with the first 9 being uniform. Uniform in the case of LBP is interpreted as at most two bitwise transitions. For instance the descriptor "00000000" has zero transitions, "00000111" has just one transition, 00111000 has two transitions, while 01001101 has five transitions.

The first nine descriptors in Figure 3.12b first row are all uniform, each of which stands for a type of contour, such as the first one (0-valued) is a bright spot, the last one of the row (8) is a dark spot, while the intermediate (1-7) represent different types of edges. In terms of experiments 4.2, 4.3, and 4.4 University of Oulu implementation[1] was used, while for experiments 4.6 and 5.2.1 the MATLAB implementation was used[2].

[1]http://www.cse.oulu.fi/wsgi/MVG/Downloads/LBPMatlab
[2]https://www.mathworks.com/help/vision/ref/extractlbpfeatures.html

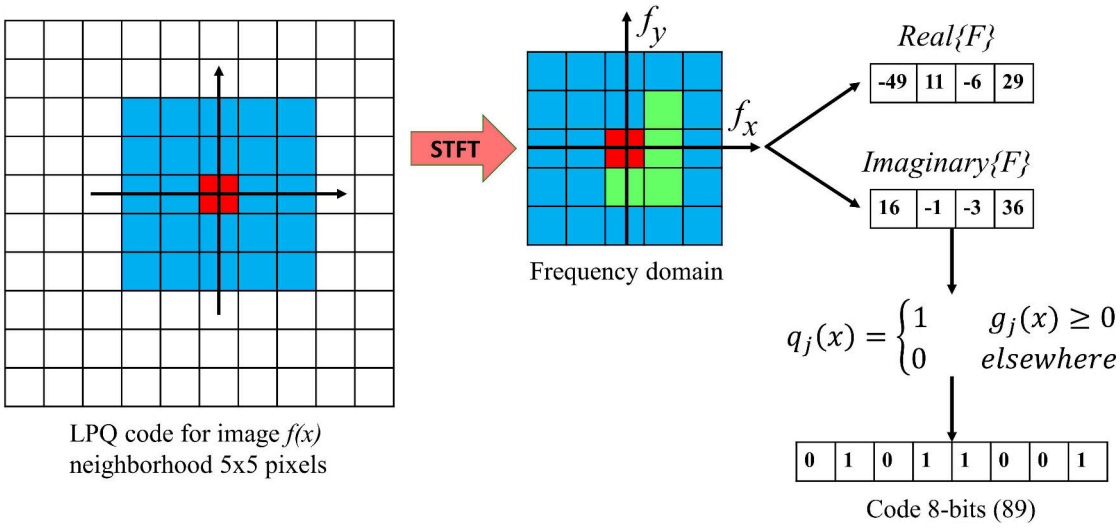


Figure 3.13: Example of the LPQ pipeline

The different implementations were employed based on the current availability of the respective toolboxes.

3.3.1.2 Local Phase Quantization

LPQ [154] is computed in the frequency domain, based on the Fourier transform, for each pixel. Local Fourier coefficients are computed, while their phase information results in binary coefficients after scalar quantization. The final descriptor corresponds to the histogram of the binary coefficients. More specifically, for an $N \times N$ image $f(x, y)$ the local phase is computed using the short-term Fourier transform (STFT) based on the following equation:

$$\hat{f}_{u_i}(x) = (f * \phi_{u_i})(x) \tag{3.12}$$

ϕ_{u_i} is a complex valued $m \times m$ mask defined in the discrete domain by:

$$\phi_{u_i} = \left\{ e^{-j2\pi u_i^T y} | y \in \mathbb{Z}^2; \|y\|_\infty \leq r \right\} \tag{3.13}$$

where $r = (m-1)/2$, and u_i a 2-D frequency vector. STFT for the specific implementation of LPQ[101] is computed at four frequencie points: $u_1 = [\alpha, 0]^T$, $u_2 = [0, \alpha]^T$, $u_3 = [\alpha, \alpha]^T$, and $u_4 = [\alpha, -\alpha]^T$, where $\alpha = 1/m$. The implementation used for LPQ feature extraction was the one from the University of Oulu [1][101]. An illustrated example

[1] http://www.cse.oulu.fi/wsgi/MVG/Downloads/LPQMatlab

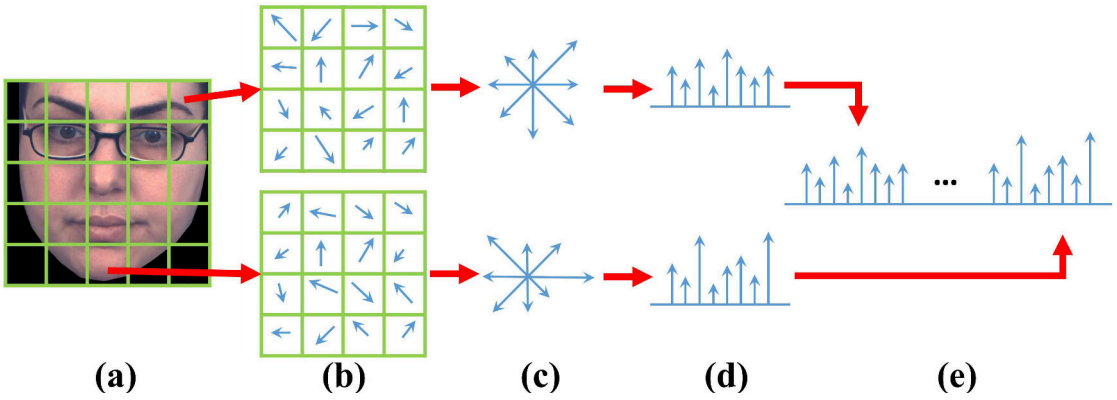


Figure 3.14: Example of HOG pipeline

for the LPQ pipeline is shown in Fig.3.13 in similar manner with [34].

3.3.1.3 Histogram of Oriented Gradients

HOG [62] entails counting gradient orientations in a dense grid. Each image is divided into uniform and non-overlapping cells, the weighted histogram of binned gradient orientations for each cell is computed, and subsequently combined to form the final feature vector. More specifically, for an image L with size $N \times N$ pixels, divided in cells as shown in Fig.3.14.(a), the orientation θ for each pixel $p = (p_x, p_y)$ is computed (c.f Fig.3.14.(b)) based on the following equation:

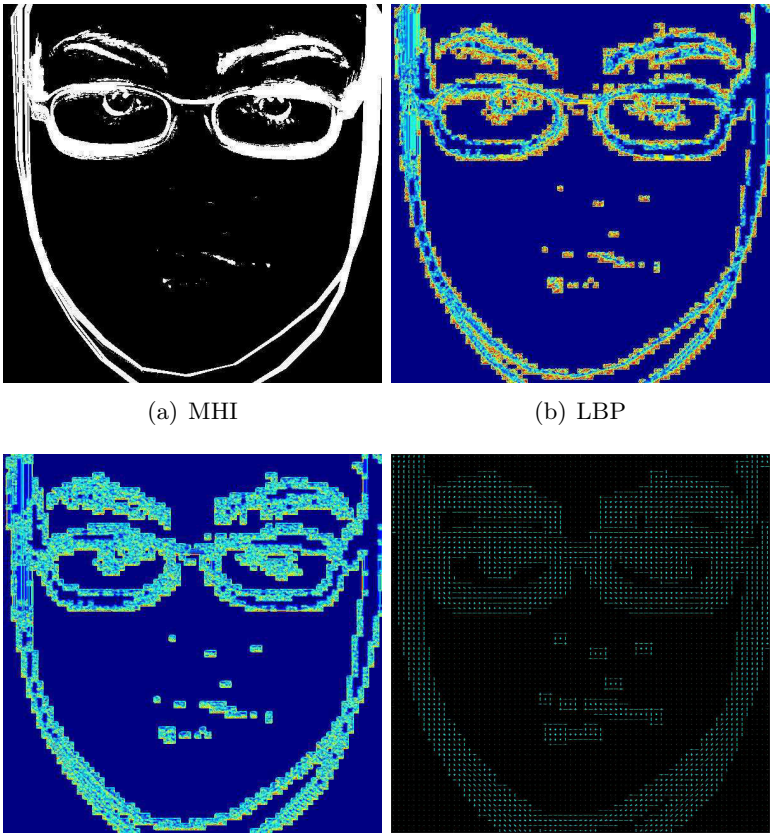
$$\theta(p) = tan^{-1} \frac{L(p_x, p_y + 1) - L(p_x, p_y - 1)}{L(p_x + 1, p_y) - L(p_x - 1, p_y)} \tag{3.14}$$

The estimated orientations are accumulated in a histogram of a predetermined number of bins (c.f. Fig.3.14.(c)-(d)). The output corresponds to the concatenated individual histograms resulting in a single spatial HOG histogram as shown in Fig.3.14.(e) [47]. The MATLAB implementation was used in terms of the experimental tests in subsections 4.4, 4.6, and 5.2.1. [1].

3.3.1.4 Image Histogram

Additionally, the combined histogram, mean and standard deviation of the motion-image gray values were also considered as a single descriptor [Hist-Mean-Std]. Specifically for

[1] https://www.mathworks.com/help/vision/ref/extracthogfeatures.html



(a) MHI (b) LBP

(c) LPQ (d) HOG

Figure 3.15: Visualization of appearance-based features

the histogram, zero values (absence of movement) are disregarded, and only the bins of the remaining 255 gray values are considered, with the addition of mean and standard deviation. This feature was not visualized like the rest appearance-based features as it does not entail spatial distribution. Again, the MATLAB implementation was used[1].

3.3.2 Transfer Learning from Pretrained Networks

Deep learning, which has become increasingly popular during recent years, is a self-learning tool designed to identify patterns in several sets of data samples, extracted from multiple processing layers. Each layer is composed by representation-learning methods, and is processed in a higher and more abstract level [131]. Convolutional Neural

[1]https://www.mathworks.com/help/images/ref/imhist.html

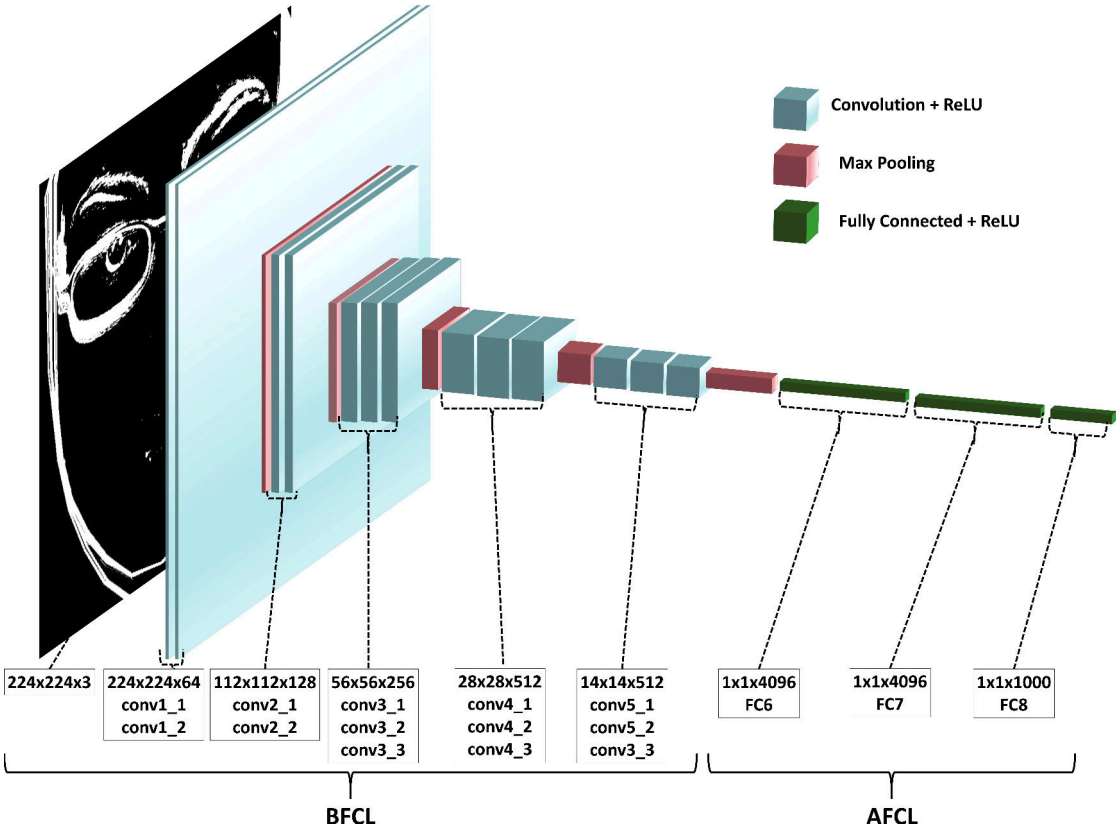


Figure 3.16: Architecture of VGG16

Networks (CNN) is a particular deep feedforward network with higher generalization efficiency than other fully connected networks. There are typically 2 types of layers: the convolutional layer, and the pooling layer. In the convolutional layer, all units are connected to the weights (also known as filter banks), while the weighted sum is inserted to the Rectified Linear Unit (ReLU).

The CNN architecture used in the present work was employed in the participation that won the 2014 ImageNet competition (ILSVRC) [173] [130]. Visual Graphic Geometry (VGG) is a CNN variant proposed by Simonyan and Zisserman [188]. Using VGG, they achieved 92.7% top-5 test accuracy on the ImageNet Dataset, which comprises over 14 million images in 1000 classes. In the proposed work VGG16 and VGG19 are employed. Fig. 3.16 illustrates the microarchitecture of VGG16. It consists of 16 layers with 13 convolutional layers, while VGG19 has 19 layers with 16 convolutional layers. Both VGG16 and VGG19 have 3 fully connected layers, and all convolutional layers

Figure 3.17: Visualization of Pool Relu1_1 activations

involved have filters of size 3-by-3.

The RGB image, with pixel values ranging between 0–255, is normalized by subtracting the mean pixel value. The input to VGG (a fixed-size 224×224 RGB image) passes through a stack of convolutional layers, where the very small filters are of receptive field size 3×3 to capture the notion of left/right, up/down, and center. The convolution stride is fixed to 1 pixel; the spatial padding of a convolutional layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1 pixel for 3×3 convolutional layers.

Spatial pooling is carried out by five max-pooling layers, which follows some of the convolutional layers (not all the convolutional layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride 2. A stack of convolutional layers (which has a different depth in different architectures) is followed by three Fully Connected layers: the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the soft-max layer. The configuration of the fully connected layers is the same

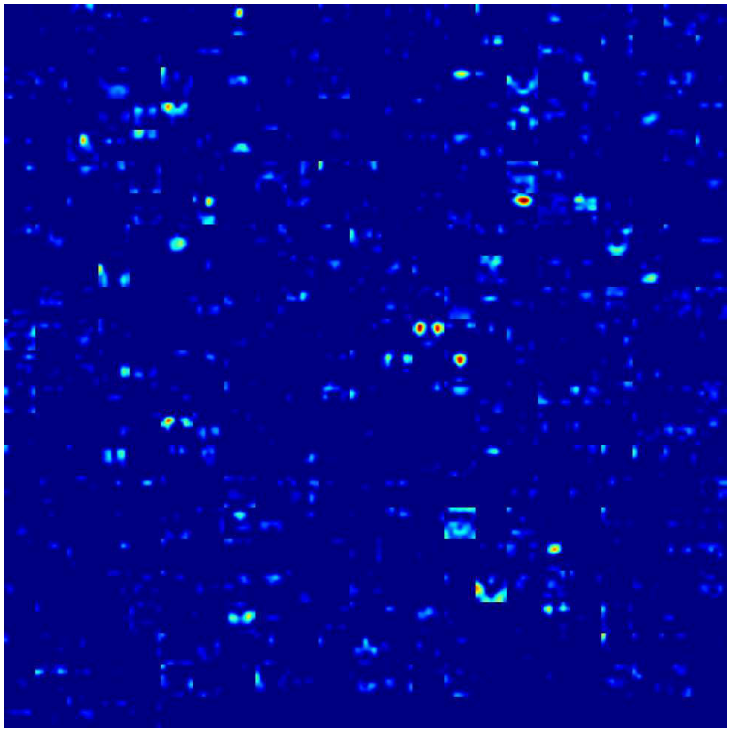


Figure 3.18: Visualization of Pool 5 activations

in all networks. All hidden layers are characterized by non-linearity afforded by ReLU [130].

In the present work a pre-trained VGG16 network was employed for each motion history image separately. The particular version of the network was chosen, as it has shown excellent results in related medical applications. Again, different implementations were employed based on the current availability of the respective toolboxes. Two different implementations have been employed during experiment 5 (c.f. subsection 4.6): a) Before Fully Connected Layer (BFCL), and b) After Fully Connected Layer (AFCL); in the main study (c.f.4.6) only the AFCL was used. BFCL provides the features to the fully connected layer of the VGG16, as shown in Fig. 3.16. In AFCL there are three fully connected layers in VGG16. Layers 1 and 2 operate on a feature matrix of size 1×4096 and layer 3 on a feature matrix of size 1×1000.

In order to achieve optimal outcome from deep neural networks training and fine tuning is required. However in order to have accurate training a high number of training samples is needed, in the order of more than 100,000. In clinical applications, and espe-

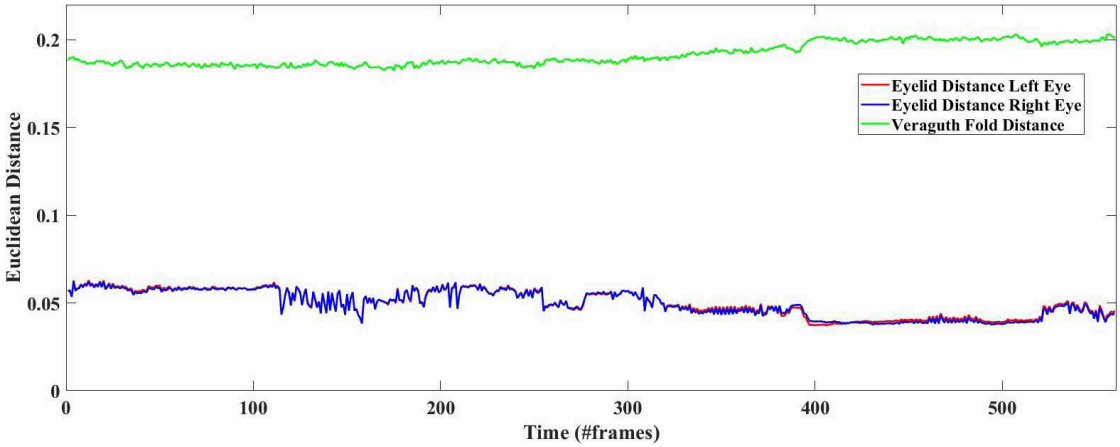


Figure 3.19: Example of time-series from eye region distances

cially in mental health related fields, obtaining data is not so straight-forward. Thus, in the proposed work the pretrained networks from the MATLAB Neural Network Toolbox[1] were utilized; this technique is known as transfer learning. In particular the MATLAB model used hereby is trained on a subset of the ImageNet database (from ILSVRC). Both VGG16 and VGG19 are trained on more than a million images, for classifying 1000 object categories (e.g. keyboard, pencil, animals, etc), enriched by a wide range of feature representations. Figures 3.17 and 3.18 illustrate the processing outcome from ReLU1_1 and Max Pooling level 5.

3.3.3 Statistical Descriptors

The time-series derived based on the motion representation method described in subsection 3.2.3 (e.g. Fig. 3.19) were further processed as signals to extract several statistical features. The measures selected included mean, median, mode, range, standard deviation, variance, skewness, kurtosis, energy, entropy, correlation, and interquartile range [18]. Additional statistics included max, min, mad, mean frequency, and band power [166]. MATLAB implementations were employed for extracting all the statistical descriptors.

[1] https://www.mathworks.com/help/nnet/ug/pretrained-convolutional-neural-networks.html

3.4 Machine Learning

The machine learning level is concerned with exploiting the extracted features in order to train models that generalize well and are able to recognize signs of depression. Given the high dimensionality of the features there is the need to employ a reduction algorithm. Further, before the training/testing of the models there is another issue to address, this of the cross validation which is used so as to deal with less biased measures. Finally, in the proposed work two approaches were investigated, the categorical (classification) and the continuous (regression). Next, the different steps are described.

3.4.1 Dimensionality Reduction with PCA

In the present work, Principal Component Analysis (PCA) was employed to achieve dimensionality reduction. PCA is one of the most popular methods for this purpose, and is based on the linear transformation of the original feature vector, into a set of uncorrelated principal components. For a data set of size $N \times M$ (i.e., N samples and M features) PCA identifies a $M \times M$ coefficient matrix (component loadings) that maps each data vector from the original space to a new space of M principal components. However, by properly selecting a smaller set of $K < M$ components, the dimensionality of the data can be reduced while still retaining much of the information (i.e., variance) in the original dataset.

3.4.2 Cross Validation

Cross validation is a technique employed in order to establish the experimental conditions is such manner that minimum possible bias is introduced to the models. For this reason the data samples are partitioned to train and test sets, so that samples that are used for training are not included in the testing process. The performance of a model can be objectively evaluated when it is introduced with unseen data. Furthermore, by experimenting with several configurations of train/test sets the outlier effect on the performance evaluation is reduced.

Some techniques employed by the proposed work for performing cross validation involve the k-folds, Leave-One-Out (LOO), and Leave-One-Subject-Out (LOSO). Implementation of k-folds partitions the data samples in k sets, e.g. for $k = 10$ the dataset is divided in 10 different non-overlapping subsets, each time one subset is kept out, while

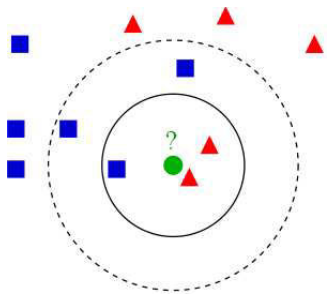


Figure 3.20: Example of kNN. Taken from:[11]

the rest are used for training the model, and afterwards tested on the left-out sample, and this process is repeated 10 times. LOO on the other hand holds out one sample every time, trains with the rest, and again tests the performance on the left-out sample, and this process is repeated N times, where N the total number of samples. Finally, LOSO is a particular type of LOO, and applied in cases where more than one samples of a given subject/participant are included in the dataset. This process is repeated as many times as the number of subjects included in the dataset (rather than samples as in LOO) each time all samples of the given subject are held out from the training set, and then used for testing the performance.

3.4.3 Gender Dependency

Gender-based classification for depression has been reported to substantially improve performance [23] [194]. In the present work, in addition to gender-independent classification/regression, a gender-based model was also implemented by building two separate classifiers, one for male and another for female participants. The classifier for male subjects was trained on feature-sets extracted from data of male participants and the female classifier with feature-sets extracted from data of female participants.

3.4.4 kNN

kNN is considered to be one of the fastest and simplest algorithms commonly used for supervised learning. kNN considers the known feature vectors, and for every new vector computes the Euclidean Distance to find the nearest known class, based on the Equation 3.11. The class is computed according to the nearest k neighbours, that is to which it has the smallest Euclidean distance. k, the number of neighbours has to be specified.

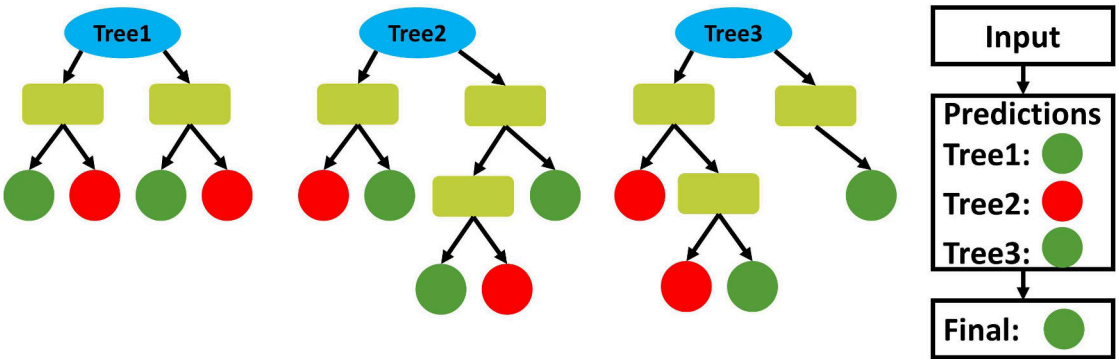


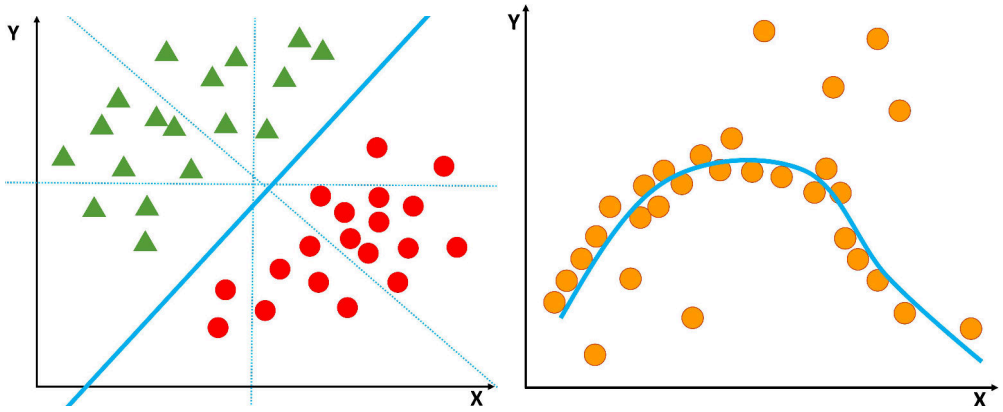
Figure 3.21: Example of RF[1]

However, kNN is well known for its sensitivity to the data local structure [83] [72]. For example at Fig. 3.20, there is a two-class classification problem, red-triangles and blue-squares, and the new object, green-circle, needs to be classified in either of them. The two concentric (on the new object) circles correspond to $k = 3$, and $k = 5$; it is apparent how the classification results is influenced by this slight change. In the case $k = 3$, the new object would be classified as a red-triangle, while on the case $k = 5$ to the blue-squares. Based on this simple example, and how the classification result changes based on k value, it is obvious that the selection of k can be critical. Due to this fact, for the tests explained in the next Chapter different values of k were tested, in order to evaluate their performance and relevance to the depression assessment. In this work, the MATLAB implementation was employed for conducting the experimental tests[2].

3.4.5 Random Forest

Random Forest is a flexible, popular, easy to use supervised machine learning algorithm. The idea behind RF is creating an ensemble of decision trees in a random manner, as its name suggests, based on the "bagging" method, which is nothing more than combining learning models in order to increase the overall performance. More specifically, RF builds a series of decision trees and combine their outputs for a more accurate prediction[2] (c.f. Fig.3.21). The implementation employed in the present work is the one from MATLAB[3].

[2]https://www.mathworks.com/help/stats/fitcknn.html
[2]https://machinelearning-blog.com/2018/02/06/the-random-forest-algorithm/
[3]https://www.mathworks.com/help/stats/fitctree.html



(a) Classification (b) Regression

Figure 3.22: Example of SVM for both classification and regression approaches

3.4.6 SVM

SVM is another supervised machine learning algorithm that can be employed for both classification and regression problems. In terms of classification SVM is intended for binary problems, with the main purpose to find an optimal hyper-plane that separates samples of the two classes in feature space. In the example of Fig.3.22.a several hyper-planes have been tried (dashed lines), yet only one (the thick line) is chosen as it optimally differentiates the two classes. An additional consideration when choosing the optimal hyper-plane is establishing maximal margins from the samples.

Support Vector Regression (SVR) is in generally based on the same principles with the SVM that is used for classification, yet it encompasses some differences. The regression nature of the algorithm is intended in predicting real-value numbers, as opposed to discrete categories, which has infinite possibilities. Therefore it is a more complicated problem, which targets the construction of a function that by minimizes the error between the actual and the estimated value (c.f. Fig.3.22.b). Implementations for SVM are from MATLAB both for classification[1] and regression[2] approaches.

[1]https://www.mathworks.com/help/stats/fitcsvm.html
[2]https://www.mathworks.com/help/stats/fitrsvm.html

3.5 Fusion

Many types of features were extracted, such as appearance-based features, transfer-learning-based VGG features, as well as the statistical features from time-series of geometrical features. In an attempt to test the additional value of the different descriptors, fusion of the features was tested. Specifically, the feature-level fusion was used, which implements the concatenation of the different vectors, in several combinations. Feature-level fusion was employed in experiment 3 [161] (subsection 4.4), experiment 5 [162] (subsection 4.6), and in the main study (subsection 5.2.1).

3.6 Summary

The employed pipeline was described in this Chapter, along with the several methods used for the different steps. Specific experimental setups employing the described algorithms are presented in the next Chapter.

Chapter 4

Preliminary Experimental Evaluation

In this chapter the different preliminary experimental tests are presented, with the specific configuration of each methodology, as well as the derived results. The data collection was a long-term process, which took more than 4 years to be completed, in the meanwhile the several proposed algorithms and methodologies were tested in available benchmark datasets. The preliminary experiments took place in order to provide some insight on the proposed methodologies, until the data collection was completed. Next the two benchmark datasets employed for the preliminary tests are described, followed by the description of the respective experiments.

4.1 Employed Datasets

In spite the fact that automatic depression assessment is highly desirable, clinical data are not open to the research community, given the sensitivity of personal data involved. However, two datasets with depression annotation, both based on self-reports and volunteer non-diagnosed participants, were made available in terms of the AVEC challenges in 2013, 2014, and 2016 respectively. Both require a signed End-User License Agreement before granting permission for download access.

4.1.1 AVEC

This dataset was introduced in the 3rd and 4th AVEC [210], at the time being the only freely available dataset, annotated for depression, which included video recordings of participants. The dataset consisted of video feeds of undiagnosed volunteers performing the following tasks in the German language: vowel pronunciation, solving a task out loud, counting from 1 to 10, reading novel excerpts, singing, and describing a scene displayed in pictorial form. During AVEC'13 the complete recording was used for testing the performance of the different approaches, while in AVEC'14 two tasks were selected: the Northwind (reading a novel passage) and Freeform (answering a series of questions-both neutral and potentially emotionally challenging) [210]. In the present work the subset of the two tasks, as employed in AVEC'2014, was used.

Three data partitions were provided by the challenge: a training, a development, and a test set, for a total of 300 recordings. Depression annotation however was provided only for 200 of the recordings; test set labels which were withheld for the challenge needs, were released later. Depression annotations of the video recordings were the participants' scores on the Beck Depression Inventory-II (BDI-II). Certain participants undertook the tasks on more than one time points, and had therefore more than one BDI-II scores. The following cutoffs are used to interpret individual BDI-II scores, standardized by Beck [33]:

- 0-13: minimal depression
- 14-18: mild depression
- 19-28: moderate depression
- 30-63: severe depression

Despite the AVEC'13 and AVEC'14 datasets being focused on continuous depression assessment, different approaches have utilized it to address classification of portrayed persons into high- and low-depression severity groups according to the standard BDI score cut-offs above [24] [182]. In the present several approaches have been made, such as categorical assessment of 2 classes (depressed vs non-depressed), 4 classes as the ones standardized by Beck, while in the main experiment continuous assessment was also attempted.

4.1.2 DAIC

During AVEC'16 [212] a different dataset was utilized, the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [93]. This dataset contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. These interviews were conducted by a virtual character, while collecting audio and video recordings as well as questionnaire responses. However the virtual agent was not functioning independently, but was controlled by a human interviewer in another room, as the purpose at the time was not to have a fully functional intelligent agent, but rather evaluate the efficiency of the interaction of the character with humans.

DAIC-WOZ provided only the features extracted with the OpenFace software [31], instead of the original video recordings, due to data protection issues. The shared data include 189 sessions, ranging between 7-33min (with an average of 16min). Transcript of the interaction, audio recordings, and OpenFace features are available for each recording.

4.2 Experiment 1

Results of the first complete experiment conducted in terms of the PhD work was published in IEEE International Conference on Signal and Image Processing Applications (Pampouchidou et al. 2015) [159]. The methodology involved the Tanaka face-parts detection, LCBP-TOP, kNN for classification, and k-fold for cross-validation. The parameterization for each of the algorithms is described next. The dataset used for experiments described in this subsection was the AVEC, focusing on the eye-pair region.

Curvelet transform was computed in each image for 8 orientations and 4 scales, resulting in a total of 41 pseudo-images such as the ones visualized in Fig.3.6. For the planes related to the time axis the curvelet transform was computed only for scale and orientation equal to 1. The same process was applied in every row (XZ) and column (YZ) of the pseudo image, resulting to $N + M$ number of images, where N the number of rows and M the number of columns in the pseudo image.

LBP was computed for 41×windowSize images for plane 1, plus the $N + M$ images of planes 2 and 3, for the parameter combination [8,1] and [16,2] (neighbourhood and radius respectively) [153]. The algorithm was tested for the following window sizes: {5, 15, 30, 60, 90, 120, 150, 180} for 30 fps videos. Length of the feature set varied according

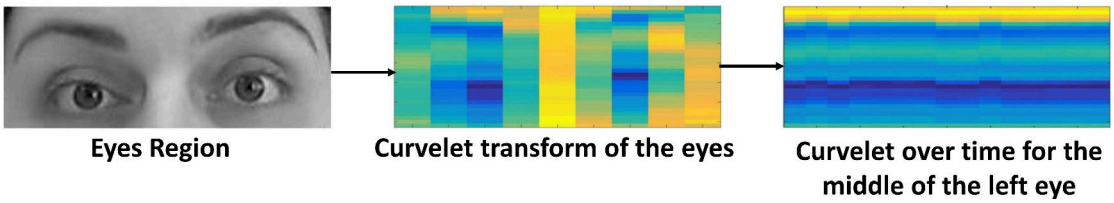


Figure 4.1: Experiment 1: LCBP-TOP on eye region, example for the column passing vertically from the middle of the left eye

to the length of the window; a window of 30 frames resulted on 1230 XY images (41 pseudo-images per frame), while for XZ plane the number of images was M and for YZ it was N; LBP ran for each one of them in two combinations of parameters as described above, resulting in 840 features for XY, 252 for XZ and 28 for YZ, for a total of 1120, while for the window of 180 frames the resulting feature set was of size 5320.

Table 4.1: Results of Experiment 1 for four levels of depression

Window Size (frames#)	Classification accuracy
5	41.36
15	48.52
30	**55.42**
60	47.83
90	49.31
120	48.66
150	46.32
180	50.33

Finally a Nearest Neighbor classifier was used to train/test the method adopted for detecting self-reported severity of depressive symptomatology according BDI-II scores. As shown in Table 4.1, best results were obtained by integrating facial activity data over for30 frames (1 sec). According to the confusion matrix computed for the 30 frame window in Table 4.2, more than 50% of the samples were correctly classified in each class, significantly exceeding random levels (25%) in the case of four classes.

Table 4.2: Experiment 1: Confusion matrix for best result (4 classes)

	Minimal	**Mild**	**Moderate**	**Severe**
Minimal	**51.0%**	19.6%	20.2%	9.2%
Mild	10.7%	**63.4%**	16.2%	9.7%
Moderate	11.2%	19.3%	**55.0%**	14.6%
Severe	11.9%	22.1%	10.5%	**55.5%**

4.3 Experiment 2

The work of this experiment was published in the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Pampouchidou et al. 2016) [160]. Again the AVEC dataset was used, while an effort was made to find solutions in crucial issues that emerged during the first experiment.

The first issue that needed to be addressed was face detection, as previously there was a high rate of false detections (˜ 20%). In order for any false classification to be attributed solely to the feature extraction method 100% accurate face detection had to be established; thus a semi-automatic face detection algorithm was implemented. Face region was manually initialized, and then tracked with the Kanade-Tomasi-Lucas tracker as described in [204], which is set to fail and to be reinitialized if the tracked points are below a threshold (20 points). Tracking fails when face goes out of the field of view, because of occlusions (e.g. hand in front of the face), or when illumination becomes too inadequate even for a human observer to distinguish facial features. Such issues are met in about 20 videos out of the total of 200, and these videos were disregarded.

The extracted facial region was resized to 256x256 pixels, followed by the Curvelet transform, with Orientation and Scale parameters were set to 1 [46], resulting to a 43x43 CurveFace. LBP descriptors were extracted for two sets of [Radius, Neighbourhood] [153]; LBP1=[1,8] and LBP2=[2,16]. LBP1 gave a 10-bin histogram, and LBP2 an 18-bin histogram; the two were concatenated to a 28 element feature vector for each frame in the *Frame-based Classification*.

Video-based Classification was performed next by computing the XZ and YZ planes, for a window of 30 subsequent frames, with an overlap of 15 frames. Therefore, a set of 30 CurveFaces of 43x43, results in 43 XZ planes of 30x43, and 43 YZ of 43x30. LBP1

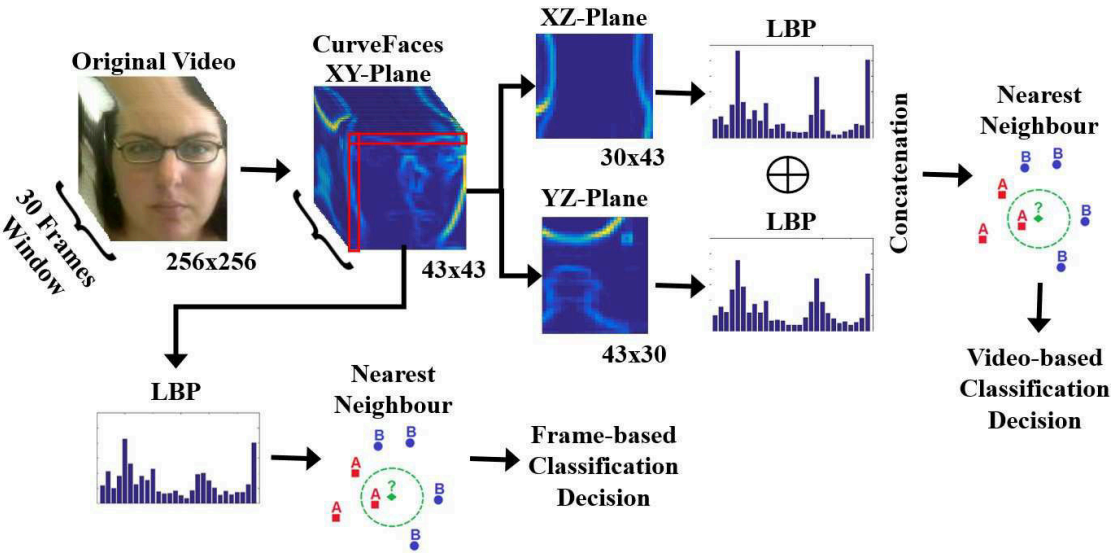


Figure 4.2: Experiment 2: Pipeline for the LCBP-POP

and LBP2 are again applied, to provide 43 pairs $\left(\bigcup_{i=1}^{43}(XZ_i \bigoplus YZ_i)\right)$ of LBP1$\oplus$ LBP2 descriptors. That is for every window 43 different feature vectors of 56 elements are being extracted, each of which is being treated as an individual sample for the classifier.

In an effort to make the problem binary, all three possible pairs of depression severity subgroups were initially considered: {minimal/mild}, {mild/moderate}, {moderate/severe}. However, the 4 subsets were highly unbalanced, with the 'minimal' class having as many recordings as all the rest together. Consequently random data-sampling was used in order to keep equal number of samples from each class.

Two main sets of tests were conducted, with either 20-Fold or LOO cross-validation methods. The 20-Fold classification was applied to individual samples, with the sets being partitioned 20 times. In this manner, 20 different randomly selected train/test sets were used in cross-validation. In the LOO method, all recordings belonging to the same participant were excluded from the training process, and were only used for testing. There were 58 different participants for 200 video recordings. For the LOO method classification of the videos was based on the class that was attributed to the majority of the samples. A Nearest Neighbour classifier was used. The proposed framework is illustrated in Fig.4.2. With a careful observation of the XZ and YZ planes, along with the sequence of CurveFaces, the motion patterns formed for the first row and column

Table 4.3: Experiment 2 results (%)

	Leave-One-Out		20-Fold	
	Frame	**Video**	**Frame**	**Video**
Minimal/Mild	60.5	**74.5**	**97.6**	83.8
Mild/Moderate	59.0	63.5	96.9	**85.4**
Moderate/Severe	**72.5**	**74.5**	95.8	81.3

Table 4.4: Experiment 2: Multi-class confusion matrix based on the Frame-based algorithm and 20-Fold cross validation (%)

	Minimal	**Mild**	**Moderate**	**Severe**
Minimal	**89.3**	3.6	3.7	3.4
Mild	1.9	**95.3**	1.4	1.4
Moderate	1.3	1.2	**96.1**	1.5
Severe	0.8	1	0.8	**97.4**

can be observed in both X and Y axes respectively.

Results for both *Frame-based* and *Video-based* classification algorithms, and all three cut-offs, for both cross validation methods are summarized in Table 4.3. Table 4.4 presents the confusion matrix across the four classes derived from the *Frame-based* algorithm (20-Fold cross-validation).

In sum, the technical advances achieved in Experiment 2 were: a) establishing an accurate face detection, b) implementing overlapping window to involve more information on the motion patterns, c) choosing the Pairwise-Orthogonal-Planes instead of the Three-Orthogonal-Planes in order to reduce the dimensionality of the features, d) testing video versus frame based representation, and e) applying additional cross-validation methods by using LOO in addition to k-fold.

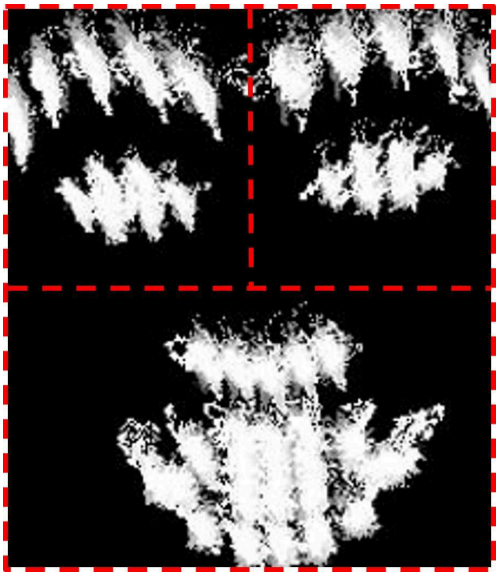


Figure 4.3: Experiment 3: LMHI

4.4 Experiment 3

This work was part of the participation to the 6th International Workshop on Audio/Visual Emotion Challenge (Depression Sub-Challenge), in terms of the ACM Multimedia Conference (Pampouchidou et al. 2016) [161]. The dataset employed was the DAIC-WOZ, and the participation was among the finalists of the challenge. Addressing the AVEC'2016 was even more challenging than the previous years, as video recordings were not available, and visual-based approaches were limited to using the provided landmarks.

Given the special circumstances the LMHI methods was applied. The landmarks used for LMHI, according to the numbering of Fig.3.2, were those corresponding to eyebrows {18-27}, eyes {37-48}, nose-tip {32-36}, and mouth {49-68}. Before computing the LMHI all landmarks were co-registered, using affine transformation, by aligning the points corresponding to the temples, chin, and inner and outer corners of the eyes {1, 9, 17, 37, 40, 43, 46}.

The gray value was defined by a step s, which corresponded to the maximum pixel value (255) divided by the total number of frames. Thus in every frame the gray value was computed by s multiplied by the frame count (i.e. for the 4th frame the gray value was 4 x s). The morphological operation of erosion with a structural element of disk and size 2 was applied in order to remove outliers (very distant movements) and the image

Table 4.5: Experiment 3 results: F1 score reflecting binary classification (high vs. low self-reported depressive symptomatology) in gender-independent and gender-based modes

Gender-independent	Gender-based
0.5 (0.9)	0.17 (0.83)

was cropped to the non-black pixels (non-zero). The resulting LMHI was further resized to fit the average size of the LMHI. An example of the resulting LMHI is illustrated in Fig.4.3, where the amount of movement is indexed by the multitude of bright pixels.

LBP, as well as HOG, were extracted based on LMHI. LBP was computed for two sets of parameters: radius and neighborhood for {1, 8} and {2, 16}, resulting in a total of 28 features covering the entire face. The LMHI was further partitioned using corresponding half ratios to represent three composite regions: "nose+mouth", "left eye+eyebrow", and "right eye+eyebrow". These regions were partitioned by red dashed-lines in Fig.4.3. For each subregion 28 additional features were computed. Further, HOG was computed for the entire face area with 1080 features, as well as for the remaining sub-regions of "nose+mouth", giving 360 features, "left eye+eyebrow" with 144, and "right eye+eyebrow" with 144 features as well. Additionally, the LMHI histogram bins were computed resulting in 255 additional features (black was excluded). Mean and standard deviation of the pixel values constitute the final two features, resulting in a total of 2097 LMHI features.

The performance of the proposed method was evaluated through training on the training set and subsequent testing with the development and test sets, as set by the challenge organizers. In addition, the algorithms were assessed using the LOO procedure on the joined training and development sets. Performance of each modality, in gender-independent and gender-based models, as well as the favorable comparison to the baseline performance, are reported in Tables 4.5 and 4.6.

4.5 Experiment 4

The fourth experiment was set in a different direction than the previous, as instead of appearance-based features, geometrical features with statistical descriptors were tested. The results of this experiment were published in the 39th Annual International Confer-

Table 4.6: Experiment 3: Comparison of the proposed method to the challenge baseline [212] F1-score: *depressed* (*not-depressed*)

Partition	Baseline	Proposed
Development	**0.58 (0.86)**	0.50 (0.90)
Test	**0.50 (0.90)**	0.18 (0.75)

Note: F1 score reflecting binary classification (high vs. low self-reported depressive symptomatology)

Table 4.7: Experiment 4: Experimental results(F1-Score)

Window	15	30	45	60
Gender independent	0.519	0.551	0.465	**0.586**
Gender based	0.577	0.519	0.546	0.580

ence of the IEEE Engineering in Medicine and Biology Society (Pampouchidou et al. 2017) [163].

The dataset employed was the AVEC'14, while several parameters of the algorithm were tweaked; facial landmark distances for example were computed based on euclidean and cityblock distance, and the time window used to extract the landmark motion features was systematically manipulated (set at 15, 30, 45, and 60 frames, which given the frame rate (30 fps) correspond to 0.5, 1, 1.5, and 2 sec). PCA was also tested for several sets by keeping 50, 60, 70, 80, 90, 100, 150, and 199 components. Further, two classification algorithms were tested, nearest neighbour and decision tree, using default parameters and leave one out cross validation. The parameters involved in the best-performing feature combinations were fixed to: euclidean distance, PCA components equal to 60 for each video and audio, and nearest neighbour classifier. Table 4.7 presents video classification for the different frame windows with optimal classification results for the gender independent mode at 60 frames window (2 seconds).

4.6 Experiment 5

Experiment 5 entailed a quantitative comparison of motion history image variants on the AVEC dataset for binary classification. Results were published in the EURASIP Journal on Image and Video Processing (Pampouchidou et al. 2017) [162]. The overall

Table 4.8: Experiment 5: Experimental results employing appearance-based descriptors (F1-score %)

	LBP{1,8}#	LBP{2,16}#	HOG	LPQ	Hist + Mean + Std	Feature Fusion
MHI	-*	-*	**81.9**	59.3	**81.9**	36.6
LMHI	-*	66.4	64.9	45.8	64.8	**72.7**
GMHI	-*	-*	**80.0**	69.8	**80**	74.0

LBP parameters in brackets correspond to {radius, neighborhood} respectively.
* Dash represents unavailable F1-score due to zero depressed individuals classified correctly.

pipeline of the methodology proposed in terms of this experiment is illustrated in Fig. 4.4.

OpenFace was employed for detection of 2D facial landmarks extracting aligned facial images of size 112×112 pixels [31]. In the present work, only successfully detected frames were retained for further processing.

Regarding specific parameters of the motion representation algorithms, the value of ξ was set to 25 for MHI, and to 8 for GMHI. These thresholds were chosen empirically, so that the static-background did not present movement in the motion image. This effect was noted at lower ξ values, where differences in pixel intensity were attributed to illumination variations. Setting the appropriate threshold, ensures that the movement represented by the motion images is meaningful, and can be attributed solely to movements. LBP was tested with two sets of [radius, neighborhood], namely [1,8] and [2,16]. The Gaussian kernel was chosen for the SVM classifier, with the expected proportion of outliers in the training data set to 10%. The $log(x/(1-x))$ transform function was applied. The number of principal components retained was also selected empirically: k=100 for the appearance-based descriptors and k=60 for VGG. LOSO was used for cross-validation, as a non-biased and person-independent method, given that the dataset contained more than one recordings of the same subject (ranging from 2 to 6).

Performance of the different configurations of the proposed algorithm is summarized in Tables 4.8 – 4.12. Table 4.8 presents performance of the various appearance-based descriptors for each of the three different motion images. The descriptors were tested individually, and combined with feature level fusion (concatenated). Table 4.9 presents

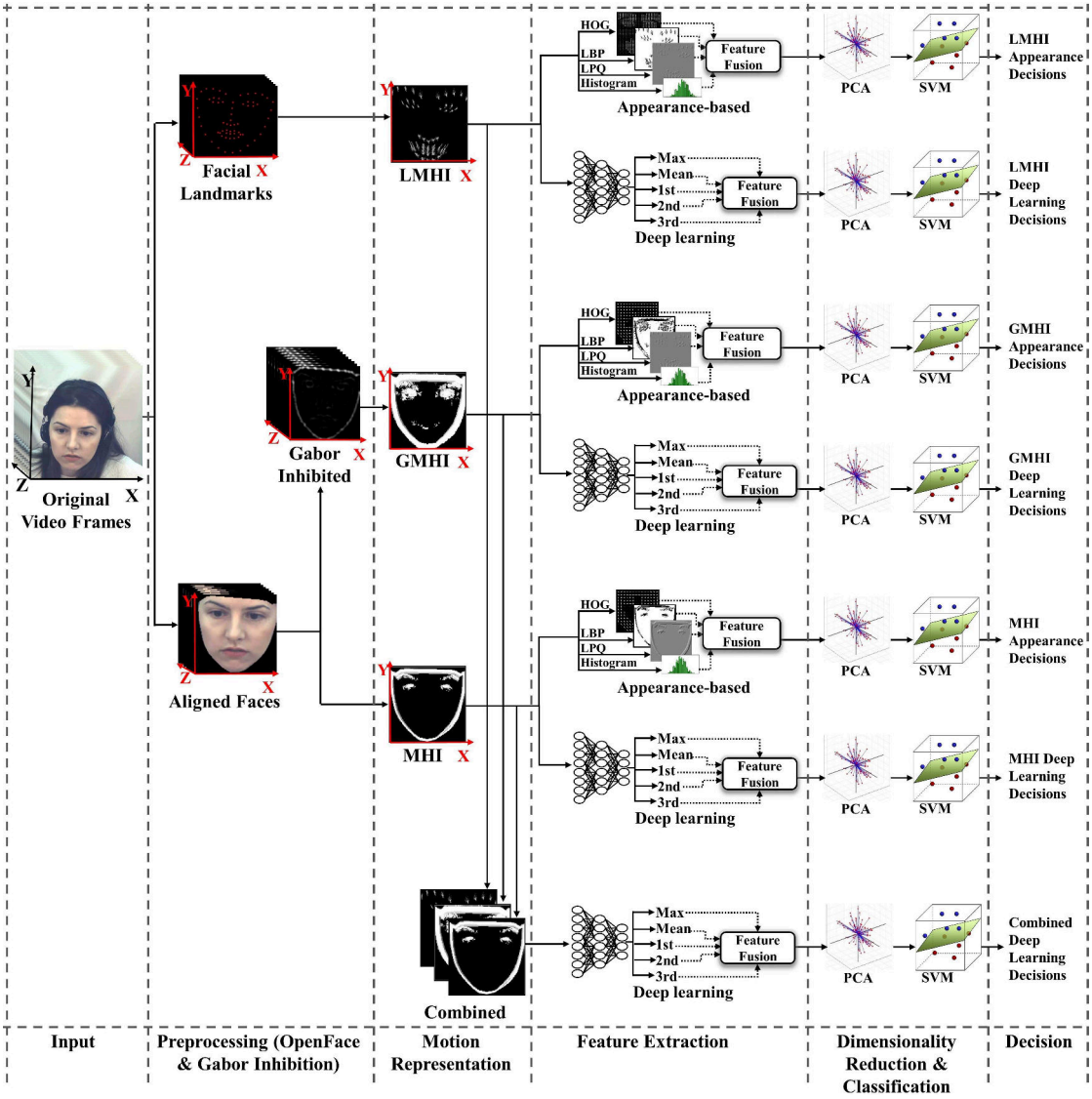


Figure 4.4: Experiment 5: Flow of the motion history variants methodology. Dashed arrows in feature extraction indicate that features are considered individually or in combination with feature fusion. MHI: Motion History Image, LMHI: Landmark Motion History Image, GMHI: Gabor Motion History Image, HOG: Histogram of Oriented Gradients, LBP: Local Binary Patterns, LPQ: Local Phase Quantization, PCA: Principal Components Analysis, SVM: Support Vector Machine

Table 4.9: Experiment 5: Experimental results employing VGG features, before and after fully connected layer (F1-score %)

	Before		After			Feature
	Max	**Mean**	**FC6**	**FC7**	**FC8**	**Fusion**
MHI	87.1	63.0	51.5	84.8	70.1	**87.4**
LMHI	64.0	**67.7**	56.6	18.2	66.4	64.0
GMHI	**85.7**	62.7	55.6	76.0	65.5	84.3
Combined	64.6	52.1	51.8	**74.7**	46.3	65.1

Table 4.10: Experiment 5: Confusion matrix for the best F1-Score of the proposed approach (VGG-MHI feature fusion/2 classes)

Self-Reported \ **Predicted**	**Non-depressed**	**Depressed**
Non-depressed	**102**	2
Depressed	20	**76**

the performance of VGG for the different configurations. The confusion matrix corresponding to the best-performing model (VGG feature fusion) is presented in Table 4.10.

Additional performance metrics for the best performing model are reported in Table 4.11, whereas Table 4.12 compares the present findings to previously published results using similar data sets. It should be noted that the results presented by Senoussaoui et al. [182] and Alghowinem et al. [24] were obtained using different organization of the

Table 4.11: Experiment 5: Additional performance metrics for the best-performing approach (VGG-MHI feature fusion / %)

F1	87.4
Accuracy	89.0
Sensitivity	79.1
Specificity	98.1
Precision	97.4
Cohen's Kappa	77.8
.95 Confidence Interval	± 8.6

Table 4.12: Experiment 5: Performance comparison with the literature (%)

Approach	Accuracy	Average Recall	F1
Sennoussaoui et al. [182]	82.0	-	-
Alghowinem et al. [24]	-	81.3	-
Pampouchidou et al. [160]	74.5	-	-
Pampouchidou et al. [163]	-	-	58.6
Proposed (VGG feature fusion)	**89.0**	**88.6**	**87.4**

AVEC datasets, thus a direct comparison with the present results is not possible. In the cross-corpus approach of Alghowinem et al. [24] the used set was carefully selected from the original dataset (AVEC'13) in terms of the total number and duration of recordings per participant, in order to match the other two datasets. On the other hand, in [182] the algorithm was applied to the training data set provided by the challenge organizers (AVEC'14) and tested on the development dataset. Although the AVEC dataset has been widely employed in approaches for continuous depression assessment, the aforementioned approaches, to the best of the authors' knowledge, are the only ones attempting categorical depression assessment on the specific dataset. Results are reported for F1-score, unless indicated otherwise, which is given by equations 2.5, 2.3, and 2.4

4.7 Summary

Classification results using previously available data sets are listed in Table 4.13 for each of the five preliminary experiments. In total 4 conference papers were produced, three published by IEEE, while the one by ACM was in terms of the AVEC'16 Depression Sub-Challenge, with the participation being included in the finalists of the challenge. Finally, one article was produced as well, in Springer EURASIP Image and Video Processing.

Table 4.13: Summary of achievements during preliminary experiments

Publication	Type	Name	Dataset	No Classes	Acc	Kappa
Pampouchidou et al. [161] (2016)	Intern. Conf.	ACM AVEC	DAIC-WOZ	2	F1=0.5	39.00%
Pampouchidou et al. [159] (2015)	Intern. Conf.	IEEE ICSIPA	AVEC	4	55.42%	41.61%
Pampouchidou et al. [160] (2016)	Intern. Conf.	IEEE EMBC	AVEC	2	74.50%	13.00%
Pampouchidou et al. [163] (2017)	Intern. Conf.	IEEE EMBC	AVEC	2	F1=58.6	20.84%
Pampouchidou et al. [162] (2017)	Article	Springer EURASIP	AVEC	2	89%	**78.00%**

Chapter 5

Main Experiment

The main experiment in terms of the PhD is consisted by two main parts: a) the data collection from the clinical study, and b) the experimental evaluation of the developed methodology on the clinically valid dataset. The respective sections follow below based on this rationale. The work described in this Chapter has been submitted for publication the the IEEE Journal of Biomedical and Health Informatics.

5.1 Data Collection

The work described in this section concerns the creation of a database, comprising recordings of human facial expressions, speech, and physiological signals, after experimental induction of discrete emotions that are considered relevant for assessing depressive symptomatology. The collected data are used in Section 5.2 for the main study of developing computational methods capable of recognizing non-verbal signs, able to differentiate mentally healthy individuals from those suffering from depression.

Although the primary aim of the dissertation did not involve construction of an emotion database, the undertaking of original data collection lies in the field of affective computing and faces similar challenges. Data collection procedures required careful design having a strong impact on the capacity to address the main research questions [122]. As already shown in Ch.2, the methods employed for collecting relevant symptomatology from individuals, along with the annotations, are the most important steps. Table 5.1 summarizes all steps that took place in order to successfully complete the data

collection performed in terms of the presented thesis, while the details of the different steps are described in some detail within the following sections.

5.1.1 Ethics and Data Protection

Obtaining the required bioethics permissions was the first and prerequisite step in order to begin the data collection. Privacy and ethics are fundamental principles when it comes to data collection involving video recordings, or other types of sensitive data [57]. Obtaining ethical approval, informed consent, and establishing the confidentiality of the data are some key aspects. Securing approvals from the Bioethics Committee of the University Hospital of Heraklion (Decision 296, Session 7/06-04-2016) and the National Data Protection Authority (Protocol No ΓΝ/ΕΞ/392-2/21-04-2016) lasted approximately one year. In addition to providing prospective participants with an extensive description of the study details in layman terms, they were explicitly informed that they could withdraw at any point during or after completing the experiment and request deletion of all their data.

5.1.2 Participants

Participants were healthy volunteers aged 20-65 years without history of mental or neurological disorder and patients suffering from MDD as diagnosed by their treating psychiatrists at the Psychiatry Outpatient Clinic, University Hospital of Heraklion. Healthy volunteers were recruited through announcements on a Facebook page, flyers posted at several sites (patient waiting areas at the University Hospital, University Campus, FORTH labs and common areas), as well as through personal referrals. Patients were informed regarding the study by their physicians during regular appointments and if they consented verbally they were explained the details of the study both verbally and in writing by a research assistant (psychologist). In most cases written consent and testing were obtained at the same time to avoid bring the patient in on a separate day.

The final sample included 65 participants (control group n=45, patient group n=20). Socio-demographic information and clinical indices can be found in Table 5.2 for each group, while individual data on the same variables can be found in Appendix A. Although an effort was made to keep demographics in similar levels, and recruit participants in the same age range the patient group was older and had completed fewer years of formal

Table 5.1: Summary of work toward data collection

1. Obtaining necessary research permissions from the
   - Bioethics Committee, University of Crete Hospital
   - National Data Protection Authority
2. Protocol design
   - Building web-based application to administer self-report questionnaires and scales
   - Building the emotion perception application
   - Selecting stimulus video-clips
     - **o** Initial selection of video-clips for pilot testing
     - **o** Post-processing of the video-clips
       - ***** Video-editing of clips to appropriate length
       - ***** Adding Greek subtitles
       - ***** Adding white screen at the beginning of each clip to serve as baseline
     - **o** Pilot testing the video-clips (n=10)
     - **o** Statistical Analysis of emotion self-ratings from pilot study
     - **o** Final selection of video-clips based on pilot study results
3. Technical setup
   - Selection and purchasing of suitable, portable biosignal recording device
   - Setting up the experiment room (camera, lights, biosignal device, PC controlling stimuli and questionnaire/scale administration and PC performing data acquisition and storage)
4. Participant recruitment
   - Control group: through dedicated project webpage, email list, flyers
   - Patient group: through referrals by clinicians in University Hospital, Psychiatric Outpatient Clinic
5. Recordings
   - Technical part: handling recordings
   - Psychologist: guiding the participants through the protocol, and conducting interviews
6. Post-Processing
   - Data Compression
   - Selection of informative parts
   - Condition- and diagnosis-blind annotation of facial videos by experts (psychologists)

Table 5.2: Socio-demographics and clinical data for both groups, control (n=45) and patients (n=20)

	Control	**Patients**	**P-value**
Age [M(SD) in years]	39.96 (7.87)	49.7 (12.68)	<.001
Age range in years	24-56	24-70	–
Men	17 (37.8%)	3 (15.0%)	>.05
Education [M(SD) in years]	16.69 (5.04)	10.1 (4.77)	<.001
BDI-score [M(SD)]	6.49 (5.62)	21.8 (14.39)	<.001
STAI2-score [M(SD)]	40.27 (9.27)	52.95 (10.13)	<.001
Blinded expert judgment[1] **[M(SD)]**	2.43 (1.25)	4.85 (1.08)	<.001

[1]Conducted by psychologists based on participants' face videos alone

education than the control group. The two groups did not differ significantly on the percentage of women which was higher than the percentage of men in both groups and especially among patients (85 vs. 15%) in accordance with the literature [40]. As expected, self-reported depression and anxiety were higher for the patients (BDI-II, STAI, and annotation). Three participants did not complete one task (each for different reasons and a different task), bringing the total number of recordings to 322 (out of a possible of 325: 65 participants × 5 tasks).

5.1.3 Psychological measurements and experimental procedures

In designing the study special care was given to two key aspects:

(a) Tools employed to assess depression-related symptoms in everyday life of participants (including those not meeting formal criteria for MDD or other mood disorder), and

(b) Defining the experimental conditions to elicit specific emotions in the laboratory under the assumption that the quality and intensity of such emotions and facial expressions of emotions will be altered in the presence of significant depressive symptomatology [64]. We adopted techniques that involved "human-human interaction" as well as "human-computer interaction" [122] in both "social" and "nonsocial" context as described in Subsections of 5.1.3.2 and summarized in Table 5.3.

5.1.3.1 Tools assessing depression-related symptoms

Two self-report questionnaires were administered to all participants to assess depression and anxiety symptomatology, namely the Greek adaptations of the BDI-II and the STAI Form Y. The BDI-II [85] has been adjusted according to DSM-IV criteria for MDD, comprising 21 questions, scored on 0-3, 0-4, or 0-5 point scales. These questions assess a wide range of emotional and behavioral signs of depression such as body image, hypochondriasis, difficulty in working, sleep loss, appetite loss, thoughts of self-punishment, suicidal ideation, and reduced libido. The higher the score, the more severe the depressive symptoms, while the standardized cutoffs for BDI-II are:

- 0-13: minimal depression
- 14-18: mild depression
- 19-28: moderate depression
- 30-63: severe depression

The Trait-Anxiety Scale (STAI Form Y-1) [81] was designed as a self-assessment tool of persistent symptoms (feelings, somatic complaints and behaviors) that are considered as core manifestations of anxiety as a characteristic of the individual. It consists of 20 items rated on a 1-4 point scale, with higher total scores indicating higher levels of anxiety.

5.1.3.2 Emotion elicitation paradigm

5.1.3.2.1 Non-social context: Selection of stimuli

In the context of a pilot study involving n=10 healthy volunteers aged 26-to-39 (5 men) we evaluated thirteen video clips that were considered as suitable for eliciting each one of four of Ekman's six basic emotions [joy (4 clips), disgust (3 clips), sadness (3 clips), and fear (3 clips) [74]). Clips were chosen from popular movies, TV series, or YouTube. Participants in the pilot study were asked to view each video clip (presented in a different random order across participants). Participants in the pilot study were asked to view each video clip (presented in a different random order across participants). After watching each clip they were asked to rate their emotional experience on 15 dimensions (according to Gross and Levenson [95]) on a 0-8 point scale. In Fig. 5.1.a the three

video clips that were initially selected to induce sadness are contrasted based on group-average ratings on Gross and Levenson's 15 emotion dimensions. Whereas sadness is the dominant emotion elicited by all three clips, the clip from the movie "Forrest Gump" was excluded as it ranked lower in the sadness dimension while also elicited higher levels in pleasure and interest. After having collected recordings from about half of the control population (n=24), the stimuli were again evaluated, to make another selection, and keep only one for the patients population. As it can be seen in Fig. 5.1.b again both clips rank high in terms of sadness, yet the clip from the movie "Terms of Endearment" was chosen because it elicited lower levels of two potentially confounding emotional states, namely arousal and tension. These states are considered as indices of psychological stress that can alter the intended facial expression profiles for sadness.

Regarding the joy clip, a scene from a popular Greek comedy series (Para pente) was chosen over a clip of a laughing baby from YouTube, a greek movie (Mpakalogatos), and another comedy series (Peninta-Peninta), using a similar rationale as for the sadness clip. Finally, 30 frames (1 sec) were added to the beginning of each clip to serve as baseline. The remaining video clips (disgust, fear) did not elicit distinct emotions (e.g., the disgust clip elicited high levels of disgust but also similar levels of repulsion, arousal, interest, tension / c.f. Fig. 5.2) Therefore these clips were not used in the main study.

5.1.3.2.2 Social context

In addition to the video clips, an interpersonal method for eliciting positive and negative emotions was employed as well in the form of a semi-structured interview with the research assistant. Initially (prior to viewing the Joy clip) the participant was asked to describe a positive personal experience from his/her life; if the participant hesitated he/she was encouraged by the research assistant to go into more details, and asked probing questions guiding him/her to relive this experience as vividly as possible. In a similar manner, prior to viewing the Sadness clip participants were asked to describe a negative personal experience, which involved sadness or distress. The research assistant was closely monitoring the participant's emotional state and was instructed to terminate the interview in case of extreme emotional responses. A neutral baseline for the positive/negative experience description was chosen in the form of reading aloud a 260-word narrative text describing an excursion in the country.

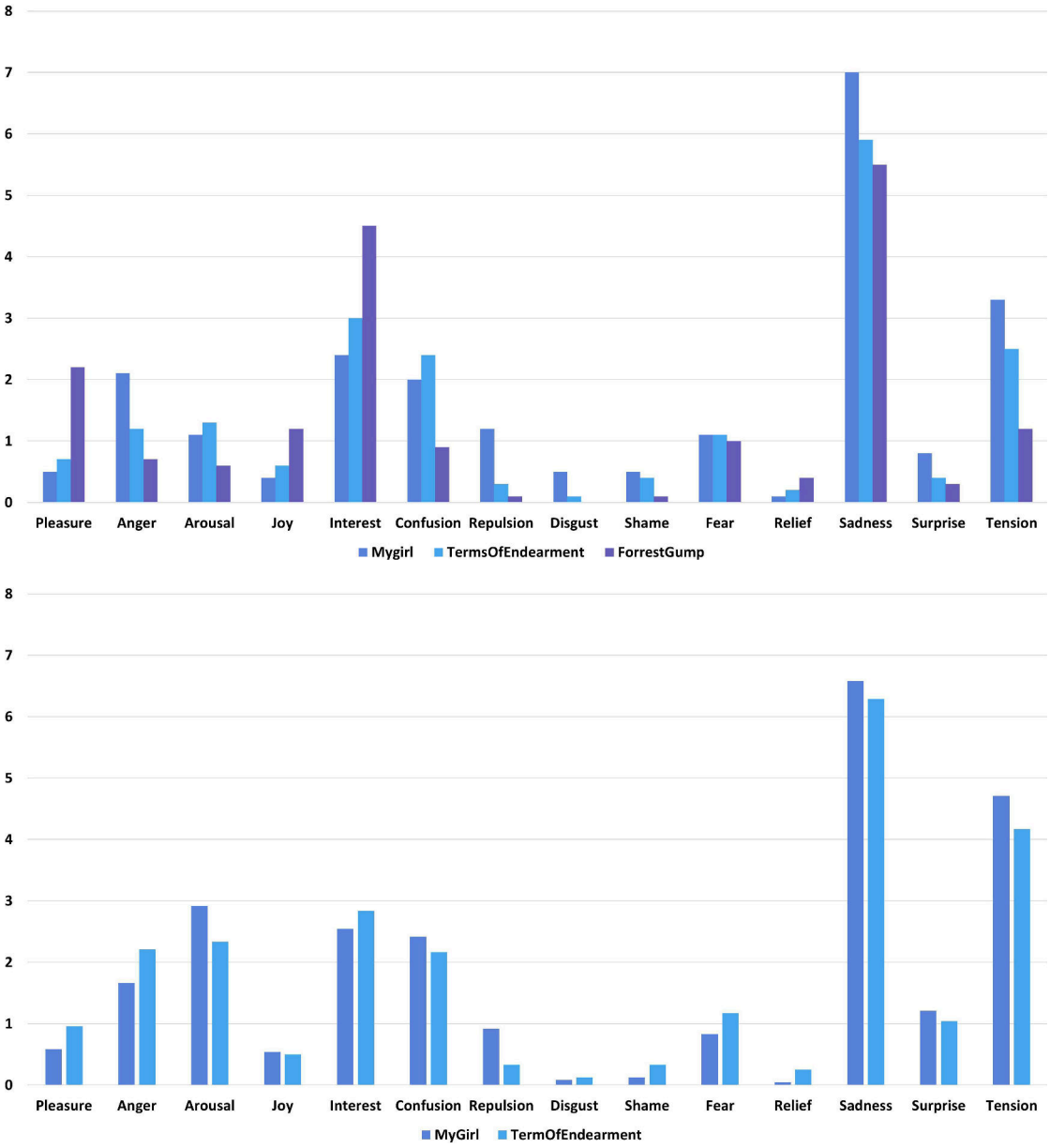


Figure 5.1: Average participant ratings of Sad video clips on the 9-point emotion scale of Gross and Levenson [95]. Upper panel: Ratings of the three video clips tested in the pilot study; Lower panel: Corresponding ratings by control group participants in the main study (n=24)

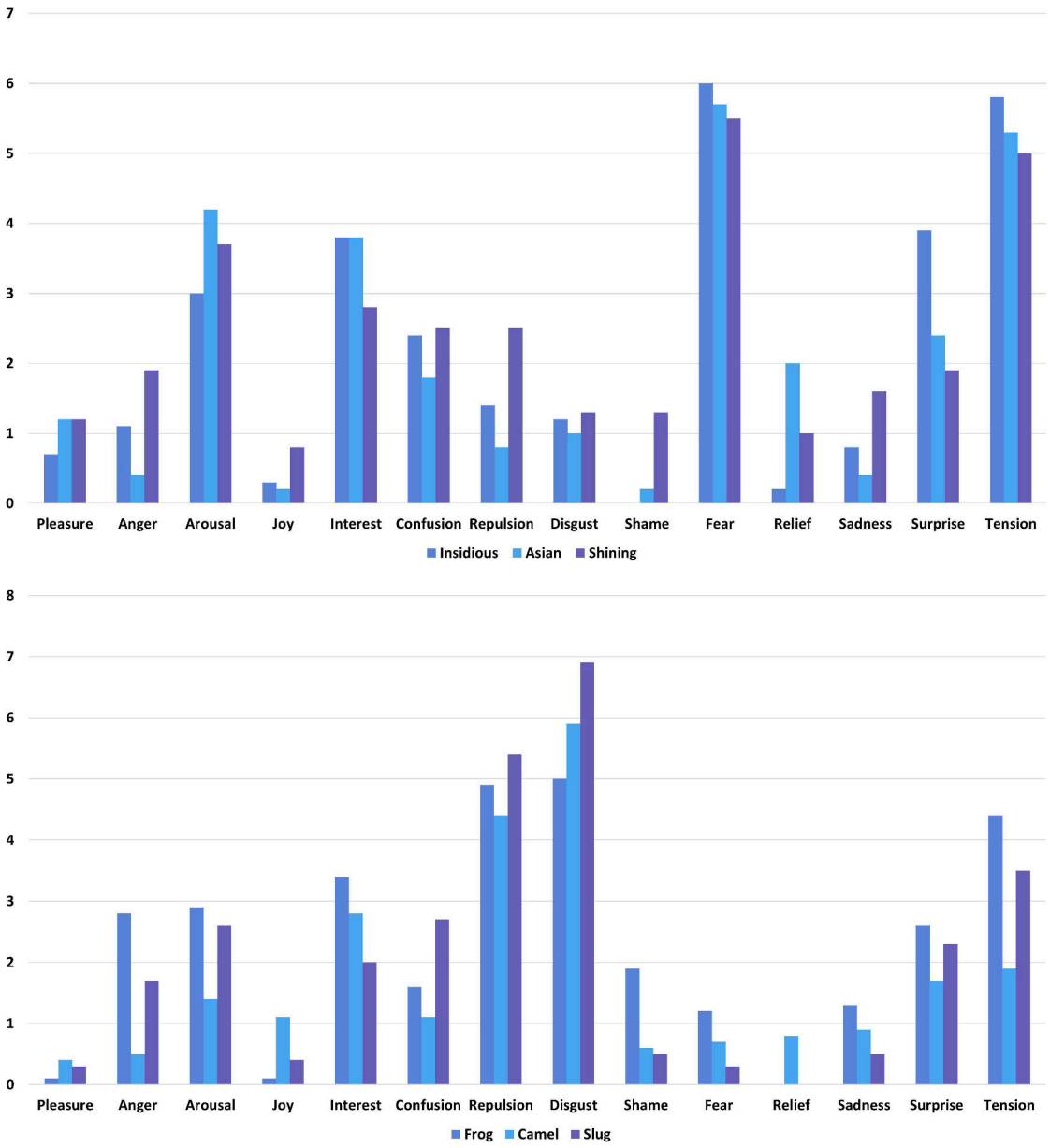


Figure 5.2: Average pilot study participant ratings of Fear (upper panel) and Disgust video clips (lower panel) that were not selected for the main study, as they failed to elicit distinct emotions

On two predetermined occasions, the participant was guided through relaxation exercises, involving controlled breathing and brief mindfulness techniques to ensure that emotional states and stress levels returned to neutral prior to the description of positive experience/joy clip and, again, prior to the description of negative experience/sadness clip.

5.1.4 Technical Setup

The technical setup involved a camera, lights, microphone, and biosignal recording device. As shown in Fig. 5.3, the participant was seated in front of the PC which presented the stimuli, rating scales and questionnaires and registered his/her responses. Two researchers conducted the experiment: one operated the stimulus delivery and recording devices (AP) and the second (psychologist) interacted with the participant (obtained consent, provided instructions and additional explanations if needed, performed guided relaxation, and conducted the semi-structured interviews). Additional biosignals (voice, heart rate, galvanic skin response) were also recorded but will not be considered further in this thesis. Facial video data analyzed for the purposes of this dissertation originated from conditions 7-8 and 11-13 (shown in bold in Table 5.3). The order of conditions is presented in Table 5.3 and the total duration of the experiment ranged between 60-90 min.

The camera employed for the video recordings was the Grasshopper®3, which provides high-performance and high-quality imaging (see Table 5.4). Chosen camera was of high specification in order to investigate how the quality of the recordings affects the outcome, as comparing to the reported datasets (see Table 2.2 which are of lower standards. The camera specs support maximum specifications of 90 frames per second (fps) and 2048×2048 image resolution. Although the PC involved was high performing (32GB RAM, SSD, SATA3), still full specifications of the camera could not be supported. During 10 benchmark tests it was specified that the average speed of writing to the disk was 413.9 MB/s. Furthermore, several configurations of fps and resolution were tested, to check whether frames were being lost during acquisition due to bottle-neck problem; 80 fps and 1920×1920 was the maximum configuration without losing any frames for 10 additional tests, therefore there were the finalized settings for the camera. Indirect lights were used to establish controlled illumination; the lights did not point directly on the participants, but to the wall opposite of them, to avoid extreme brightness.

Table 5.3: Data collection protocol in the main study

#	Task	Method
1	Acquaintance	Orally
2	Read and sign information sheet and consent form	In writing
3	Demographics & Clinical History	Webpage
4	Relaxation (breathing exercise)	BioTrace
5	Prolonged **/a/** utterance	Orally
6	Positioning in front of the camera (aligned with a spot on the wall)	–
7	**Description of positive experience**	Orally
8	**Joy clip**	Webpage
9	Emotion ratings for Joy clip	Webpage
10	Relaxation (breathing exercise)	BioTrace
11	**Reading aloud a neutral text**	Webpage / Orally
12	**Description of negative experience**	Orally
13	**Sadness clip**	Webpage
14	Emotion ratings for Sad-clip	Webpage
15	Complete STAI	Webpage
16	Complete BDI-II	Webpage
17	Prolonged **/a/** utterance	Orally

Table 5.4: Camera Specifications

Resolution	2048×2048	**Frame Rate**	90 FPS
Megapixels	4.1 MP	Chroma	**Color**
Sensor Name	CMOSIS CMV4000-3E5	**Sensor Type**	CMOS
Readout Method	Global shutter	**Sensor Format**	1$$
Pixel Size	5.5 μm	**Lens Mount**	C-mount

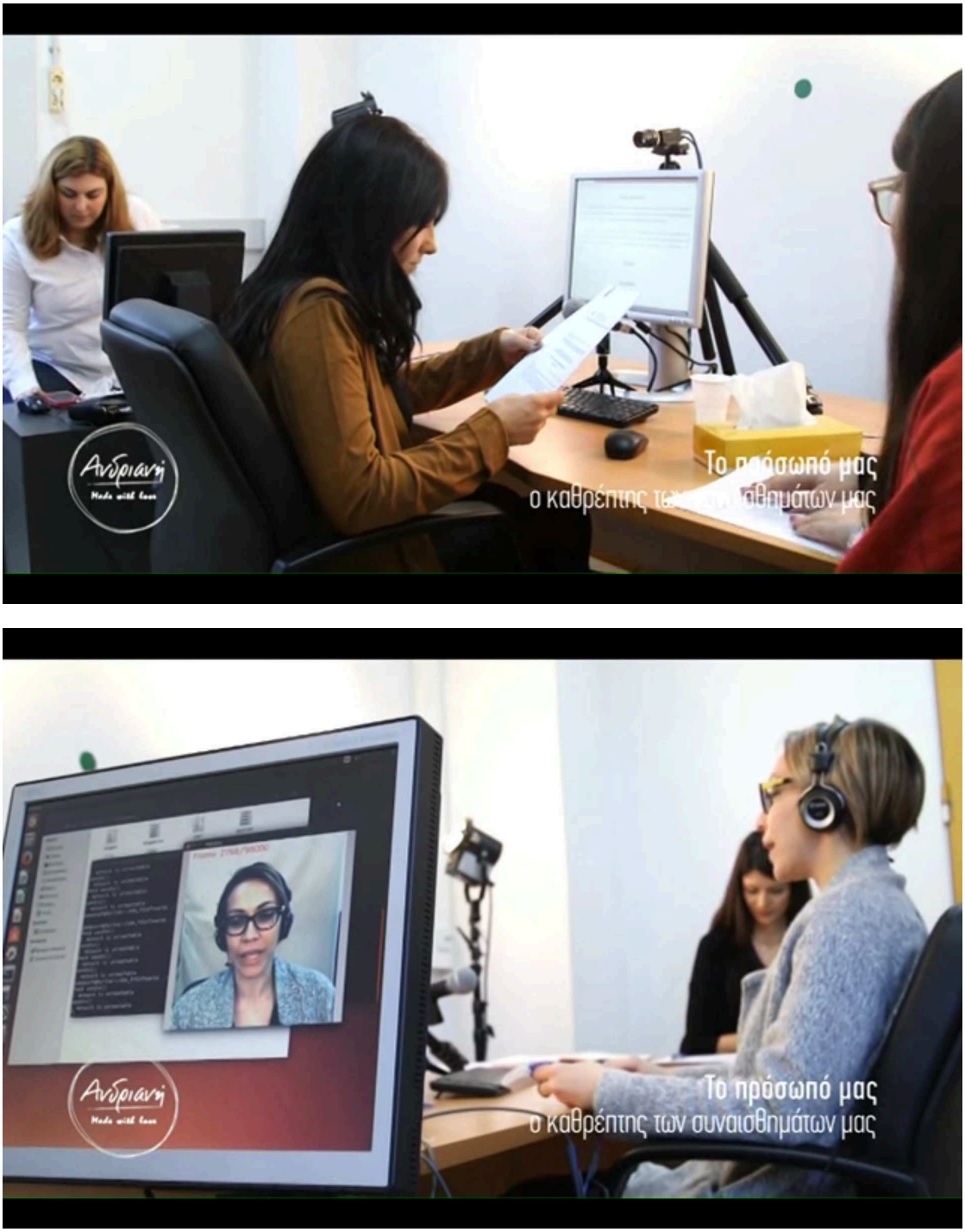


Figure 5.3: Data collection experimental setup

5.1.5 Blinded Expert Annotation

Although based on the data collection protocol, 3 types of labels were already available (diagnosis, BDI-II, STAI), still it was interesting to investigate the annotation by experts. As also supported by [56] *"It makes more sense to construct a labeling that is rich enough to support any foreseeable use of the basic material."*, who continues supporting that a parameter to be examined is redundancy. In this case however, given the nature of the rest of the labels against the annotation, it can be supported that there is no redundant information, as they entail entirely different aspects for depression manifestations. More importantly, the motivation to move forward with the annotation was to have a label directly comparable to the output of a video processing algorithm, as both diagnosis and self-reports involve much more information (verbal, full-body, repeated sessions, reporting bias, etc.) which cannot be derived solely by a single-visual video recording. Thus, the aim would be to compare the ability of an expert to detect visual signs of depression from the entire set of facial video recordings, versus the proposed automatic methodology based on digital image processing and machine learning. It is worth noting that it is the first time that experts' annotations are used for automatic depression assessment, as to the best of the author's knowledge all relevant approaches to date employ either the diagnosis, self-reports, or HAM-D scores attributed by clinicians in the time of examination.

CARMA[1] [86], a user-friendly software, was employed for this procedure; among the annotation tools reviewed in [122] CARMA was the only one which was fully-functional and corresponded to the needs of our research. Within CARMA the video is displayed in the center (only visual information in our case), while a slider right next to the display can be adjusted real-time and in a continuous manner by the annotator, stored externally in a file of comma separated values for further statistical analysis. The rating scale was set to 0-8 points with 0 indicating complete absence of depressive signs, and 8 indicating the most severe signs of depression. Annotation was performed by two experts (psychologists with at least one year of experience with patient interviews at the University Psychiatry Clinic) who were blinded with respect to clinical diagnosis and emotion elicitation condition. In case of deviation between raters equal to or greater

[1] https://github.com/jmgirard/CARMA/

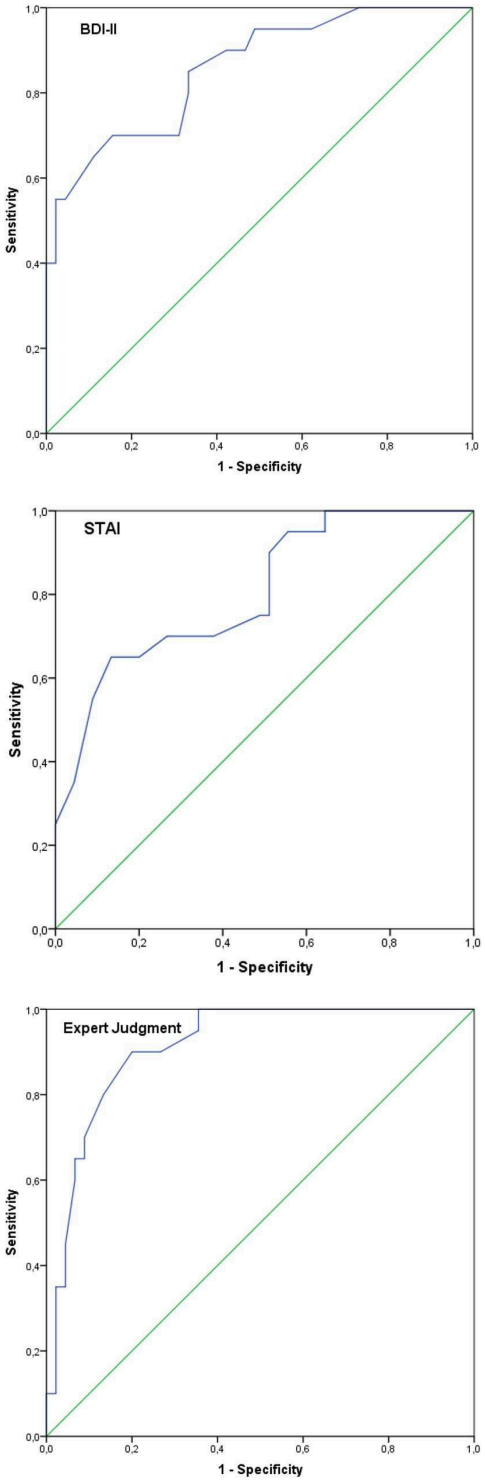
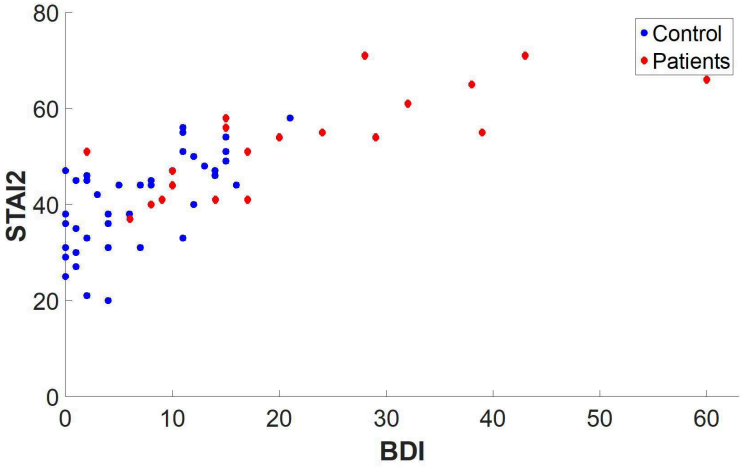
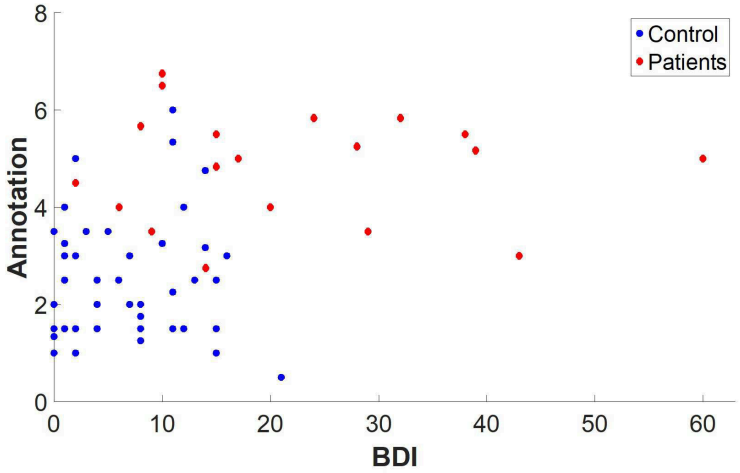


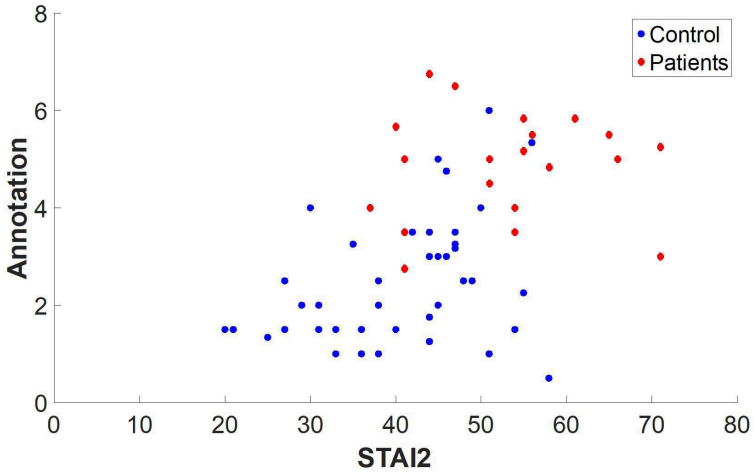
Figure 5.4: Dispersion of BDI-II, STAI, and expert judgment values per group. Median (dark horizontal line), interquartile range (boxes), and range of values on BDI-II (upper panel), STAI (middle panel), and blinded expert judgment (lower panel) per participant group. The 13/14 point cutoff on BDI-II, 39/40 point cutoff on STAI, and 3.4/3.5 point on expert judgment are indicated by a blue horizontal line



(a) BDI/STAI2 rho:0.768



(b) BDI/Annotation rho:0.424



(c) STAI/Annotation rho:0.527

Figure 5.5: Bivariate scatter plots of STAI, BDI-II and expert annotations across participant groups.

than 1.5 points on the scale, a third annotator was employed. The final annotation score registered was the average of all available annotations for each elicitation condition.

5.1.6 Statistical Results

The dispersion of BDI-II, STAI, and expert judgment values per group is shown in Fig. 5.4; it shows that the best separation between groups was achieved by expert judgment ratings of depression relying solely on facial expressions. This impression was proven by sensitivity (Receiver Operating Characteristic) analyses revealing higher Area Under the Curve for expert judgment (AUC=0.913, SE=0.035, p<0.001; Fig. 5.4 lower panel) as compared to BDI-II (AUC=0.857, SE=0.051, p< 0.001; Fig. 5.4 upper panel) and STAI (AUC=0.804, SE=0.059, p< 0.001; Fig 5.4 middle panel). These analyses confirmed that the optimal cutoff reported in the Greek validation studies [85] [81] for BDI-II (13/14 points) was applicable in the current data set as indicated by 70% sensitivity and 85% specificity for BDI-II. Although sensitivity associated with the standard cutoff of 39/40 points on STAI was very high (90%), specificity was quite low (55%) which is expected given that the standard cutoff was establish for identifying persons with anxiety disorders [85] [81]. The optimal cut-off value for expert judgment (i.e., the value associated with highest sensitivity (90%) and specificity (80%)) was determined at 3.4/3.5 points (on the 0-8 point scale). Bivariate scatter plots shown in Figure 5.5 further suggest that self-reported values of depressive and anxiety symptoms were only moderately correlated with expert judgments of depression, whereas the correlation between the former (STAI and BDI-II scores) was very high (r = .768).

5.2 Experimental Tests

After the data collection was completed, the question of evaluating the proposed methodology arose. Based on the preliminary evaluation of the different algorithms and experimental setups, the MHI combined with VGG features proved to be the best performing, thus there is a great motivation to proceed with this methodology for the final and main experimental test. Next, the particulars of experimental setup are being explained, followed by two set of tests, those of a) categorical assessment, and b) continuous assessment.

5.2.1 Experimental Setup

The video recordings were successfully preprocessed by OpenFace to derive aligned faces, which were then submitted to the MHI algorithm (c.f. subsection 3.2.2.1) to derive one motion image for each of the 322 recordings available.

The motion images were then processed through different feature extraction methods (i.e., HOG, LBP, VGG16, and VGG19, see subsection 3.3). Both HOG and LBP were applied with the default MATLAB parameters and fully connected layers were employed for the VGG networks (FC7 and FC8). PCA was used next for dimensionality reduction, for different values of coefficients (i.e., 5, 10, 20, 40, 45, 50). The recordings were considered in different setups a) across all tasks, and b) for each task (individual stimulus) separately, for both gender-based and gender-independent modes. Finally SVM was employed for both categorical (classification) and continuous (regression) approaches.

Given that video recordings were of high specifications, the processing had high requirements too. However the first steps (preprocessing and feature extraction) were performed offline on high-performing PC's (i.e. Intel (R) Core (TM) i7-4720HQ CPU @ 2.60 GHz, 8 GB RAM, 64-bit Windows 10 Home (C) Microsoft Corporation, 512 GB SSD). The cross validation pipeline, employing LOSO, for PCA and SVM was executed on a remote Virtual Machine (VM) with 8 vCPU (virtual CPUs), 256 GB RAM, and 100 GB hard disk, running Ubuntu 16.04 LTS. The VM was running on a physical server of quite high overall specifications (i.e. 512 GB RAM, Intel Xeon E5-2690 v3 2.6GHz 12 cores / 24 threads), and at the time it was not especially loaded with other VMs while running the described tests.

5.2.2 Categorical Assessment

Categorical analyses were performed for four different types of binary labels: clinical diagnosis (depression, healthy), BDI-II score, STAI score, and blinded experts judgment (using cut-offs listed in 5.1.6). Analyses were performed separately across genders as well as for each gender and task separately. Results (Cohen's Kappa values) of different approaches are presented in Fig. 5.6, whereas Table 5.5 lists the best-performing approach for each classification scheme (diagnosis, STAI, BDI-II, expert judgment) in gender-based and gender-independent modes.

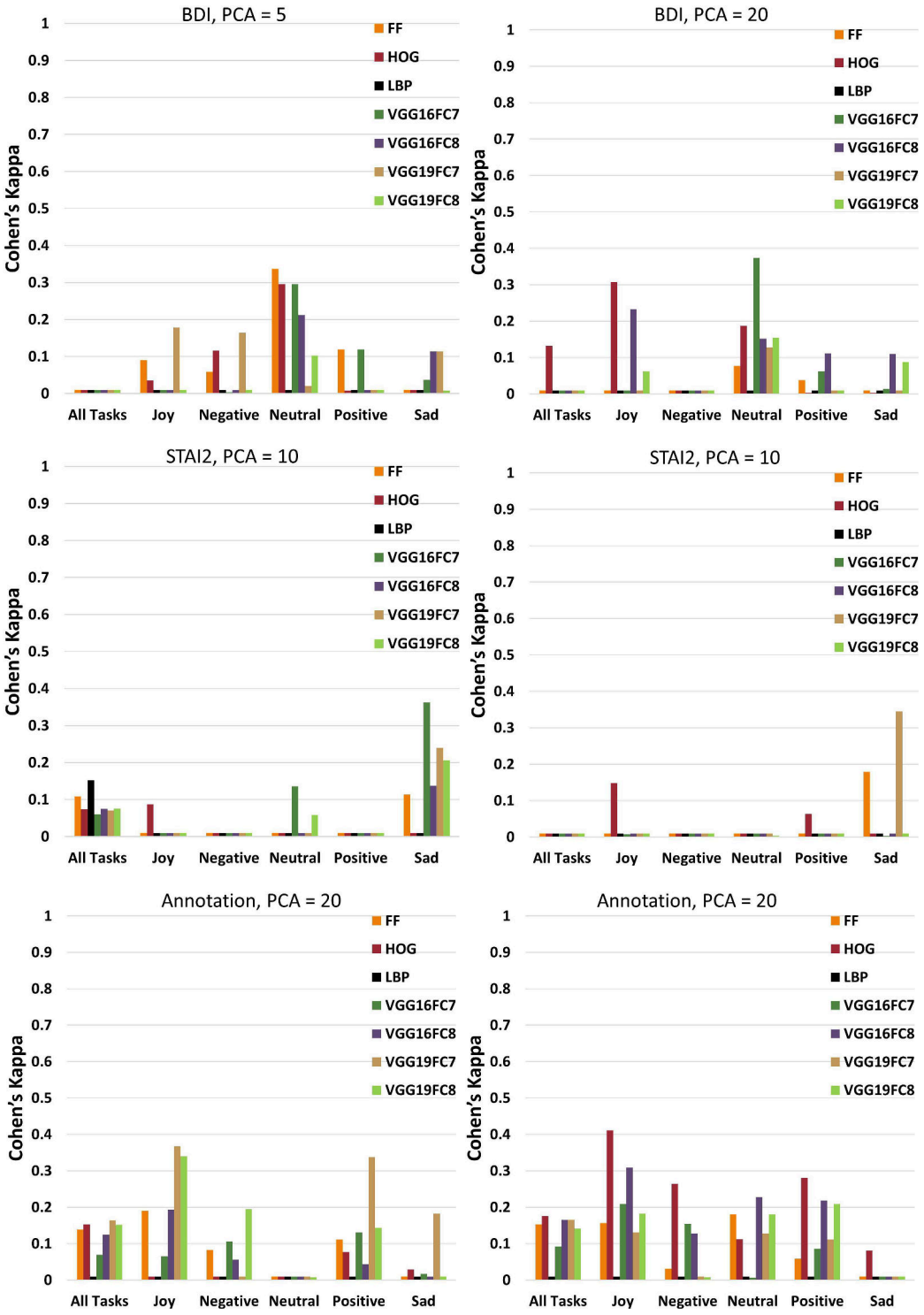


Figure 5.6: Average performance (Cohen's Kappa) of the categorical assessment schemes by number of PCA components retained during preprocessing for each experimental condition. Results for gender-dependent and gender-independent schemes are displayed in the left- and right-hand columns, respectively. Colors represent different video-based features.

Continue on the next page

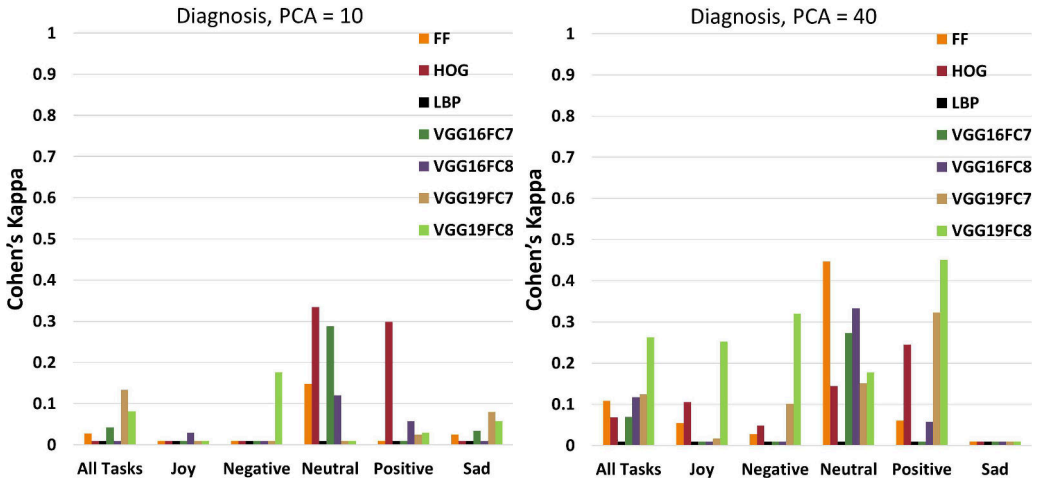


Figure 5.6: Average performance (Cohen's Kappa) of the categorical assessment **(cont.)**

Table 5.5: Best-performing classification schemes in gender dependent and gender-independent modes.

Label	Gender mode	Cond	Feature	PCA	Kappa	F1	Acc	Prec	Rec
Diag	Based	Neu	HOG	10	33.4%	41.7%	**78.1%**	**100.0%**	26.3%
Diag	Indep	Pos	V19FC8	40	**45.1%**	61.5%	76.9%	63.2%	60.0%
BDI	Based	Neu	FF	5	33.7%	44.4%	76.6%	85.7%	30.0%
BDI	Indep	Neu	V16FC7	20	37.3%	56.4%	73.4%	57.9%	55.0%
STAI	Based	Sad	V16FC7	10	36.2%	81.7%	73.8%	77.6%	86.4%
STAI	Indep	Sad	V19FC7	10	34.5%	**82.1%**	73.8%	76.5%	**88.6%**
Exp	Based	Joy	V19FC7	20	36.8%	57.1%	72.3%	60.0%	54.5%
Exp	Indep	Joy	FF	5	41.1%	57.9%	75.4%	68.8%	50.0%

Cond:Condition, Acc: Accuracy, Prec: Precision, Rec: Recall, Diag:Diagnosis, Exp: Expert Judgment, Indep: Independent, Neu: Neutral, Pos: Positive, FF: Feature Fusion. PCA: Number of extracted components.

5.2.3 Continuous Assessment

Continuous assessment (regression) was performed separately for each of the three continuous labels (i.e., BDI-II, STAI, and expert judgment score) in gender dependent and gender-independent modes. Analyses were conducted for each condition separately as well as on combined data across conditions. Participant scores were normalized on a 0-100 scale to account for the different score ranges (i.e. BDI-II: 0-63, STAI: 0-80, and annotation: 0-8). Fig. 5.7 displays average normalized RMSE values demonstrating that the proposed method succeeds best in predicting the self-reported score on the STAI. Best-performing scheme in terms on both RMSE and MAE among those listed in Table 5.6 is the gender-independent model conducted on video recordings from the Neutral condition. Additional indices of model performance are presented in Fig. 5.8 - 5.13, in the form of absolute and log-transformed differences between actual and predicted label values (BDI-II score, STAI score, or expert judgment) using the following formula:

$$\log_2 \frac{A_{norm}}{P_{norm}} \tag{5.1}$$

where A_{norm} is the normalized actual value and P_{norm} the normalized predicted value. Fig. 5.8 - 5.13 also include Bland-Altman plots [38] using the following formula:

$$M = \alpha - \beta, A = \frac{\alpha + \beta}{2} \tag{5.2}$$

where α the predicted value and β the actual value. M corresponds to the difference between actual and predicted values (plotted on the y axis) and A to their average (plotted on the x axis) .

For future clinical applications of automated assessment methods, false negative results (i.e., failure to detect significantly high scores on a psychopathological trait, are most critical. In our results, such events correspond to participants where our models significantly underestimated self-reported or expert-judgment ratings as indicated by scores >1.96 SDs from the sample mean. With respect to BDI-II scores (Fig. 5.8 - 5.9), the best performance was achieved by VGG features derived from the Neutral condition in gender-independent mode. In this analysis, there were only two false negative cases (the model significantly underestimated self-reported depression severity for two participants who scored > 55 points on BDI-II). With respect to STAI scores (Fig. 5.10 - 5.11),

Table 5.6: Best-performing continuous assessment schemes in gender-based and gender-independent modes.

Label	Gender mode	Cond	Feature	PCA	RMSE	MAE	Normalized RMSE	Normalized MAE
BDI	Based	Neu	HOG	20	10.59	7.46	16.81	11.84
BDI	Indep	Neu	V19FC7	40	10.54	7.86	16.73	12.48
STAI	Based	Neu	HOG	20	10.53	8.56	13.16	10.71
STAI	Indep	Neu	HOG	20	9.94	7.88	**12.42**	**9.85**
Exp	Based	Neg	HOG	20	1.61	1.37	20.17	17.07
Exp	Indep	Joy	V19FC7	40	1.47	1.21	18.39	15.06

Cond: Condition, Exp: Expert Judgment, Indep: Independent, Neu: Neutral, Neg: Negative. PCA: Number of extracted components. RMSE: Root Mean Square Error, MAE: Mean Absolute Error

both best-performing models underestimated self-reported anxiety severity in a two participants scoring >50 points in the scale. Both models relied on HOG features derived from the Neutral condition. With respect to expert judgment values (Fig. 5.12 - 5.13) was achieved marginally by HOG features derived from the Negative Experience Recall condition in gender-based mode. In this analysis, the model significantly underestimated expert ratings of depression severity in two participants.

The generalizability of the proposed method was tested on the AVEC'14 dataset comprised of 300 video recordings from 83 participants using continuous assessment against the only available continuous variable (i.e., individual BDI-II total scores). Comparison of results presented in Tables 5.6 and 5.7 show that our method relying on HOG features in gender-based mode produces similar results across datasets: RMSE=10.59 and MAE=7.46 in our dataset (neutral condition); RMSE=11.45/10.74 and MAE=7.81/9.92 (test/development data sets) in the comparable condition of the AVEC dataset (Northwind/passage reading).

Results using the AVEC data set permit comparison of our method with previously reported approaches listed in Fig. 5.14. This comparison shows that although the proposed methodology does not outperform the other approaches, still succeeds in performing at the same levels as state-of-the-art methods, and on top of that it sustains the same performance over two different datasets, with completely different specifications,

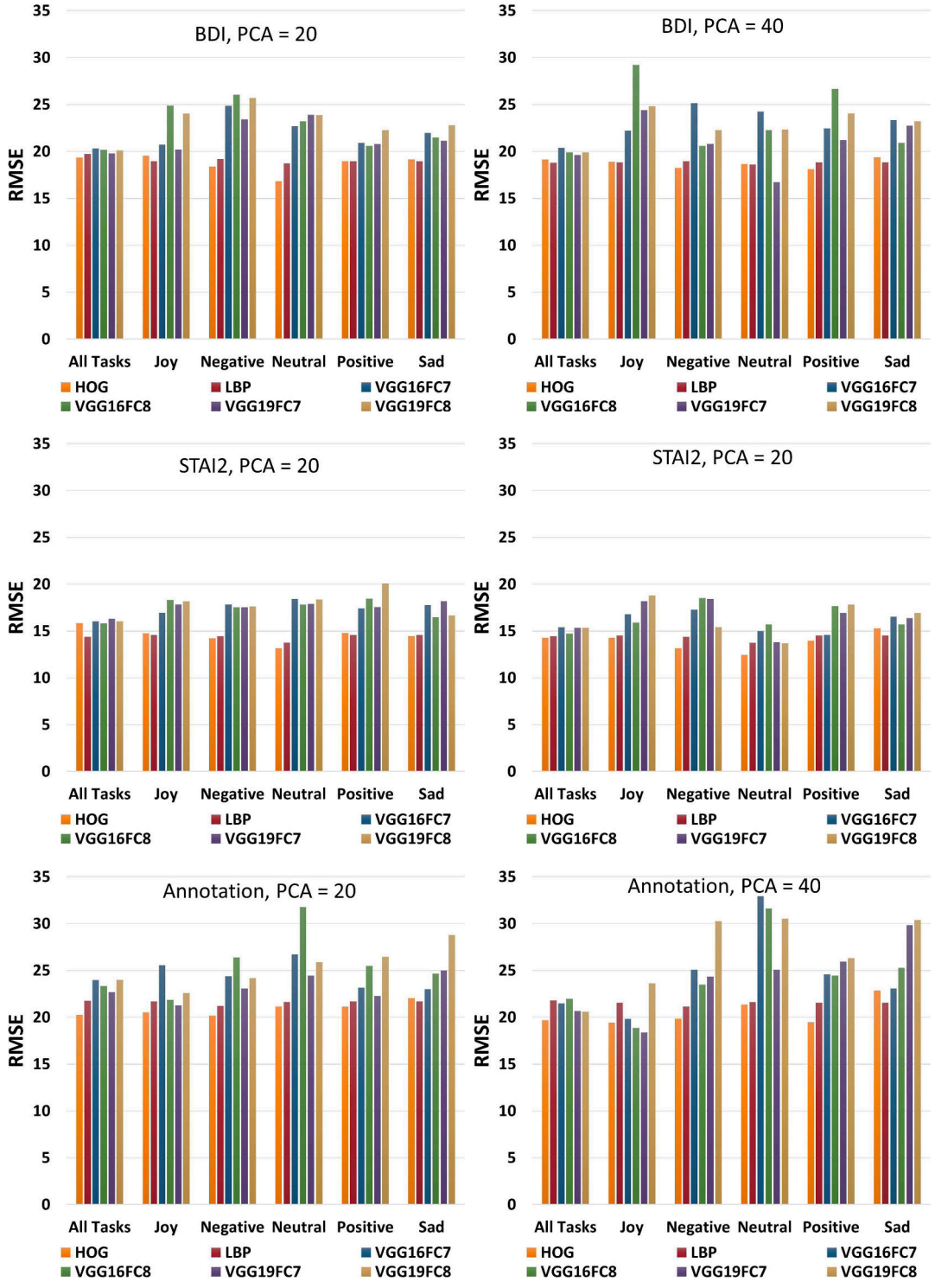
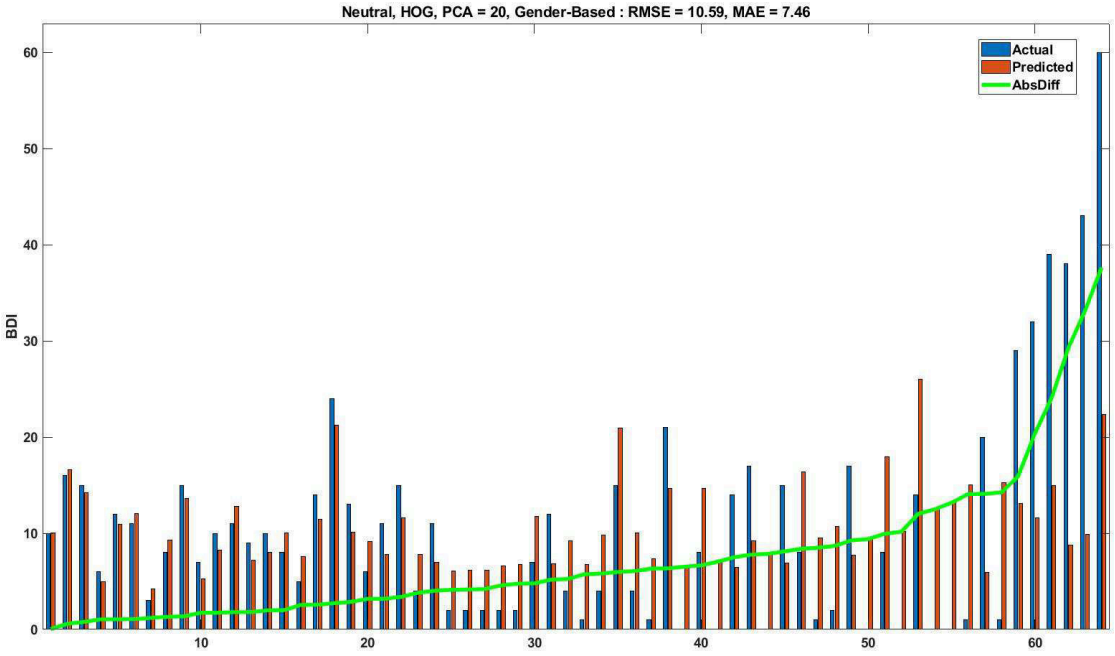
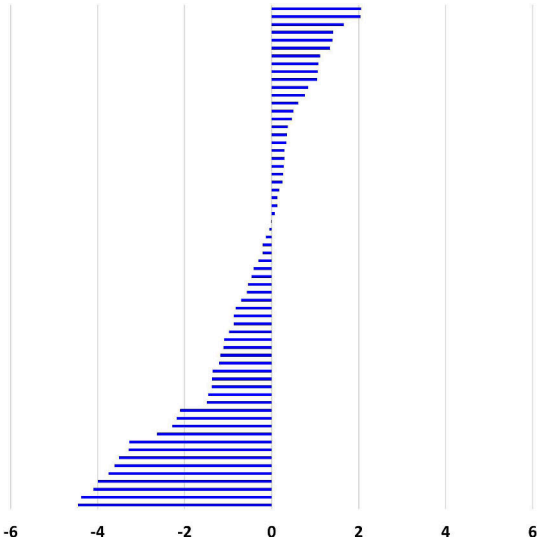


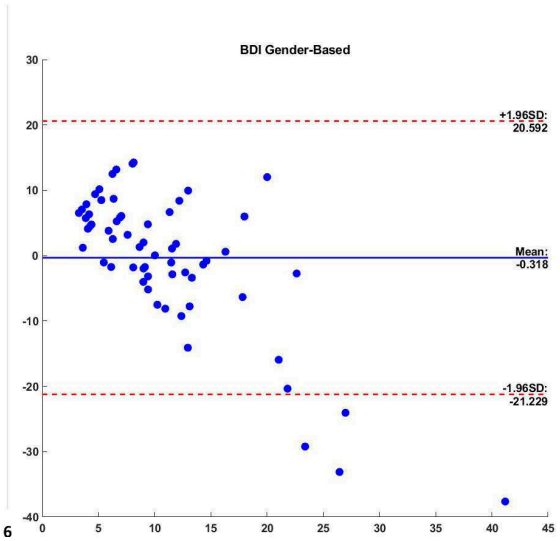
Figure 5.7: Average normalized RMSE for the continuous assessment schemes predicting BDI-II scores (upper panel), STAI scores (middle panel), and expert judgments of depression (lower panel) by number of PCA components retained during preprocessing for each experimental condition. Results for gender-based and gender-independent schemes are displayed in the left- and right-hand columns, respectively. Colors represent different video-based features.



(a)

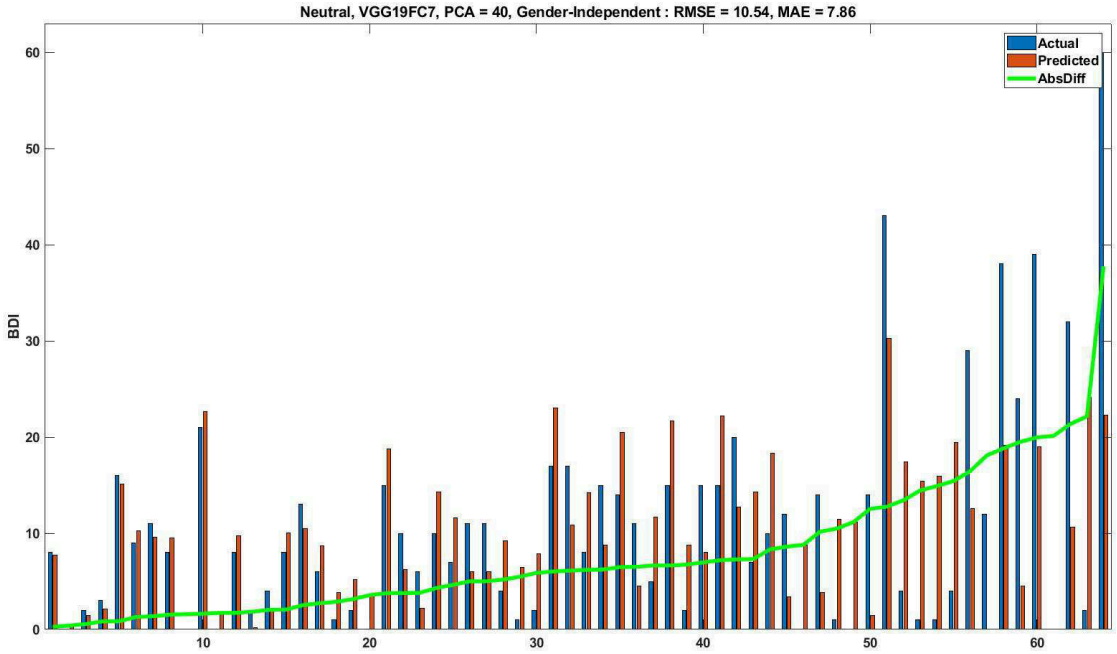


(b) Log(Actual_ norm/ Predicted_ norm)

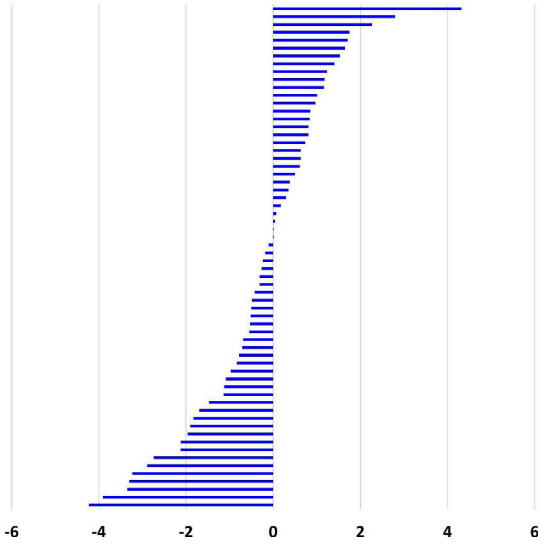


(c)

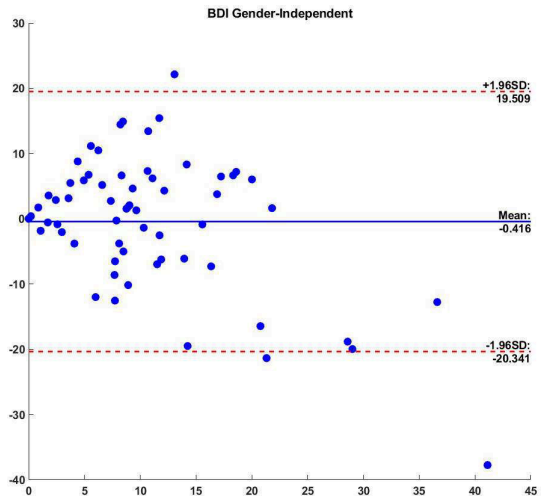
Figure 5.8: Continuous assessment: Prediction of BDI-II scores in the main experiment (gender-based mode). Upper panel: Absolute differences between actual and predicted BDI-II scores for each participant. Predicted values were computed in the context of continuous assessment analysis using recordings from the Neutral condition (HOG features). Lower panel: (b) log transformed differences between actual and predicted values; (c) Bland-Altman plot displaying the distribution of differences (y axis) over the range of BDI-II scores (x axis).



(a)

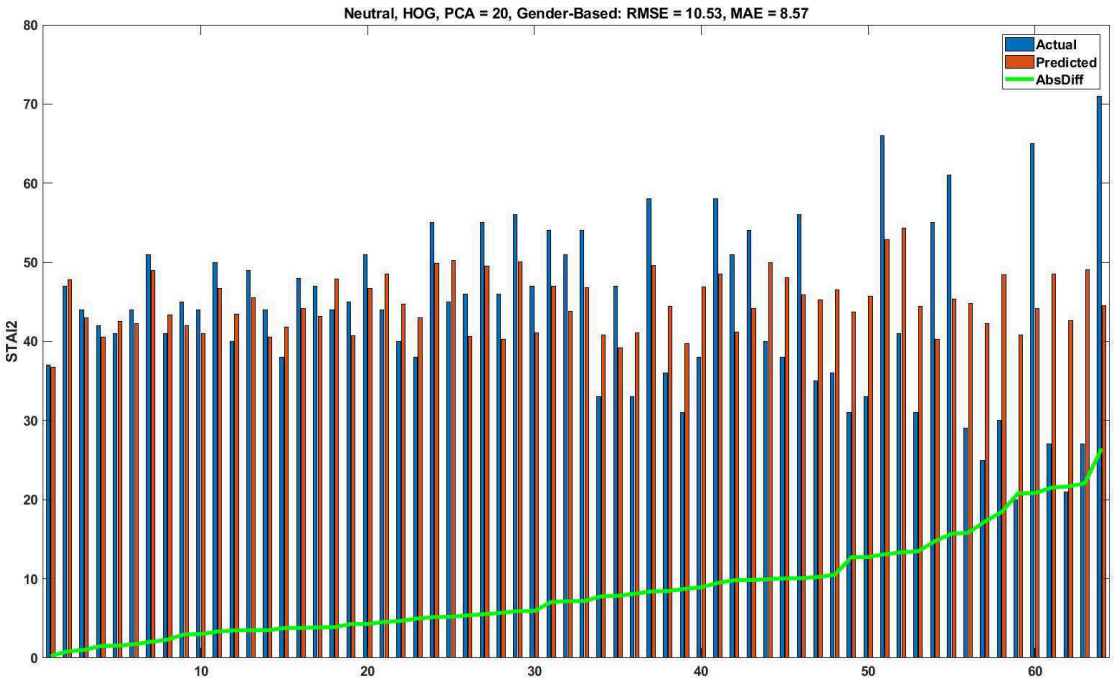


(b) Log(Actual_ norm/ Predicted_ norm)

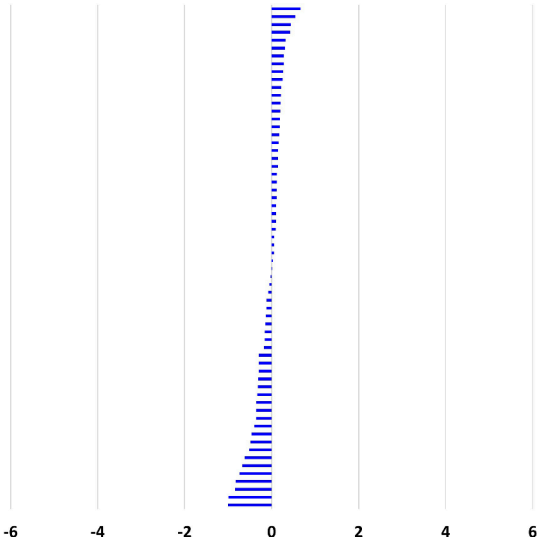


(c)

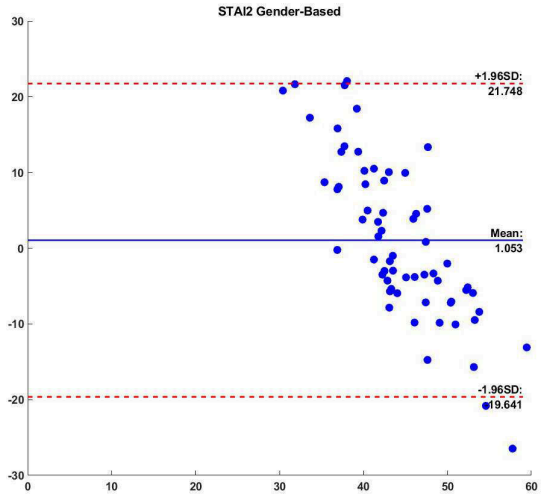
Figure 5.9: Continuous assessment: Prediction of BDI-II scores in the main experiment (gender-independent mode). Upper panel: Absolute differences between actual and predicted BDI-II scores for each participant. Predicted values were computed in the context of continuous assessment analysis using recordings from the Neutral condition (VGG features). Lower panel: (b) log transformed differences between actual and predicted values; (c) Bland-Altman plot displaying the distribution of differences (y axis) over the range of BDI-II scores (x axis).



(a)

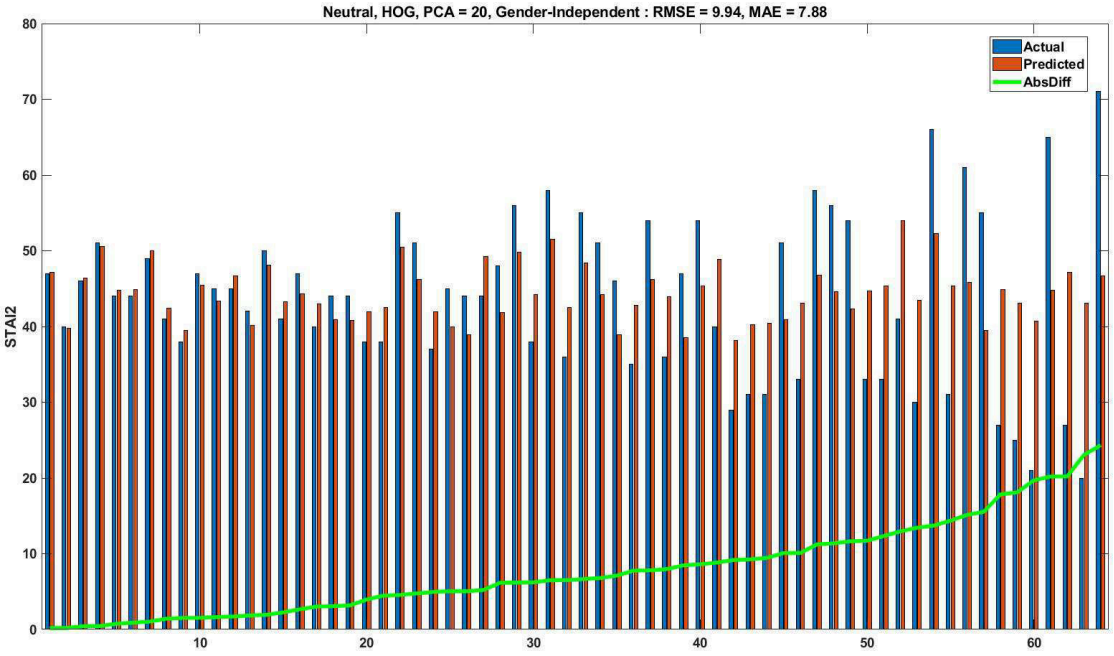


(b) Log(Actual_ norm/ Predicted_ norm)

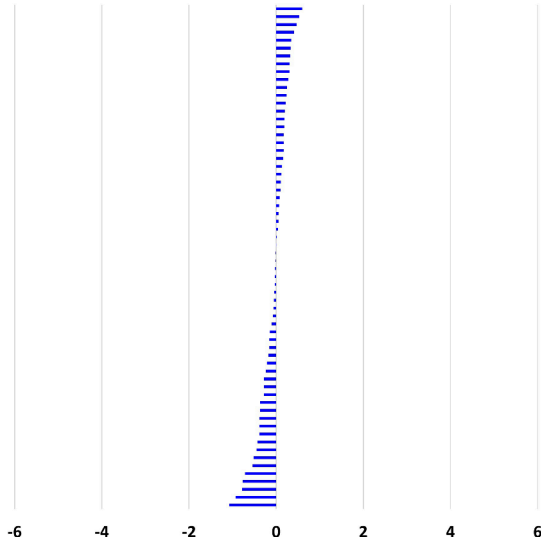


(c)

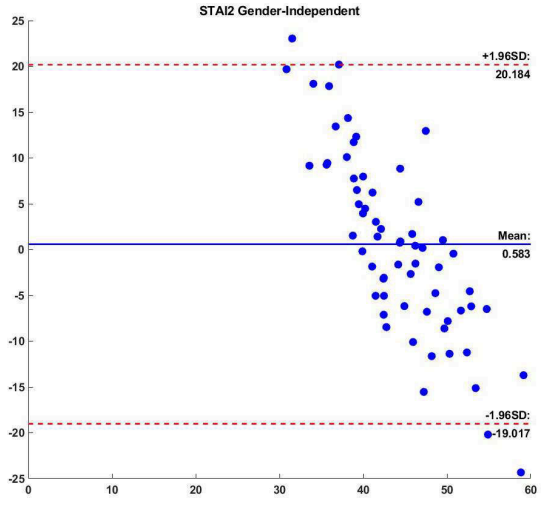
Figure 5.10: Continuous assessment: Prediction of STAI scores in the main experiment (gender-based mode). Upper panel: Absolute differences between actual and predicted STAI scores for each participant. Predicted values were computed in the context of continuous assessment analysis using recordings from the Neutral condition (HOG features). Lower panel: (b) log transformed differences between actual and predicted values; (c) Bland-Altman plot displaying the distribution of differences (y axis) over the range of STAI scores (x axis).



(a)

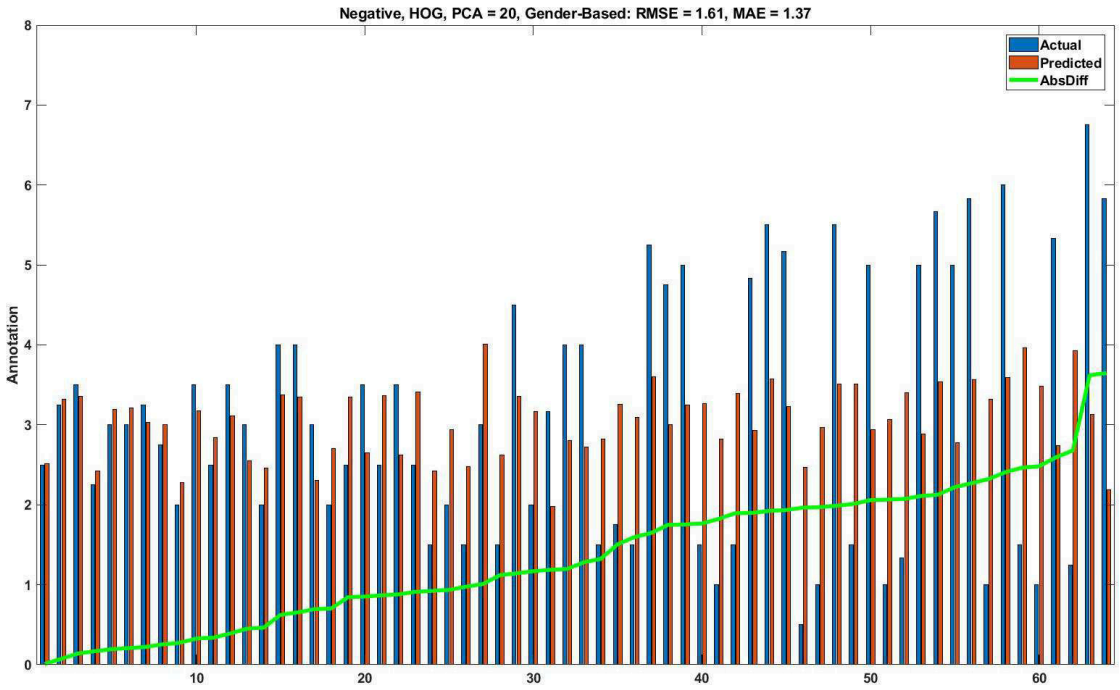


(b) Log(Actual_ norm/ Predicted_ norm)

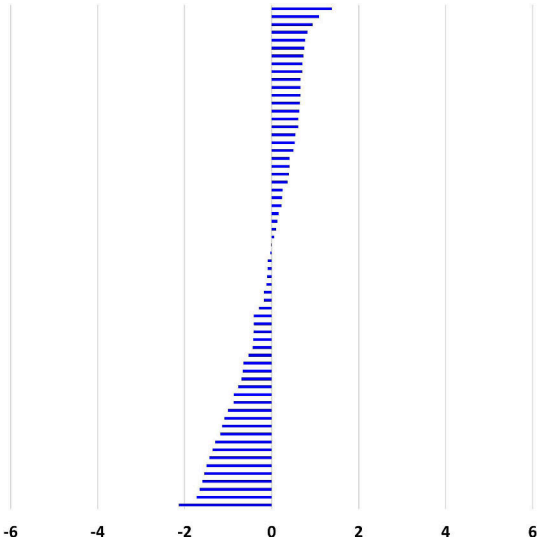


(c)

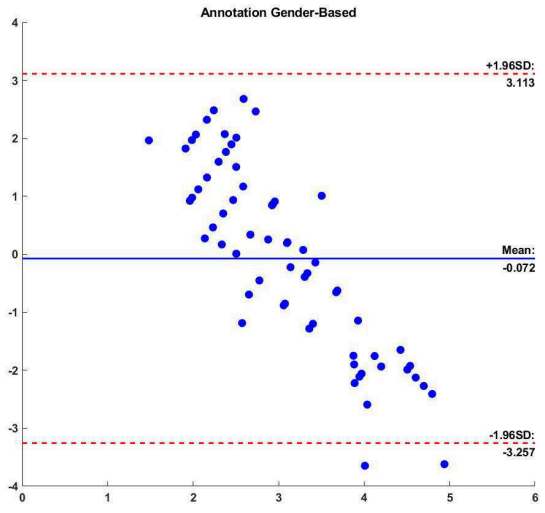
Figure 5.11: Continuous assessment: Prediction of STAI scores in the main experiment (gender-independent mode). Upper panel: Absolute differences between actual and predicted STAI scores for each participant. Predicted values were computed in the context of continuous assessment analysis using recordings from the Neutral condition (HOG features). Lower panel: (b) log transformed differences between actual and predicted values; (c) Bland-Altman plot displaying the distribution of differences (y axis) over the range of STAI scores (x axis).



(a)

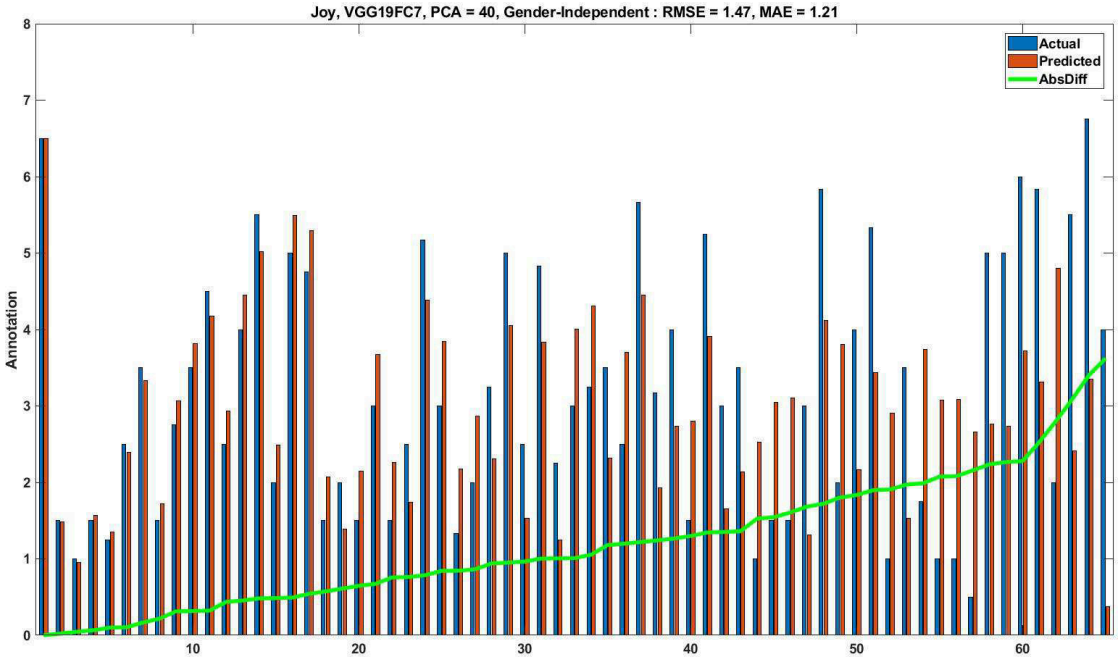


(b) Log(Actual_ norm/ Predicted_ norm)

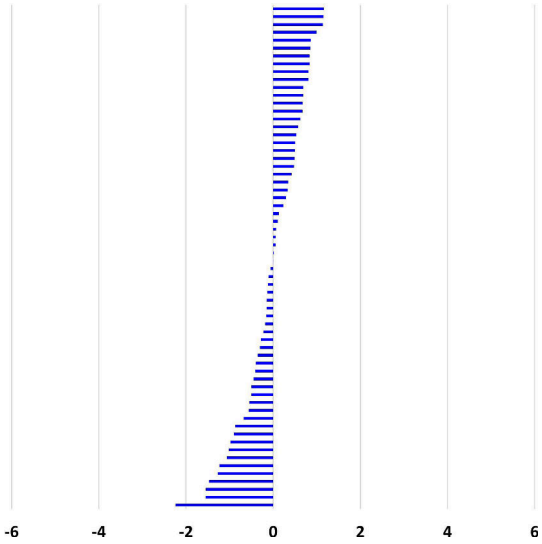


(c)

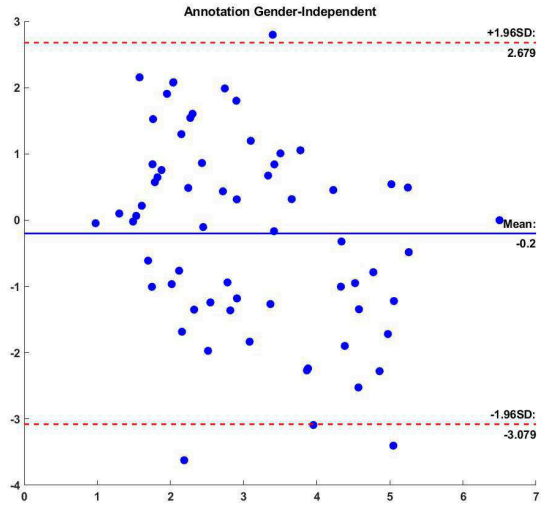
Figure 5.12: Continuous assessment: Prediction of expert judgment of depression in the main experiment (gender-based mode). Upper panel: Absolute differences between actual and predicted expert judgment for each participant. Predicted values were computed in the context of continuous assessment analysis using recordings from the Negative Experience Recall Condition (HOG features). Lower panel: (b) log transformed differences between actual and predicted values; (c) Bland-Altman plot displaying the distribution of differences (y axis) over the range of expert judgment values (x axis).



(a)



(b) Log(Actual_ norm/ Predicted_ norm)



(c)

Figure 5.13: Continuous assessment: Prediction of expert judgment of depression in the main experiment (gender-independent mode). Upper panel: Absolute differences between actual and predicted expert judgment for each participant. Predicted values were computed in the context of continuous assessment analysis using recordings from the Joy film (VGG features). Lower panel: (b) log transformed differences between actual and predicted values; (c) Bland-Altman plot displaying the distribution of differences (y axis) over the range of expert judgment values (x axis).

Table 5.7: Best-performing continuous assessment schemes conducted on various partitions of the AVEC'14 dataset in gender-based and gender-independent modes.

Test Partition	Task	Gender mode	PCA	Feature	RMSE	MAE
Test	Freeform	Based	10	HOG	10.63	8.58
Test	Northwind	Based	10	HOG	11.45	9.92
Test	Freeform	Indep	20	HOG	**10.15**	**8.48**
Test	Northwind	Indep	40	HOG	10.95	9.22
Development	Freeform	Based	5	HOG	9.20	**7.81**
Development	Northwind	Based	40	HOG	10.74	8.91
Development	Freeform	Indep	45	HOG	**9.15**	7.83
Development	Northwind	Indep	40	HOG	10.97	9.33
LOSO	Both	Based	90	HOG	10.96	8.89
LOSO	Both	Indep	100	HOG	**10.89**	8.87

LOSO: LOSO cross validation on the entire dataset

protocols, and of different cultures. It is remarkable that across different setups for the AVEC dataset HOG is always the best performing feature set.

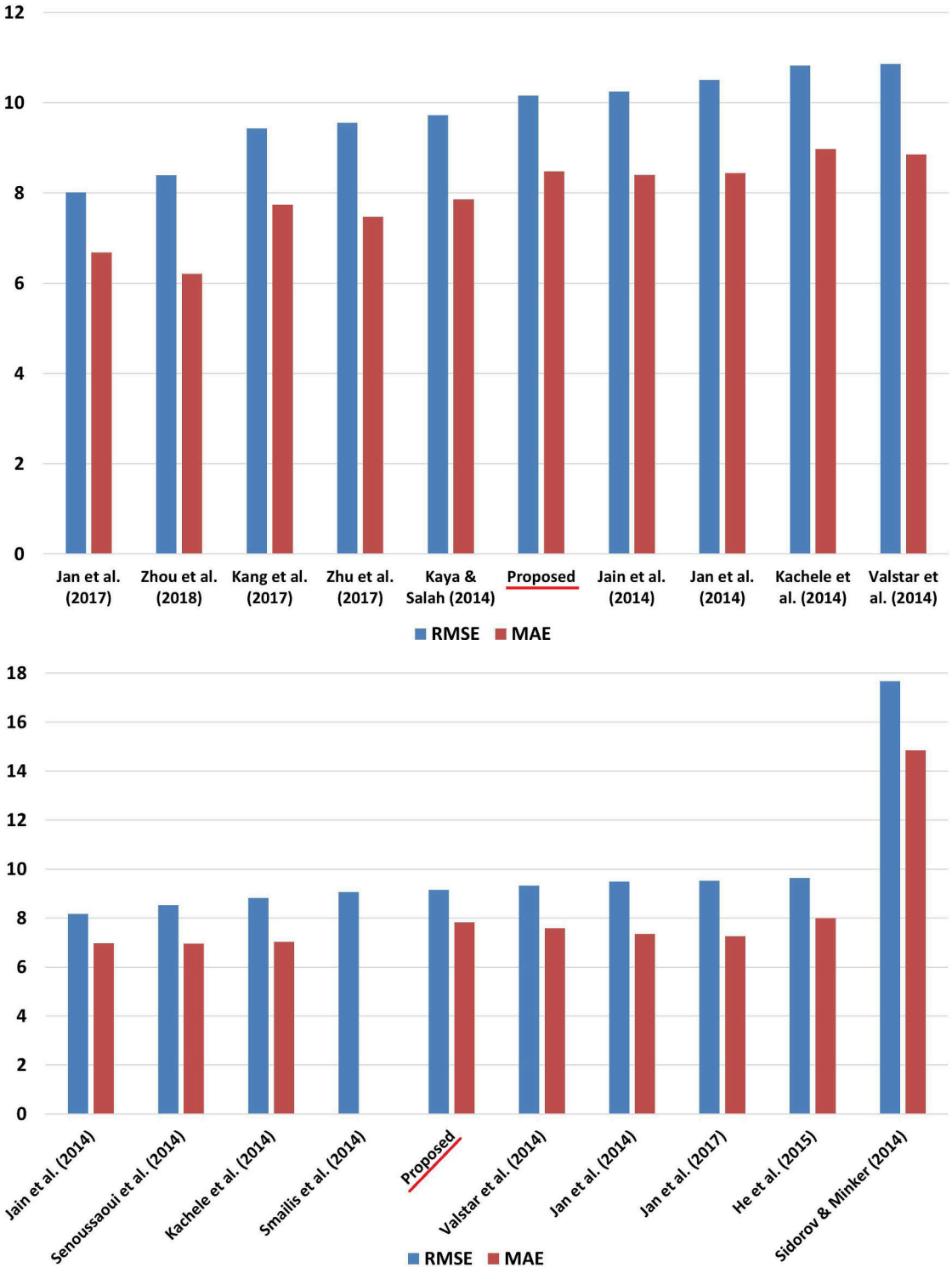


Figure 5.14: Comparison of our approach (Proposed) with previously reported results on the AVEC'2014 Test (upper panel) and Development (lower panel) datasets.

Table 5.8: Comparison of several specifications between the AVEC dataset and our dataset

	AVEC	**Our Dataset**
Number of recordings	300	322
Number of participants	83	65
Age [M (SD)]	31.5 (12.3)	42.4 (11.92)
Male rate	32.67%	30.77%
BDI-II [M (SD)]	15.06 (11.9)	11.2 (11.73)
Participants	Volunteers	Volunteers & Patients
Country	Germany	Greece
Protocol	Non-social	Interpersonal & Non-social
Setup	Independent	Controlled
Illumination	Independent	Controlled indirect lighting
Image Resolution (pixels)	640×480	1920×1920
Facial image size (pixels)	112×112	600×600
Frame Rate (fps)	30	80

Chapter 6

Discussion

The main outcomes of the research performed described in the present work are discussed below organized in three sections: a) Issues pertaining to algorithm development and performance, b) Data related issues, and c) Plan of future work.

6.1 Algorithm Development and Performance

Categorical vs Continuous Assessment One of the main issues addressed in the present work concerned a comparison of categorical versus continuous assessment. Initially, we pursued classification problems which are computationally more straightforward. However, computational approaches developed and applied to the AVEC dataset with apparent success failed to generalize to the data obtained in the context of the main experiment. Such poor performance motivated adoption of regression approaches which, eventually, appeared to perform comparably across datasets.

Importantly, one of the depression related labels (clinical diagnosis, based on clinically acceptable BDI-II cut-off, or based on expert judgment of presence of depression based on visual cues) were predicted at an acceptable level using categorical assessment. This failure however may not be so crucial after all, as the machine learning algorithm does not need to make a diagnosis, just a recommendation for further exploration; as already made clear, the proposed methodology is supposed to serve as a decision support tool, and not as a standalone system. In addition, the relatively poor classification performance may reflect the continuous nature of depressive symptomatology. This problem is compounded by the potential moderating role of clinical factors (such as

type of medication, type and duration of treatments) and person-specific characteristics (illness-related cognitions and personality characteristics) that may affect the intensity and quality of facial expression of depression symptoms (e.g., negative mood, apathy, helplessness). Furthermore, depression does not have clear categories, and thus cannot be easily treated as a classification problem. The apparent continuum of severity of depression symptomatology is expected to be reflected in a corresponding continuous manner in facial motion dynamic patterns across patients rendering continuous assessment more appropriate.

Low vs high level features Although a trend toward utilizing high-level features has been observed, and promoted by AVEC'16, where features were provided after preprocessing to enable high-level feature extraction, it seems that they are not the best performing. The high-level features are more appealing though, as they enable direct interpretation and give insights more easily. In terms of the low level features, LBP did not perform well in terms of classification, for the given set of parameters, resulting in near zero recognition most of the times. A potential improvement to LBP may entail adopting larger radius and neighborhood settings, as the ones selected here may represent only micro-movement patterns, or even employing another variant of LBP, such as weighted LBP. Remarkably, LBP, although failing in classification, it was among the best performing in the main experiment for predicting individual BDI-II scores in the gender-based mode. Overall the best performing low level descriptor in many different experimental setups was HOG, which can be justified by the fact that it retains spatial information, and does not construct a unique histogram for the whole image.

Deep learning The deep learning approach outperformed the other approaches in most settings, both in Experiment 5 (see section 4.6) and in the main study. However, given that only the generic VGG was employed in the proposed work, based on its previously reported performance, it is highly probable that after training and tuning the network, the rate of correct depression assessment could improve significantly. The highly competitive performance of deep learning-based features has been noted in several research areas. The fact that deep learning does not rely on prespecified rules, in a manner similar to human cognition, may account for its superiority. The performance of deep learning algorithms with a relatively small data set, such as the one tested in

the present work, generates great promises regarding their capacity to provide clinically meaningful results for depression assessment provided with sufficiently large training data sets.

Motion representations Results pertaining to the visual features based on LMHI as extracted from the available data set in Experiment 3 (see section 4.4) were rather surprising given that analyses highlighted a single visual feature (LMHIFaceHOG) as the most significant. This could be explained by the fact that LMHIFaceHOG incorporates motion information, and by being registered and resized, it minimizes appearance-based variation. Furthermore, LMHI requires significantly less information (image frames vs. selected landmarks), and ensures participant anonymity by retaining only facial landmarks - an important attribute in studies with clinical samples.

Among additional motion history-based algorithms evaluated (MHI and GMHI) in Experiment 5 (see section 4.6) the original MHI algorithm performed better than the proposed variants. The rest of the proposed motion representations, namely the LCBP-TOP and LCBP-POP, as well as the one based on facial geometry, did not perform as well as the motion images. Finally, in terms of the window based approaches, the duration of windows varied for the different approaches, which means it is highly depending on the specifics of the approach (e.g. method, dataset, etc).

Cross validation An important issue to consider when evaluating published reports on automatic depression assessment concerns the use of relatively small samples and suboptimal cross validation methods, which are highly susceptible to model overfitting.

Performance metrics The experience gained through the experimental tests leads to the conclusion that performance metrics need to be considered in combination. Most of the reported approaches in the literature report only one metric (e.g. accuracy) which does not reliably show the capacity of a model, especially in cases of highly unbalanced datasets. Further metrics such as Cohen's Kappa encompasses most pertinent information and it is not surprising that the majority of complementary metrics typically follow Kappa values across studies. The F1-score, which has been very popular in previous reports, does not necessarily reflect accurate recognition across all classes (i.e., a high F1-score may be associated with good recognition in one class and poor recognition in

another) and does not consider the level of chance. Thus it is highly relevant to report a set of metrics, rather than choosing a single one.

6.2 Data Related Issues

Sample size Many approaches, along with some of the ones proposed hereby, report very high detection accuracy rates, a fact that clearly demonstrates the clinical potential of the field, but sample sizes are often too small to enable the generalizability of these results. In order for a system to be fully evaluated and acknowledged as an assessment tool, it must be tested on considerably larger sample sizes, featuring a wider variety of demographic characteristics, clinical diagnosis methods, and ethnic-cultural backgrounds.

Comorbidity & subtypes Comorbid diagnoses, should be carefully recorded and used to evaluate potential misclassifications, given the high comorbidity rates between PTSD, anxiety and depression [178]. In addition the capacity to distinguish between different depression subtypes, and MDD from other mood disorders also needs to be addressed [195]. Individual variability due to comorbid personality disorders or characteristics, as well as the influence of ethnicity and culture requires further exploration [25].

Multiple sessions Furthermore, as reported section 1.1, a one-off clinical assessment may not be sufficient neither for diagnosis or registration of facial features, as development of rapport with the participant is necessary. For this to be achieved several sessions over a fixed interval (e.g. 7 weeks as in [53] [90] [89]) are advisable. Existence of baseline data would also be useful, but unfortunately this is not possible in most cases. However, given the importance of symptom/sign stability for depression diagnosis, repeated recordings over several days or weeks would render results more clinically relevant. Finally, multiple sessions can benefit the remission assessment, allowing long-term monitoring of the recovery process, as well as providing a personalized model based on the built rapport.

Real-world vs laboratory setting In the field of automatic facial expression recognition (AFER) approaches are moving toward real-world conditions [143], as exemplified by the Emotion Recognition in-the-Wild (EmotiW) challenge series [66] [67] [68]. The manner in which the AVEC dataset was constructed also supports this idea, as the recordings took place in independent setups and on personal computers. This choice, however, impacted performance, as shown in Table 2.5: approaches for categorical assessment of depression based on AVEC demonstrated lower than average performance, when compared to approaches based on other datasets.

Additionally, although current in-the-wild approaches may be considered as promising, they are not yet sufficiently reliable even for AFER as supported in [180] [143]. Therefore, at present, such approaches do not appear to meet minimum requirements for a clinical decision support system. On the other hand, the strict requirement for standardization of data collection [60] may impose potentially serious limitations, such as questionable originality and genuineness of the data and lack of variance in contextual information. Although standardized medical equipment typically operates under highly controlled conditions, collection of data indicative of depressive symptomatology is highly susceptible to the dynamic nature of behavioral and underlying psychological processes of the person being evaluated.

Stimulus In terms of the benchmark datasets employed for the preliminary experiments, the DAIC-WOZ presented a better interpersonal context for depression assessment in view of the extant literature supporting the better suitability of interviews for detecting signs of depression. However, in the tests performed in terms of the main study the processing of clips recorded during the neutral text reading (non social context) turned out to be the best performing in several setups. Some disadvantages of the neutral text reading task involved some reading problems from the participants, language issues, sight problems, etc.

Testing clips combined across tasks was never ranked among the best performing. It is obvious that comparison, either in terms of categorical assessment or continuous, must be under strictly defined protocol, and during executing the same task. Furthermore, with the exception of diagnosis, which has their best performance in both Neutral and Positive stimuli, for gender-based and gender-independent respectively, the rest labels have their best in the same stimulus without regard to the gender dependency mode.

More specifically, for BDI the best performance is with neutral stimulus in both gender-based and gender-independent modes. In the same rationale, the highest performance for STAI prediction was again with the neutral stimulus for both gender modes. In terms of the expert judgment values, the best result was for the Negative Experience Recall condition in gender-based mode.

Annotation During the preliminary experimental tests the need for clinical data, involving diagnosed patients and not just volunteers as in the benchmark datasets, is stressed. It was assumed that a score on a single self-report instrument may not be sufficient to establish reliable and valid classification of individuals according to depression status. Other factors can readily elevate self-report scores on depression scales, such as the presence of significant, long-term life stressors and anxiety. Thus, a clinical interview would be more reliable for building a robust dataset.

In the proposed work, the data were collected from both non-diagnosed individuals to form the control group, and diagnosed MDD patients. In addition, a team of psychologists rated each video and provided an overall assessment of depressive signs, to provide an alternative benchmark which would be more closely matched to the source of information employed by the proposed methodology (facial images). The annotation was another novelty by the proposed work, as it has never before been provided by other datasets, which employ just a self-report score (BDI, PHQ9, etc), or the score of a clinically-administered instrument (e.g. HAM-D). The parallel use of STAI was an additional novelty.

However, the results did not prove the hypothesis, as the training and testing based on the diagnosis did not provide good results, which can again be attributed to the fact that depressed status is not uniform. There are several and very different ways that patients appear or behave, which is highly relevant to their severity of depression. Although all patients included in the dataset were medicated and diagnosed, they were in different stages of remission, as well as with different levels of severity, thus considering them as a uniform class is false.

STAI scores were predicted much better than any other variable, which implies that anxiety can be portrayed and evaluated more accurately by the specific methodology. Results are well above chance for all metrics involved in categorical assessment of STAI, and also in terms of continuous assessment it performs best too. Although the best

Cohen's Kappa is for diagnosis, yet it is still low (45.07%). In addition, and although if we consider each metric independently, again diagnosis has the best accuracy and precision, still it fails for F1-score and for Recall. A better insight of the results was given by checking which setup had the overall best performance, a rule of "which setup has all metrics above 70%" was set (excluding the kappa). The STAI was the only one to conform with this rule, for both gender based and gender independent. This seems quite promising, in both cases for the Sad stimulus.

Gender dependency The gender-based model did not perform very well, while the gender-independent performed much better. This can be explained by the fact that by separating male/female subjects then the remaining training samples for each model are quite low. Therefore gender dependency was not properly evaluated, and again there is a need for a greater sample, as even the male rate is quite low. It seems that it was not really meaningful to attempt gender-based approaches in the first place, as the training sample was too low.

Specifications Based on the literature review it was predicted that video acquisition conditions are of great importance, as they often affect processing tasks such as face detection, and that to this end, illumination, image resolution, frame rate, are some crucial factors to be considered. However, this prediction was not corroborated in the main study, as the categorical approach using the same methodology did not perform as well as in the AVEC dataset, while for the continuous assessment the perform in the same level. Finally, by constructing a high-specification dataset the computational cost is raised significantly, as higher resolution frames and at a higher frame rate produce a considerably higher amount of data to be processed, which directly affects the pipeline; for instance for the example of the motion history image, having a frame of $1200x1200$ at a frame rate of 80 fps would be much more complex to be computed compared to an image of $480x640$ at a frame rate of 30 fps.

Failed predictions We made an effort to account for the failed predictions of the best performing method, which is the continuous prediction of the STAI self-reported score for the neutral stimulus in the gender-independent mode. The most notable failure concerns underestimation of STAI scores, given the higher associated clinical risk. Among the four

patients in this category, two suffered from severe depression and were treated with high doses of combinations of antidepressants which may have affected the dynamics of facial expressions. Another patient spoke Greek as a second language and experienced some difficulty in reading the text, while the fourth patient also had difficulty reading due to reduced visual acuity. These issues can be considered as limitations of the proposed methodology, namely: individual medical treatments, language and vision problems.

6.3 Future Plans

In this section potential solutions to problems identified in the current experimental work are proposed.

Algorithms In future work improvement of feature selection methods is probably the best avenue to enhance classification performance. Inspection of the bivariate and partial correlation matrix between individual features and using probability-based statistic indices (such as Fisher's z) to identify significant associations may help optimize feature selection.

Although performance achieved here in terms of categorical assessment for the AVEC'14 outperforms related work, there is still room for significant improvements. Future work may attempt training and tuning the VGG, testing different versions of the network, assessing the performance of additional classifiers, as well as attempting decision fusion for the different motion representations and feature combinations. Additional improvements could involve applying windowing to the different motion image variants, using codebook approaches for dimensionality reduction (e.g. Bag of Words).

Multimodal Although this thesis is focused on image processing, work conducted independently in our lab employing multimodal approaches showed promising results [161] [163] [164]. These preliminary results are in agreement with the systematic literature review presented in [165]. Therefore, combining audio based features [187] could potentially improve overall performance. In addition, during the data collection physiological signals were collected too, at the same time with visual and audio signals; blood volume pulse (BVP), and galvanic skin response (GSR) could also add value to the performance.

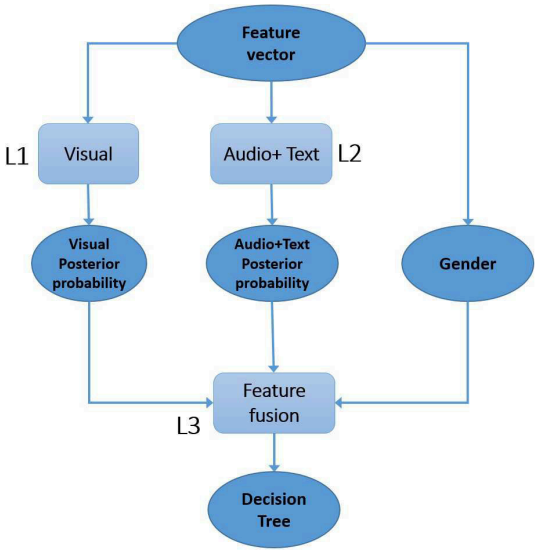


Figure 6.1: Posterior probability classification model for fusion

Combining different types of signals though is another challenge to be addressed. Different fusion schemes can be found in literature, such as the feature level fusion (mere concatenation of individual features). Decision level fusion is another fusion method, which considers different ways of combining the predictions from individual models as derived from the different signals, such as combining them through AND and OR operands, or using weighted fusion methods. Stacking is another method for implementing decision fusion [183], and also the Posterior Probability Classification Model [79] (c.f. Fig. 6.1) which was also used in [161] during our participation the AVEC'16 challenge.

Data Efforts to develop methods capable of differentiating mood disorder types, and also depression from other psychiatric disorders, such as various anxiety conditions, are not many. Although some notable initial attempts in this direction can be found in [194] [195] by focusing on distinguishing PTSD from depression, and another in [234] attempting to distinguish cases of unipolar depression versus bipolar disorder, still they are quite limited, and require further exploration.

Additional manifestations may need to be considered, but would need also additional specifications for the setup to be adjusted. Pupil-related features, for example, would be interesting to investigate in an non-obtrusive manner, that is without using head-mounted eye-tracker devices, but still highly constrained environment would be

required, with a very expensive camera in order to achieve resolution and illumination conditions able to monitor the pupil activity. Monitoring body gestures would also require additional cameras and a different setup; in order to be able to monitor the body, we should not have a desk or other occlusions, rendering an interview setup as the most appropriate.

As already established above, the small number of subjects in each dataset is another issue. There is a need to test the validity in terms of a clinical study, including potentially thousands of samples. This is an issue, as the recruitment of patients, and having them to commit is very hard. This is the reason that there is still a long way from standardizing such a system. An approach for cross-corpus / cross-cultural method evaluation by Alghowinem et al. [24] highlighted this issue as well.

Chapter 7

Conclusions

Research on automatic depression assessment has come a long way from Cohn et al. [53] and McIntyre et al. [145], when they introduced the field in 2009, with several novel approaches. The proposed work provided a number of insights, while identifying many questions open to further investigation. Depression diagnosis itself is an active and controversial topic in clinical psychology and psychiatry. Given the aforementioned outstanding issues, the development of automated, objective assessment methods may be valuable for both research and clinical practice. Furthermore, in this Chapter we discuss questions posed and summarize major contributions achieved within the present thesis.

7.1 Research Questions Addressed

Several research questions that were set at the beginning of this research were addressed in this thesis as described in some detail below.

Construct a clinically valid dataset, comprising diagnosed patients as well as healthy control individuals This question was successfully addressed, overcoming several, mostly expected, obstacles in this line of work. In the end data were collected from 65 individuals (45 controls and 20 patients), providing 322 video recordings across 5 experimental conditions. Facial image, speech, and BVP and GSR were recorded through video, audio, and physiological signals respectively.

Compare interpersonal vs non-social contexts for video recording Different types of stimuli were tested, both of non-social and interpersonal contexts, involving watching video clips selected to elicit specific emotions, answering questions from interviewer, and reading out loud a neutrally charged passage. However, the results of the tests did not provide a clear answer to the question, of which context is the best to evaluate the depression state. Although for our dataset it was the neutral task which provided the best performance in most settings, for the AVEC dataset the best results came for the Freeform task, which entailed responding to various questions. Given the many differences between the two datasets, it is not so straightforward to explain why this happened.

Evaluate the proposed methodology against several independent sources of information pertaining to participant psychoemotional status: Clinical Diagnosis, Self-reported symptoms of depression and anxiety, Expert judgment of visible depression manifestations. Results indicated that the proposed method was more sensitive to facial features more closely associated with self-reported anxiety (STAI scores). This fact could imply that anxiety manifestations are better detectable using the proposed methodology than manifestations of depression. The proposed method performed well for both gender-based and gender-independent modes, with best results with video recordings obtained during a non-challenging, emotionally or cognitively, condition (reading a neutral passage).

Develop video-based methodology to extract features correlating with signs of depression Several methodologies were developed in the context of this thesis, which were initially evaluated on benchmark datasets. The best performing methodology was also tested on our dataset, and proved to have consistent performance across two completely different datasets. The consistency in performance was also in terms of predicting BDI-II scores.

Experiment with different video acquisition parameters Different video acquisition parameters were set during the data collection intentionally, to evaluate their contribution to model performance. At this moment it is not clear whether higher specifications contribute to higher performance, as the best performing algorithm with the

AVEC dataset did not outperform results based on our dataset (which was obtained at higher video specifications).

Investigate categorical vs continuous depression assessment While both approaches were investigated, greater emphasis was given to the categorical assessment. Development of the focusing on classification spanned several years, whereas continuous assessment was systematically pursued during the final stages of the thesis research. It should be noted however that continuous assessment may be more suitable to the nature of measures used to diagnose and quantify depression severity.

7.2 Major Contributions

The major contributions of this thesis can be summarized as follows:

- Systematic literature review of relevant approaches, published in IEEE Transactions on Affective Computing (I.F. 4.585 / Q1) [165].
- Several methodological contributions by implementing novel motion representation algorithms, namely the Local Curvelet Binary Patterns-Three Orthogonal Planes (LCBP-TOP), Local Curvelet Binary Patterns- Pairwise Orthogonal Planes (LCBP-POP), Landmark Motion History Images (LMHI), and Gabor Motion History Image (GMHI), published in several peer-reviewed IEEE [159] [160] [163] and ACM [161] conferences, and EURASIP Image and Video Processing journal (I.F. 2.455 / Q2) [162]. All implementations will be will be soon made available in MathWorks.
- Classification using four severity classes of self-reported depressive symptomatology (minimal, mild, moderate, and severe) was performed in two of our published approaches [159] [160], which are the only ones to the best of our knowledge. The remaining published approaches attempted binary classification, three-class classification, or regression.
- Categorical assessment of depressive symptomatology was performed using deep learning methods, for the first time on the AVEC dataset.
- Construction of a clinically valid dataset in terms of depression and anxiety.

- The proposed methodology, as presented in the main study was validated across datasets (AVEC and our dataset) presenting a competitive performance for predicting individual BDI-II scores, submitted for publication to the IEEE Journal of Biomedical and Health Informatics (I.F. 3.85 / Q1).

Appendix A

Complete List socio-demographics participants profiles

In the following table the detailed list of socio-demographics participants profiles is presented. The first column correspond to the serial number of the participant. The column 'Diagnosis' takes binary values, and corresponds to the participant being a diagnosed for depression (1) or not (0). Gender takes the value 0 for male participants and 1 for female, while the column age is for how many years old they are. Years of education correspond to years of education the participant has received since the age of compulsory education (6 years old) until the day of the interview, stating the length of the mandatory attendance to a course (e.g. if the length of a bachelor degree was 3 years but the participant took 4 years to complete the course then 3 years was what was added to the overall length). The columns BDI and STAI correspond to the respective scores of the given instruments, while the annotation column corresponds to the score attributed to the participants from the blinded experts.

Table A.1: Complete socio-demographics participants profiles

#	Diagnosis	Gender	Age	Years of Education	BDI	STAI	Annotation
1	0	1	27	18	0	38	2
2	0	1	29	17	15	54	1.5
3	0	0	35	6	12	40	1.5
4	0	0	43	25	1	45	3

Continue on the next page

Table A.1: Complete socio-demographics participants profiles **(cont.)**

#	Diagnosis	Gender	Age	Years of Education	BDI	STAI	Annotation
5	0	1	30	20	21	58	0.5
6	0	0	41	16	0	29	2
7	0	1	35	22	16	44	3
8	0	1	43	16	1	30	4
9	0	0	32	19	13	48	2.5
10	0	1	53	17	11	56	5.33
11	0	0	37	15	4	38	2.5
12	0	1	49	17	1	27	2.5
13	0	0	37	16	0	25	1.33
14	0	1	54	7	15	49	2.5
15	0	1	50	12	14	46	4.75
16	0	1	51	18	1	35	3.25
17	0	0	36	23	2	45	5
18	0	0	34	25	0	38	1
19	0	1	34	14	2	33	1
20	0	1	56	14	10	47	3.25
21	0	1	26	17	8	40	1.5
22	0	0	37	16	3	42	3.5
23	0	1	51	12	11	51	6
24	0	0	39	17	11	55	2.25
25	0	1	35	16	0	36	1
26	0	1	38	16	11	33	1.5
27	0	1	37	9	12	50	4
28	0	1	32	16	2	33	1
29	0	0	41	17	0	31	1.5
30	0	0	42	12	4	20	1.5
31	0	0	39	17	7	31	2
32	0	0	46	17	0	47	3.5
33	0	0	43	24	6	38	2.5
34	0	1	40	18	4	36	1.5
35	0	1	43	33	15	51	1
36	0	1	24	14	8	45	2
37	0	0	49	20	2	21	1.5
38	0	1	43	18	5	44	3.5
39	0	1	34	19	14	47	3.17
40	0	0	51	12	1	27	1.5

Continue on the next page

Table A.1: Complete socio-demographics participants profiles **(cont.)**

#	Diagnosis	Gender	Age	Years of Education	BDI	STAI	Annotation
41	0	1	38	18	2	46	3
42	0	1	35	18	8	44	1.25
43	0	1	50	17	7	44	3
44	0	1	32	18	4	31	2
45	0	1	47	3	8	44	1.75
46	1	1	58	0	32	61	5.83
47	1	1	46	12	39	55	5.17
48	1	1	62	7	9	41	3.5
49	1	1	56	6	17	51	5
50	1	1	36	17	29	54	3.5
51	1	1	63	14	20	54	4
52	1	1	45	8	10	44	6.75
53	1	1	62	8	17	41	5
54	1	0	63	6	6	37	4
55	1	1	53	12	2	51	4.5
56	1	1	52	6	60	66	5
57	1	1	30	18	14	41	2.75
58	1	0	45	6	24	55	5.83
59	1	0	70	6	8	40	5.67
60	1	1	62	6	15	56	5.5
61	1	1	24	18	15	58	4.83
62	1	1	31	15	43	71	3
63	1	1	56	13	10	47	6.5
64	1	1	40	12	28	71	5.25
65	1	1	40	12	38	65	5.5

References

[1] ACM. URL http://dl.acm.org. 10

[2] Elsevier. URL http://www.sciencedirect.com. 10

[3] GoogleScholar. URL http://scholar.google.gr. 10

[4] IEEE. URL http://ieeexplore.ieee.org. 10

[5] MedPilot. URL http://www.medpilot.de. 10

[6] NASA. URL http://lsda.jsc.nasa.gov. 10

[7] Oxford. URL http://www.oxfordjournals.org. 10

[8] PubMed. URL http://www.ncbi.nlm.nih.gov/pubmed. 10

[9] Scopus. URL http://www.scopus.com. 10

[10] Springer. URL http://link.springer.com. 10

[11] URL http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm. 67

[12] Wiley. URL http://onlinelibrary.wiley.com. 10

[13] Mayo Clinic, Feb. 2015. URL http://www.mayoclinic.org/diseases-conditions/depression/basics/tests-diagnosis/con-20032977. 10

[14] Survey of Health, Ageing and Retirement in Europe, Feb. 2015. URL http://www.share-project.org. 2, 10

[15] World Health Organization - Regional Office for Europe, Feb. 2015. URL http://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/news/news/2012/10/depression-in-europe/depression-in-europe-facts-and-figures. 10

[16] National Comorbidity Survey - Harvard Medical School, Sept. 2016. URL http://www.hcp.med.harvard.edu/ncs/. 10

[17] World Health Organization, May 2017. URL http://www.who.int/mental_health/management/depression/en/. 1, 10

[18] D. Acharjee et al. Activity Recognition System Using Inbuilt Sensors of Smart Mobile Phone and Minimizing Feature Vectors. *Microsystem Technologies*, 22(11): 2715–2722, 2016. ISSN 1432-1858. 65

[19] R. Adorni, A. Gatti, A. Brugnera, K. Sakatani, and A. Compare. Could fNIRS Promote Neuroscience Approach in Clinical Psychology? *Frontiers in Psychology*, 7:456, 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.00456. URL http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00456. 4

[20] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 23(2):255–281, Mar 2012. ISSN 1432-1769. doi: 10.1007/s00138-010-0298-4. URL http://dx.doi.org/10.1007/s00138-010-0298-4. 51

[21] S. Alghowinem. *Multimodal Analysis of Verbal and Nonverbal Behaviour on the Example of Clinical Depression*. PhD thesis, The Australian National University, 2015. 15

[22] S. Alghowinem, R. Göcke, M. Wagner, G. Parker, and M. Breakspear. Eye Movement Analysis for Depression Detection. In *International Conference on Image Processing*, pages 4220–4224. IEEE, 2013. ISBN 9781479923410. doi: 10.1109/ICIP.2013.6738869. 4, 21, 22, 23, 28, 29, 31

[23] S. Alghowinem, R. Göcke, M. Wagner, G. Parker, and M. Breakspear. Head Pose and Movement Analysis as an Indicator of Depression. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 283–288.

IEEE, 2013. ISBN 9780769550480. doi: 10.1109/ACII.2013.53. 4, 21, 22, 23, 28, 29, 31, 67

[24] S. Alghowinem, R. Göcke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear. Cross-Cultural Detection of Depression from Nonverbal Behaviour. In *International Conference on Automatic Face and Gesture Recognition*, pages 1–8, Ljubljana, Slovenia, 2015. IEEE. 4, 21, 22, 23, 28, 29, 30, 31, 32, 34, 72, 83, 84, 126

[25] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear. Multimodal Depression Detection:Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2016. ISSN 1949-3045. doi: 10.1109/TAFFC.2016.2634527. 11, 21, 22, 23, 28, 31, 120

[26] T. R. Almaev and M. F. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 356–361. IEEE, 2013. 44, 47

[27] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders - 4th edition (DSM-IV^{TM})*. American Psychiatric Association, 1994. 11

[28] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Publishing, 2013. 1, 2

[29] S. Arbabzadeh-Bouchez, A. Tylee, and J.-P. Lépine. A European perspective on depression in the community: the DEPRES study. *CNS spectrums*, 7(2):120–126, 2002. 6

[30] R. M. Bagby, A. G. Ryder, D. R. Schuller, and M. B. Marshall. The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight? *American Journal of Psychiatry*, 161(12):2163–2177, 2004. 3

[31] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1–10, 2016. 16, 43, 73, 81

[32] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 568–573 vol. 2, June 2005. 23

[33] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri. Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. *Journal of Personality Assessment*, 67(3):588–597, 1996. 72

[34] M. Belahcene, M. Laid, A. Chouchane, A. Ouamane, and S. Bourennane. Local descriptors and tensor local preserving projection in face recognition. In *6th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6, Oct 2016. doi: 10.1109/EUVIP.2016.7764608. 60

[35] Y. S. Ben-Porath. *Handbook of Psychology*, chapter Assessing Personality and Psychopathology With Self-Report Inventories, page 553Ű577. John Wiley & Sons, Inc., 2003. ISBN 9780471264385. 3

[36] S. Bhatia. Multimodal Sensing of Affect Intensity. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 567–571, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4556-9. doi: 10.1145/2993148.2997622. URL `http://doi.acm.org/10.1145/2993148.2997622`. 20, 21

[37] 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*, 2017. 20, 21

[38] J. M. Bland and D. G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies*, 47(8):931 – 936, 2010. ISSN 0020-7489. doi: https://doi.org/10.1016/j.ijnurstu.2009.10.001. URL `http://www.sciencedirect.com/science/article/pii/S0020748909003204`. 105

[39] A. Bobick and J. Davis. Real-time recognition of activity using temporal templates. In *Applications of Computer Vision, 1996. WACV '96., Proceedings 3rd IEEE Workshop on*, pages 39–42, Dec 1996. doi: 10.1109/ACV.1996.571995. 51

[40] H. R. Bogner and J. J. Gallo. Are higher rates of depression inwomen accounted for by differential symptomreporting? *Social Psychiatry and Psychiatric Epidemiology*, 39(2):126–132, Feb 2004. ISSN 1433-9285. doi: 10.1007/s00127-004-0714-z. URL `https://doi.org/10.1007/s00127-004-0714-z`. 90

[41] M. Brown, A. Glendenning, E. A. Hoon, and A. John. Effectiveness of Web-Delivered Acceptance and Commitment Therapy in Relation to Mental Health and Well-Being: A Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, 18(8):e221, Aug 2016. 7

[42] E. Candès. *Ridgelets: Theory and Applications*. PhD thesis, Standord University, Department of Statistics, 1998. 44

[43] E. Candès. Curvelets: A Surprisingly Effective Nonadaptive Representation for Objects with Edges. Technical report, Standford University, Department of Statistics, 2000. 44

[44] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. *Communications on pure and applied mathematics*, 57(2):219–266, 2004. 46

[45] E. Candès, L. Demanet, D. Donoho, and L. Ying. Fast Discrete Curvelet Transforms. *Multiscale Modeling and Simulation*, 5(3):861–899, 2006. ISSN 1540-3459. 45, 46

[46] E. Candes, L. Demanet, D. Donoho, and L. Ying. Fast discrete curvelet transforms. *Multiscale Modeling & Simulation*, 5(3):861–899, 2006. 47, 75

[47] P. Carcagnì, M. Del Coco, P. L. Mazzeo, A. Testa, and C. Distante. Features descriptors for demographic estimation: A comparative study. In C. Distante, S. Battiato, and A. Cavallaro, editors, *Video Analytics for Audience Measurement*, pages 66–85, Cham, 2014. Springer International Publishing. ISBN 978-3-319-12811-5. 60

[48] L. Chao, J. Tao, M. Yang, Y. Li, and J. Tao. Multi task sequence learning for depression scale prediction from video. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 526–531, Sept 2015. doi: 10.1109/ACII.2015.7344620. 21, 25

[49] M. Chmielewski, L. A. Clark, R. M. Bagby, and D. Watson. Method matters: Understanding diagnostic reliability in DSM-IV and DSM-5. *Journal of Abnormal Psychology*, 124(3):764, 2015. 3

[50] N. Clark, T. Herman, J. Halverson, and H. K. Trivedi. *Mental Health Practice in a Digital World: A Clinicians Guide*, chapter Technology Tools Supportive of DSM-5: An Overview, pages 199–211. Springer International Publishing, Cham, 2015. 8

[51] J. A. Coan and J. J. Allen. *Handbook of Emotion Elicitation and Assessment*. Oxford university press, 2007. 12

[52] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968. 27

[53] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De La Torre. Detecting Depression from Facial Actions and Vocal Prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009. ISBN 9781424447992. doi: 10.1109/ACII.2009.5349358. 11, 12, 13, 20, 21, 22, 23, 28, 31, 120, 127

[54] J. S. Comer. Introduction to the special series: Applying new technologies to extend the scope and accessibility of mental health care. *Cognitive and Behavioral Practice*, 22(3):253 – 257, 2015. ISSN 1077-7229. doi: https://doi.org/10.1016/j.cbpra.2015.04.002. 7

[55] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, Aug 2016. 16

[56] R. Cowie, C. Cox, J.-C. Martin, A. Batliner, D. Heylen, and K. Karpouzis. *Emotion-Oriented Systems: The Humaine Handbook*, chapter Issues in Data Labelling, pages 213–241. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-15184-2. doi: 10.1007/978-3-642-15184-2_13. URL `https://doi.org/10.1007/978-3-642-15184-2_13`. 98

[57] R. Cowie, E. Douglas-Cowie, M. McRorie, I. Sneddon, L. Devillers, and N. Amir. *Emotion-Oriented Systems: The Humaine Handbook*, chapter Issues in Data Collection, pages 197–212. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-15184-2. doi: 10.1007/978-3-642-15184-2_12. URL `https://doi.org/10.1007/978-3-642-15184-2_12`. 88

[58] A. Cruz, B. Bhanu, and N. S. Thakoor. Facial emotion recognition with anisotropic inhibited gabor energy histograms. In *2013 IEEE International Conference on Image Processing*, pages 4215–4219, Sept 2013. doi: 10.1109/ICIP.2013.6738868. 53

[59] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Göcke, and J. Epps. Diagnosis of Depression by Behavioural Signals. In *3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13)*, pages 11–20. ACM, 2013. ISBN 9781450323956. doi: 10.1145/2512530.2512535. 20, 21, 25, 36

[60] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. A Review of Depression and Suicide Risk Assessment Using Speech Analysis. *Speech Communication*, 71:10 – 49, July 2015. ISSN 0167-6393. 4, 11, 121

[61] M. Dahmane and J. Meunier. *Continuous Emotion Recognition Using Gabor Energy Filters*, pages 351–358. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-24571-8. doi: 10.1007/978-3-642-24571-8$_$46. URL `http://dx.doi.org/10.1007/978-3-642-24571-8$_$46`. 53

[62] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005. doi: 10.1109/CVPR.2005.177. 60

[63] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, and J. Epps. Investigating Word Affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 27–35, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5502-5. doi: 10.1145/

3133944.3133952. URL http://doi.acm.org/10.1145/3133944.3133952. 20, 23, 25, 37

[64] H. Davies, I. Wolz, J. Leppanen, F. Fernandez-Aranda, U. Schmidt, and K. Tchanturia. Facial expression to emotional stimuli in non-psychotic disorders: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 64:252 – 271, 2016. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2016.02.015. URL http://www.sciencedirect.com/science/article/pii/S0149763415302372. 90

[65] A. Dhall and R. Goecke. A Temporally Piece-wise Fisher Vector Approach for Depression Analysis. In *International Conference on Affective Computing and Intelligent Interaction*, pages 255–259. IEEE, 2015. 21

[66] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion Recognition in the Wild Challenge 2013. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pages 509–516, 2013. 121

[67] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon. Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 461–466, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2885-2. 121

[68] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and Image Based Emotion Recognition Challenges in the Wild: EmotiW 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 423–426, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3912-4. 121

[69] S. Dham, A. Sharma, and A. Dhall. Depression scale recognition from audio, visual and text analysis. *CoRR*, abs/1709.05865, 2017. URL http://arxiv.org/abs/1709.05865. 25, 37

[70] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn. Multimodal Detection of Depression in Clinical Interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 307–310. ACM, 2015. 21, 22, 23, 28, 30

[71] H. Dibeklioğlu, , Z. Hammal, and J. F. Cohn. Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–1, 2017. ISSN 2168-2194. doi: 10.1109/JBHI.2017.2676878. 22, 23, 27, 28, 30

[72] S. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-6(4):325–327, 1976. 68

[73] M. Dyla and H. Tairi. Content Based Images Retrieval Based on PDE Model: use of Curvelet Transform and Gabor Wavelets. *ICGST International Journal on Graphics, Vision and Image Processing*, 11:53–58, 2011. 46

[74] P. Ekman. Are there basic emotions? *Psychological Review*, 1992. 91

[75] P. Ekman. *Handbook of Cognition and Emotion*, chapter Basic Emotions, pages 45–60. John Wiley & Sons, Ltd, 2005. 4

[76] H. Ellgring. *Non-verbal Communication in Depression*. Cambridge University Press, New York, 2007. 4, 50, 55

[77] M. Falagas, K. Vardakas, and P. Vergidis. Under-diagnosis of common chronic diseases: prevalence and impact on human health. *International journal of clinical practice*, 61(9):1569–1579, 2007. 6

[78] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259 – 275, 2003. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/S0031-3203(02)00052-3. URL `http://www.sciencedirect.com/science/article/pii/S0031320302000523`. 53

[79] A. Fazlollahi, F. Meriaudeau, L. Giancardo, V. L. Villemagne, C. C. Rowe, P. Yates, O. Salvado, and P. Bourgeat. Computer-aided detection of cerebral microbleeds in susceptibility-weighted imaging. *Computerized Medical Imaging and Graphics*, 46:269 – 276, 2015. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2015.10.001. URL `http://www.sciencedirect.com/science/article/pii/S0895611115001421`. 125

[80] M. B. First. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders: Patient Edition*. Biometrics Research Department, Columbia University, 2005. 3

[81] K. N. Fountoulakis, M. Papadopoulou, S. Kleanthous, A. Papadopoulou, V. Bizeli, I. Nimatoudis, A. Iacovides, and G. S. Kaprinis. Reliability and psychometric properties of the greek translation of the state-trait anxiety inventory form y: Preliminary data. *Annals of General Psychiatry*, 5(1):2, Jan 2006. ISSN 1744-859X. doi: 10.1186/1744-859X-5-2. URL `https://doi.org/10.1186/1744-859X-5-2`. 91, 101

[82] K. N. Fountoulakis, W. Kawohl, P. N. Theodorakis, A. J. F. M. Kerkhof, A. Navickas, C. Höschl, D. Lecic-Tosevski, E. Sorel, E. Rancans, E. Palova, G. Juckel, G. Isacsson, H. K. Jagodic, I. Botezat-Antonescu, I. Warnke, J. Rybakowski, J. M. Azorin, J. Cookson, J. Waddington, P. Pregelj, K. Demyttenaere, L. G. Hranov, L. I. Stevovic, L. Pezawas, M. Adida, M. L. Figuera, M. Pompili, M. Jakovljević, M. Vichi, G. Perugi, O. Andreassen, O. Vukovic, P. Mavrogiorgou, P. Varnik, P. Bech, P. Dome, P. Winkler, R. K. R. Salokangas, T. From, V. Danileviciute, X. Gonda, Z. Rihmer, J. F. Benhalima, A. Grady, A. K. K. Leadholm, S. Soendergaard, C. Nordt, and J. Lopez-Ibor. Relationship of Suicide Rates to Economic Variables in Europe: 2000-2011. *The British Journal of Psychiatry*, 205(6):486–496, 2014. ISSN 0007-1250. 2

[83] K. Fukunaga and L. Hostetler. K-nearest-neighbor Bayes-risk estimation. *IEEE Transactions on Information Theory*, 21(3):285–293, 1975. 68

[84] S. Ghosh, M. Chatterjee, and L.-P. Morency. A Multimodal Context-based Approach for Distress Assessment. In *16th International Conference on Multimodal Interaction*, pages 240–246. ACM, 2014. ISBN 9781450328852. 4, 11, 21, 22, 23

[85] M. Giannakou, P. Roussi, M. Kosmides, G. Kiosseoglou, A. Adamopoulou, and G. Garyfallos. Adaptation of the beck depression inventory-II to greek population. *Hellenic Journal of Psychology*, 10(2):120–146, 2013. 91, 101

[86] J. M. Girard. CARMA: Software for continuous affect rating and media annotation. *Journal of open research software*, 2(1):e5, 2014. doi: http://doi.org/10.5334/jors.ar. 98

[87] J. M. Girard and J. F. Cohn. Automated Audiovisual Depression Analysis. *Current Opinion in Psychology*, 4:75–79, 2015. 6, 15

[88] J. M. Girard and J. F. Cohn. A primer on observational measurement. *Assessment*, 23:404–413, 2016. 6

[89] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald. Social Risk and Depression: Evidence from Manual and Automatic Facial Expression Analysis. In *10th International Conference on Automatic Face and Gesture Recognition*, pages 1–8. IEEE, 2013. ISBN 978-1-4673-5546-9. doi: 10.1109/FG.2013.6553748. 11, 12, 21, 22, 23, 28, 120

[90] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald. Nonverbal Social Withdrawal in Depression: Evidence from Manual and Automatic Analyses. *Image and Vision Computing*, 32(10):641–647, 2013. ISSN 02628856. doi: 10.1016/j.imavis.2013.12.007. 4, 11, 12, 13, 21, 22, 28, 120

[91] D. Goldberg. *Sadness or Depression? International Perspectives on the Depression Epidemic and Its Meaning*, chapter The Current Status of the Diagnosis of Depression, pages 17–27. Springer Netherlands, Dordrecht, 2016. 6

[92] Y. Gong and C. Poellabauer. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 69–76, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5502-5. doi: 10.1145/3133944.3133945. URL `http://doi.acm.org/10.1145/3133944.3133945`. 21, 22, 25

[93] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherere, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. S. Rizzo, and L.-P. Morency. The Distress Analysis Interview Corpus of Human and Computer Interviews. In *Language Resources and Evaluation Conference*, pages 3123–3128. ELRA, 2014. 4, 5, 11, 73

[94] C. Grigorescu, N. Petkov, and M. A. Westenberg. Contour detection based on nonclassical receptive field inhibition. *IEEE Transactions on Image Processing*, 12 (7):729–739, July 2003. ISSN 1057-7149. doi: 10.1109/TIP.2003.814250. 53

[95] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition and Emotion*, 9(1):87–108, 1995. doi: 10.1080/02699939508408966. URL `https://doi.org/10.1080/02699939508408966`. 91, 93

[96] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan. Multimodal Prediction of Affective Dimensions and Depression in Human-Computer Interactions Categories and Subject Descriptors. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 33–40. ACM, 2014. ISBN 9781450331197. 4, 21, 22, 25

[97] M. Hamilton. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56, 1960. 3

[98] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma. Automated Facial Action Coding System for Dynamic Analysis of Facial Expressions in Neuropsychiatric Disorders. *Journal of Neuroscience Methods*, 200(2):237–256, 2011. ISSN 01650270. doi: 10.1016/j.jneumeth.2011.06.023. 21

[99] S. Harati, A. Crowell, H. Mayberg, J. Kong, and S. Nemati. Discriminating Clinical Phases of Recovery from Major Depressive Disorder Using the Dynamics of Facial Expression. In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2016. 13, 21

[100] L. He, D. Jiang, and H. Sahli. Multimodal Depression Recognition with Dynamic Visual and Audio Cues. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 260–266. IEEE, 2015. 20, 21, 36

[101] J. Heikkila, V. Ojansivu, and E. Rahtu. Improved blur insensitivity for decorrelated local phase quantization. In *20th International Conference on Pattern Recognition*, pages 818–821, Aug 2010. doi: 10.1109/ICPR.2010.206. 59

[102] D. M. Hilty, T. Rabinowitz, R. M. McCarron, D. J. Katzelnick, T. Chang, A. M. Bauer, and J. Fortney. An update on telepsychiatry and how it can leverage collaborative, stepped, and integrated services to primary care. *Psychosomatics*, 59(3):227 – 250, 2018. ISSN 0033-3182. doi: https://doi.org/10.1016/j.psym.2017.12.005. URL `http://www.sciencedirect.com/science/article/pii/S0033318217302670`. 7

[103] C. Hollis, R. Morriss, J. Martin, S. Amani, R. Cotton, M. Denis, and S. Lewis. Technological innovations in mental healthcare: harnessing the digital revolution. *The British Journal of Psychiatry*, 206(4):263–265, 2015. 6

[104] B. Hosseinifard, M. H. Moradi, and R. Rostami. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Computer methods and programs in biomedicine*, 109(3):339–345, 2013. 4

[105] Z. Huang, B. Stasak, T. Dang, K. Wataraka Gamage, P. Le, V. Sethu, and J. Epps. Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, pages 19–26, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4516-3. doi: 10.1145/2988257.2988265. URL `http://doi.acm.org/10.1145/2988257.2988265`. 25

[106] M. Islam, D. Zhang, and G. Lu. Rotation invariant curvelet features for texture image retrieval. In *Proc. of IEEE International Conference on Multimedia and Expo*, pages 562–565. IEEE, 2009. 46

[107] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux. Depression Estimation Using Audiovisual Features and Fisher Vector Encoding. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 87–91. ACM, 2014. ISBN 9781450331197. doi: 10.1145/2661806.2661817. 20, 21, 23, 36

[108] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh. Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 73–80. ACM, 2014. ISBN 9781450331197. doi: 10.1145/2661806.2661812. 20, 21, 25, 36, 51

[109] A. Jan, H. Meng, Y. F. A. Gaus, and F. Zhang. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–1, 2017. ISSN 2379-8920. doi: 10.1109/TCDS.2017.2721552. 21, 25, 36

[110] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3D Face Alignment From 2D Videos in Real-time. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8, May 2015. doi: 10.1109/FG.2015.7163142. 16

[111] N. P. Jones, G. J. Siegle, and D. Mandell. Motivational and Emotional Influences on Cognitive Control in Depression: A Pupillometry Study. *Cognitive, Affective, & Behavioral Neuroscience*, 15(2):263–275, 2014. ISSN 1530-7026. doi: 10.3758/s13415-014-0323-6. 12, 21

[112] K. Josephine, L. Josefine, D. Philipp, E. David, and B. Harald. Internet-and mobile-based depression interventions for people with diagnosed depression: a systematic review and meta-analysis. *Journal of affective disorders*, 223:28–40, 2017. 7

[113] J. Joshi. Depression analysis: a multimodal approach. In *14th ACM International Conference on Multimodal Interaction*, pages 321–324, Santa Monica, California, 2012. ACM. ISBN 9781450314671. doi: 10.1145/2388676.2388747. 12

[114] J. Joshi. An Automated Framework for Depression Analysis. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 630–635. IEEE, 2013. 5, 13, 20, 21, 23, 28, 30, 31

[115] J. Joshi, A. Dhall, R. Göcke, M. Breakspear, and G. Parker. Neural-net Classification for Spatio-temporal Descriptor Based Depression Analysis. In *21st International Conference on Pattern Recognition*, pages 2634–2638. IEEE, 2012. ISBN 9784990644109. 20, 21, 23, 28

[116] J. Joshi, A. Dhall, R. Göcke, and J. F. Cohn. Relative Body Parts Movement for Automatic Depression Analysis. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 492–497. IEEE, 2013. ISBN 9780769550480. doi: 10.1109/ACII.2013.87. 5, 11, 13, 21, 23, 28, 30

[117] J. Joshi, R. Göcke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear. Multimodal Assistive Technologies for Depression Diagnosis and Monitoring. *Journal on Multimodal User Interfaces*, 7(3):217–228, 2013. ISSN 17837677. doi: 10.1007/s12193-013-0123-2. 20, 21, 22, 23, 28, 31

[118] J. Joshi, R. Göcke, G. Parker, and M. Breakspear. Can Body Expressions Contribute to Automatic Depression Analysis? In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–7. IEEE, 2013. ISBN 9781467355452. doi: 10.1109/FG.2013.6553796. 4, 11, 13, 20, 21, 23, 28, 31

[119] A. Kacem, Z. Hammal, M. Daoudi, and J. Cohn. Detecting depression severity by interpretable representations of motion dynamics. In *13th IEEE Conference on Automatic Face and Gesture Recognition*, Xi'an, China, May 2018. 20, 22, 23, 27, 30

[120] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker. Fusion of Audio-visual Features using Hierarchical Classifier Systems for the Recognition of Affective States and the State of Depression. In *3rd International Conference on Pattern Recognition Applications and Methods*, pages 671–678. SciTePress, 2014. 21, 25, 36

[121] M. Kächele, M. Schels, and F. Schwenker. Inferring Depression and Affect from Application Dependent Meta Knowledge. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 41–48. ACM, 2014. ISBN 9781450331197. 20, 21, 25

[122] M. Kächele, M. Schels, and F. Schwenker. The influence of annotation, corpus design, and evaluation on the outcome of automatic classification of human emotions. *Frontiers in ICT*, 3:27, 2016. ISSN 2297-198X. doi: 10.3389/fict.2016.00027. URL `https://www.frontiersin.org/article/10.3389/fict.2016.00027`. 87, 90, 98

[123] Y. Kang, X. Jiang, Y. Yin, Y. Shang, and X. Zhou. Deep transformation learning for depression diagnosis from facial images. In J. Zhou, Y. Wang, Z. Sun, Y. Xu, L. Shen, J. Feng, S. Shan, Y. Qiao, Z. Guo, and S. Yu, editors, *Biometric Recognition*, pages 13–22, Cham, 2017. Springer International Publishing. ISBN 978-3-319-69923-3. 21, 25, 36

[124] S. Kapur, A. G. Phillips, and T. R. Insel. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, 17(12):1174–1179, 2012. 3

[125] H. Kaya and A. A. Salah. Eyes Whisper Depression: A CCA based Multimodal Approach. In *International Conference on Multimedia*, pages 961–964. ACM, 2014. ISBN 9781450330633. doi: 10.1145/2647868.2654978. 21, 25, 36

[126] H. Kaya, F. Çilli, and A. A. Salah. Ensemble CCA for Continuous Emotion Prediction. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 19–26. ACM, 2014. ISBN 9781450331197. 20, 21, 25, 36

[127] R. C. Kessler. The Costs of Depression. *The Psychiatric Clinics of North America*, 35(1):1–14, 2012. 2

[128] B. Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004. 9

[129] H. C. Kraemer, D. J. Kupfer, D. E. Clarke, W. E. Narrow, and D. A. Regier. DSM-5: How Reliable Is Reliable Enough? *American Journal of Psychiatry*, 169 (1):13–15, 2012. 3

[130] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc. URL `http://dl.acm.org/citation.cfm?id=2999134.2999257`. 62, 64

[131] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015. doi: 10.1038/nature14539. 61

[132] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau. Computer-Aided Detection and Diagnosis for Prostate Cancer Based on Mono and Multi-Parametric MRI: A Rreview . *Computers in Biology and Medicine*, 60:8 – 31, 2015. ISSN 0010-4825. doi: http://doi.org/10.1016/j.compbiomed.2015.02.009. 20

[133] X. Li, T. Cao, S. Sun, B. Hu, and M. Ratcliffe. Classification study on eye movement data: Towards a new approach in depression detection. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1227–1232, July 2016. doi: 10.1109/CEC.2016.7743927. 21, 22, 23

[134] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (CERT). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011. 16

[135] G. M. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou. Towards an affective interface for assessment of psychological distress. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 539–545. IEEE, 2015. 4

[136] X. Ma, D. Huang, Y. Wang, and Y. Wang. Cost-sensitive two-stage depression prediction using dynamic visual clues. In S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, editors, *Computer Vision – ACCV 2016*, pages 338–351, Cham, 2017. Springer International Publishing. ISBN 978-3-319-54184-6. 21, 23, 25, 36

[137] N. C. Maddage, R. Senaratne, L.-S. A. Low, M. Lech, and N. Allen. Video-based Detection of the Clinical Depression in Adolescents. In *31st International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3723–3726, Minneapolis, Minnesota, 2009. IEEE. ISBN 9781424432967. doi: 10.1109/IEMBS.2009.5334815. 21, 23, 29, 34

[138] M. Maj. *Sadness or Depression? International Perspectives on the Depression Epidemic and Its Meaning*, chapter The Continuum of Depressive States in the Population and the Differential Diagnosis Between "Normal"' Sadness and Clinical Depression, pages 29–38. Springer Netherlands, Dordrecht, 2016. ISBN 978-94-017-7423-9. 6

[139] J. J. Maller, S. S. George, R. P. Viswanathan, P. B. Fitzgerald, and P. Junor. Using thermographic cameras to investigate eye temperature and clinical severity in depression. *Journal of biomedical optics*, 21(2):026001, 2016. 13, 21

[140] T. Mandal and Q. Wu. Face recognition using curvelet based PCA. In *Proc. of 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008. 46

[141] A. Maridaki, A. Pampouchidou, K. Marias, and M. Tsiknakis. Machine learning techniques for automatic depression assessment. In *41st International Conference on Telecommunications and Signal Processing*, Athens, Greece, July 2018. 20, 21, 23

[142] K. E. Markon, M. Chmielewski, and C. J. Miller. The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. *Psychological bulletin*, 137(5):856, 2011. 6

[143] B. Martinez and M. F. Valstar. *Advances in Face Detection and Facial Image Analysis*, chapter Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition, pages 63–100. Springer International Publishing, Cham, 2016. 121

[144] P. Martínez, G. Rojas, V. Martínez, M. A. Lara, and J. C. Pérez. Internet-based interventions for the prevention and treatment of depression in people living in developing countries: A systematic review. *Journal of affective disorders*, 234: 193–200, 2018. 7

[145] G. McIntyre, R. Göcke, M. Hyett, M. Green, and M. Breakspear. An Approach for Automatically Measuring Facial Activity in Depressed Subjects. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8. IEEE, 2009. ISBN 9781424447992. doi: 10.1109/ACII.2009.5349593. 12, 21, 28, 127

[146] G. Mcintyre, R. Göcke, M. Breakspear, and G. Parker. Facial Response to Video Content in Depression. In *13th International Conference on Multimodal Interaction. Workshop: Inferring Cognitive and Emotional States from Multimodal Measures*. ACM, 2011. 28

[147] G. J. McIntyre. *The Computer Analysis of Facial Expressions: On the Example of Depression and Anxiety*. PhD thesis, The Australian National University, Canberra, Australia, 2010. 12, 21, 23, 28

[148] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang. Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression. In *3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13)*, pages 21–30. ACM, 2013. ISBN 9781450323956. doi: 10.1145/2512530.2512532. 21, 25, 36, 51

[149] A. J. Mitchell, J. C. Coyne, et al. *Screening for depression in clinical practice: an evidence-based guide*. Oxford University Press, 2009. 3

[150] A. J. Mitchell, A. Vaze, and S. Rao. Clinical Diagnosis of Depression in Primary Care: a Meta-analysis. *The Lancet*, 374(9690):609 – 619, 2009. ISSN 0140-6736. 6

[151] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou. Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, pages 43–50, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4516-3. doi: 10.1145/2988257.2988261. URL `http://doi.acm.org/10.1145/2988257.2988261`. 21, 22, 32

[152] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition, Elsevier*, 29(1):51–59, 1996. ISSN 0031-3203. 57

[153] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 57, 58, 73, 75

[154] V. Ojansivu, E. Rahtu, and J. Heikkila. Rotation invariant local phase quantization for blur insensitive texture analysis. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008. doi: 10.1109/ICPR.2008.4761377. 59

[155] J. Olesen, A. Gustavsson, M. Svensson, H.-U. Wittchen, B. Jönsson, on behalf of the CDBE2010 study group, and the European Brain Council. The Economic Cost of Brain Disorders in Europe. *European Journal of Neurology*, 19(1):155–162, 2012. 2

[156] B. Ooi Kuan Ee. *Early Prediction of Clinical Depression in Adolescents using Single-Chanel and Multichannel Classification Approach.* PhD thesis, RMIT University, 2014. 13, 21, 23, 34

[157] B. Ooi Kuan Ee, L.-s. A. Low, M. Lech, and N. Allen. Prediction of Clinical Depression in Adolescents Using Facial Image Analaysis. In *12th International Workshop on Image Analysis for Multimedia Interactive Services*, Delft, The Netherlands, 2011. 12, 13, 21, 23, 29, 34

[158] A. Pampouchidou. Facial-gesture based interaction with expressive virtual characters. Master's thesis, University of Burgundy, 2011. 44

[159] A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang, and M. Fabrice. Designing a Framework for Assisting Depression Severity Assessment from Facial Image Analysis. In *IEEE International Conference on Signal and Image Processing Applications*, pages 578–583, 2015. 21, 23, 40, 41, 47, 73, 85, 129

[160] A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang, G. Lemaître, and F. Meriaudeau. Video-Based Depression Detection Using Local Curvelet Binary Patterns in Pairwise Orthogonal Planes. In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2016. 21, 22, 23, 32, 41, 47, 75, 84, 85, 129

[161] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias, F. Yang, and M. Tsiknakis. Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, pages 27–34, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4516-3. doi: 10.1145/2988257.2988266. URL `http://doi.acm.org/10.1145/2988257.2988266`. 21, 22, 23, 32, 43, 51, 70, 78, 85, 124, 125, 129

[162] A. Pampouchidou, M. Pediaditis, A. Maridaki, M. Awais, C.-M. Vazakopoulou, S. Sfakianakis, M. Tsiknakis, P. Simos, K. Marias, F. Yang, and F. Meriaudeau. Quantitative comparison of motion history image variants for video-based depression assessment. *EURASIP Journal on Image and Video Processing*, 2017

(1):64, Sep 2017. ISSN 1687-5281. doi: 10.1186/s13640-017-0212-3. URL https://doi.org/10.1186/s13640-017-0212-3. 20, 21, 23, 27, 29, 32, 43, 51, 70, 80, 85, 129

[163] A. Pampouchidou, O. Simantiraki, C. M. Vazakopoulou, C. Chatzaki, M. Pediaditis, A. Maridaki, K. Marias, P. Simos, F. Yang, F. Meriaudeau, and M. Tsiknakis. Facial geometry and speech analysis for depression detection. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1433–1436, July 2017. doi: 10.1109/EMBC.2017.8037103. 20, 21, 23, 43, 80, 84, 85, 124, 129

[164] A. Pampouchidou, O. Simantiraki, C.-M. Vazakopoulou, K. Marias, P. Simos, F. Yang, F. Meriaudeau, and M. Tsiknakis. Détection de la dépression par lŠanalyse de la géométrie faciale et de la parole. In *GRETSI 2017*, volume XXVIème colloque, Juan-Les-Pins, France, September 2017. 21, 43, 124

[165] A. Pampouchidou, P. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, and M. Tsiknakis. Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*, 2017. 1, 5, 9, 124, 129

[166] M. Pediaditis. *Computerised Analysis of the Clinical Image of Absence Seizures*. PhD thesis, Doctoral School Biomedical Engineering, Technische Universität Graz, Graz, November 2014. 65

[167] H. Pérez Espinoza, H. J. Escalante, L. Villaseñor Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyes-Meza. Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 49–55. ACM, 2014. 21, 22, 51

[168] P. Pichot. *New Results in Depression Research*, chapter Self-report inventories in the study of depression, pages 53–58. Springer, 1986. 3

[169] R. B. Price, D. Rosen, G. J. Siegle, C. D. Ladouceur, K. Tang, K. B. Allen, N. D. Ryan, R. E. Dahl, E. E. Forbes, and J. S. Silk. From Anxious Youth to

Depressed Adolescents: Prospective Prediction of 2-Year Depression Symptoms via Attentional Bias Measures. *Journal of Abnormal Psychology*, 125(2):267–278, 2015. 5

[170] R. Ptucha and A. Savakis. Towards the Usage of Optical Flow Temporal Features for Facial Expression Classification. In *International Symposium on Visual Computing*. Springer, 2012. 52

[171] Y. Ren, H. Yang, C. Browning, S. Thomas, and M. Liu. Performance of Screening Tools in Detecting Major Depressive Disorder among Patients with Coronary Heart Disease: A Systematic Review. *Medical Science Monitor*, 21:646 – 653, 2015. ISSN 1643-3750. doi: http://dx.doi.org/10.12659/MSM.892537. URL `www.medscimonit.com/abstract/index/idArt/892537`. 3

[172] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 3–9, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5502-5. doi: 10.1145/3133944.3133953. URL `http://doi.acm.org/10.1145/3133944.3133953`. 25, 37

[173] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3): 211–252, Dec. 2015. ISSN 0920-5691. doi: 10.1007/s11263-015-0816-y. URL `http://dx.doi.org/10.1007/s11263-015-0816-y`. 62

[174] M. A. H. M. S'adan, A. Pampouchidou, and F. Meriaudeau. Deep learning techniques for depression assessment. In *7th International Conference on Intelligent and Advanced System*, Kuala Lumpur, Malaysia, August 2018. 21, 23

[175] J. M. Saragih, S. Lucey, and J. F. Cohn. Face Alignment Through Subspace Constrained Mean-shifts. In *IEEE 12th International Conference on Computer Vision*, pages 1034–1041. IEEE, 2009. 18, 19

[176] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency. Automatic Behavior Descriptors for Psychological Disorder Analysis. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–8. IEEE, 2013. 4, 11, 12, 21, 23, 27, 31, 32

[177] S. Scherer, G. Stratou, and L.-P. Morency. Audiovisual Behavior Descriptors for Depression Assessment. In *15th International Conference on Multimodal Interaction*, pages 135–140. ACM, 2013. ISBN 9781450321297. doi: 10.1145/2522848.2522886. 4, 11, 12, 21, 29

[178] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency. Automatic Audiovisual Behavior Descriptors for Psychological Disorder Analysis. *Image and Vision Computing*, 32(10):648–658, 2014. ISSN 02628856. doi: 10.1016/j.imavis.2014.06.001. 4, 12, 21, 22, 120

[179] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117, 2015. ISSN 0893-6080. 18

[180] B. W. Schuller. *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, chapter Acquisition of Affect, pages 57–80. Springer International Publishing, Cham, 2016. 121

[181] I. Schumann, A. Schneider, C. Kantert, B. Löwe, and K. Linde. Physicians' attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: a systematic review of qualitative studies. *Family Practice*, 29(3):255–263, 2011. 6

[182] M. Senoussaoui, M. Sarria-Paja, J. a. F. Santos, and T. H. Falk. Model Fusion for Multimodal Depression Classification and Level Detection. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 57–63. ACM, 2014. ISBN 9781450331197. 20, 21, 23, 25, 29, 32, 36, 72, 83, 84

[183] S. Sfakianakis, E. S. Bei, and M. Zervakis. Stacking of network based classifiers with application in breast cancer classification. In *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, pages 1079–1084. Springer, 2016. 125

[184] M. Sidorov and W. Minker. Emotion Recognition and Depression Diagnosis by Acoustic and Visual Features : A Multimodal Approach. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 81–86. ACM, 2014. ISBN 9781450331197. doi: 10.1145/2661806.2661816. 21, 23, 37

[185] G. J. Siegle, S. R. Steinhauer, E. S. Friedman, W. S. Thompson, and M. E. Thase. Remission Prognosis for Cognitive Therapy for Recurrent Depression Using the Pupil: Utility and Neural Correlates. *Biological Psychiatry*, 69(8):726–733, 2011. ISSN 00063223. doi: 10.1016/j.biopsych.2010.12.041. 5, 11, 12

[186] J. S. Silk, R. E. Dahl, N. D. Ryan, E. E. Forbes, D. A. Axelson, B. Birmaher, and G. J. Siegle. Pupillary Reactivity to Emotional Information in Child and Adolescent Depression: Links to Clinical and Ecological Measures. *American Journal of Psychiatry*, 164(12):1873–1880, 2007. 11, 12, 21

[187] O. Simantiraki, P. Charonyktakis, A. Pampouchidou, M. Tsiknakis, and M. Cooke. Glottal source features for automatic speech-based depression assessment. In *INTERSPEECH*, To Appear - 2017. 124

[188] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 62

[189] C. Smailis, N. Sarafianos, T. Giannakopoulos, and S. Perantonis. Fusing Active Orientation Models and Mid-term Audio Features for Automatic Depression Estimation. In *9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 2016. 20, 21, 36

[190] P. Sobocki, B. Jönsson, J. Angst, and C. Rehnberg. Cost of depression in europe. *The journal of mental health policy and economics*, 9(2):87–98, 2006. 2

[191] S. Song, L. Shen, and M. Valstar. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *13th IEEE International Conference on Face and Gesture Recognition*, Xi'an, China, 2018. 21, 25, 37

[192] E. A. Stepanov, S. Lathuilière, S. A. Chowdhury, A. Ghosh, R. Vieriu, N. Sebe, and G. Riccardi. Depression severity estimation from multiple modalities. *CoRR*, abs/1711.06095, 2017. URL `http://arxiv.org/abs/1711.06095`. 25, 37

[193] E. Stockings, L. Degenhardt, Y. Y. Lee, C. Mihalopoulos, A. Liu, M. Hobbs, and G. Patton. Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of Affective Disorders*, 174:447 – 463, 2015. ISSN 0165-0327. doi: http://dx.doi.org/10.1016/j.jad.2014.11.061. URL `http://www.sciencedirect.com/science/article/pii/S016503271400785X`. 3

[194] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 147–152. IEEE, 2013. ISBN 9780769550480. doi: 10.1109/ACII.2013.31. 11, 12, 21, 22, 23, 29, 67, 125

[195] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. Automatic Nonverbal Behavior Indicators of Depression and PTSD: the Effect of Gender. *Journal of Multimodal User Interfaces*, 9(1):17–29, 2014. ISSN 17837677. doi: 10.1109/ACII.2013.31. 4, 11, 12, 21, 22, 23, 120, 125

[196] V. Štruc and N. Pavešić. Gabor-based kernel partial-least-squares discrimination features for face recognition. *Informatica*, 20(1):115–138, 2009. 40

[197] V. Štruc and N. Pavešic. Photometric normalization techniques for illumination invariance. *Advances in Face Image Analysis: Techniques and Technologies*, pages 279–300, 2011. 40

[198] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang. A random forest regression method with selected-text feature for depression assessment. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 61–68, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5502-5. doi: 10.1145/3133944.3133951. URL `http://doi.acm.org/10.1145/3133944.3133951`. 21, 25, 37

[199] T. Suto, M. Fukuda, M. Ito, T. Uehara, and M. Mikuni. Multichannel Near-infrared Spectroscopy in Depression and Schizophrenia: Cognitive Brain Activation Study. *Biological Psychiatry*, 55(5):501 – 511, 2004. ISSN 0006-3223. 4

[200] Z. S. Syed, K. Sidorov, and D. Marshall. Depression severity prediction based on biomarkers of psychomotor retardation. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 37–43, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5502-5. doi: 10.1145/3133944.3133947. URL `http://doi.acm.org/10.1145/3133944.3133947`. 21, 25

[201] M. Tanaka. *Face Parts Detection*, 2015. Available at: http://like.silk.to/matlab/detectFaceParts.html. 41, 43

[202] R. Thomas-MacLean, J. Stoppard, B. B. Miedema, and S. Tatemichi. Diagnosing Depression: There is no Blood Test. *Canadian Family Physician*, 51(8):1102–3, 2005. 3, 6

[203] Y.-L. Tian, T. Kanade, and J. F. Cohn. *Handbook of Face Recognition*, chapter Facial Expression Analysis, pages 247–275. Springer New York, New York, NY, 2005. ISBN 978-0-387-27257-3. 53

[204] C. Tomasi and T. Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991. 75

[205] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858. IEEE, 2014. 41

[206] M. Valstar. Automatic Behaviour Understanding in Medicine. In *2014 Workshop on Roadmapping the Future of Multimodal Interaction Research*, pages 57–60, Istabul, Turkey, 2014. ACM. ISBN 9781450306157. 8

[207] M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection in video. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, volume 1, pages 635–640 vol.1, Oct 2004. doi: 10.1109/ICSMC.2004.1398371. 51

[208] M. Valstar, B. Schuller, J. Krajewski, R. Cowie, and M. Pantic. Workshop Summary for the 3rd International Audio / Visual Emotion Challenge and Workshop (AVEC'13). In *21st ACM International Conference on Multimedia*, pages 1085–1086. ACM, 2013. ISBN 9781450324045. 7

[209] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge. In *3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13)*, pages 3–10. ACM, 2013. 7, 11, 36

[210] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 3–10. ACM, 2014. 7, 11, 36, 72

[211] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, R. Cowie, and M. Pantic. Summary for AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pages 1483–1484, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3603-1. doi: 10.1145/2964284.2980532. URL `http://doi.acm.org/10.1145/2964284.2980532`. 7

[212] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, pages 3–10, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4516-3. doi: 10.1145/2988257.2988258. URL `http://doi.acm.org/10.1145/2988257.2988258`. 25, 32, 37, 73, 80

[213] P. van de Ven. User-friendly ICT Tools to Enhance Self-management and Effective Treatment of Depression in the EU. Technical report, Science Engineering, Limerick, Ireland, 2010. 2

[214] C.-M. Vazakopoulou, A. Pampouchidou, F. Yang, F. Meriaudeau, K. Marias, and M. Tsiknakis. Détection de la dépression par l'analyse de la géométrie faciale et apprentissage automatique. In *CNRIUT*, Aix-en-Provence, France, June 2018. 21

[215] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 16, 40

[216] V. Vonikakis et al. Group happiness assessment using geometric features and dataset balancing. In *18th ACM-ICMI*, pages 479–486. ACM, 2016. 55

[217] T. Vos, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, R. S. Abdulkader, A. M. Abdulle, T. A. Abebo, S. F. Abera, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100):1211–1259, 2017. 1

[218] J. C. Wakefield and S. Demazeux. *Sadness or Depression? International Perspectives on the Depression Epidemic and Its Meaning*, chapter Introduction: Depression, One and Many, pages 1–15. Springer Netherlands, Dordrecht, 2016. 6

[219] J. Wang, Y. Fan, X. Zhao, and N. Chen. Pupillometry in Chinese Female Patients with Depression: A Pilot Study. *International Journal of Environmental Research and Public Health*, 11(2):2236–2243, 2014. ISSN 16617827. doi: 10.3390/ijerph110202236. 11, 12, 21

[220] L. Wang and D. He. Texture classification using texture spectrum. *Pattern Recognition, Elsevier*, 23(8):905–910, 1990. ISSN 0031-3203. 57

[221] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, and R. Verma. Automated Video-based Facial Expression Analysis of Neuropsychiatric Disorders. *Journal of Neuroscience Methods*, 168(1):224–238, 2008. ISSN 01650270. doi: 10.1016/j.jneumeth.2007.09.030. 20

[222] P. H. Waxer. Therapist Training in Nonverbal Communication. I: Nonverbal Cues for Depression. *Journal of Clinical Psychology*, 30(2):215, 1974. 4, 50

[223] L. Wen, X. Li, G. Guo, and Y. Zhu. Automated Depression Diagnosis Based on Facial Dynamic Analysis and Sparse Coding. *IEEE Transactions on Information Forensics and Security*, 10(7):1432–1441, 2015. ISSN 1556-6013. doi: 10.1109/TIFS.2015.2414392. 21, 25

[224] L. S. Williams, E. J. Brizendine, L. Plue, T. Bakas, W. Tu, H. Hendrie, and K. Kroenke. Performance of the PHQ-9 as a Screening Tool for Depression After Stroke. *Stroke*, 36(3):635–638, 2005. 3

[225] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta. Vocal and Facial Biomarkers of Depression Based on Motor Incoordination and Timing. In *4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 65–72. ACM, 2014. 20, 21, 25

[226] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri. Detecting Depression Using Vocal, Facial and Semantic Communication Cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, pages 11–18, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4516-3. doi: 10.1145/2988257.2988263. URL `http://doi.acm.org/10.1145/2988257.2988263`. 25, 37

[227] C. Winograd-Gurvich, N. Georgiou-Karistianis, P. Fitzgerald, L. Millist, and O. White. Ocular motor differences between melancholic and non-melancholic depression. *Journal of affective disorders*, 93(1):193–203, 2006. 21

[228] C. Winograd-Gurvich, N. Georgiou-Karistianis, P. B. Fitzgerald, L. Millist, and O. B. White. Self-paced and reprogrammed saccades: differences between melancholic and non-melancholic depression. *Neuroscience research*, 56(3):253–260, 2006. 21

[229] World Health Organization. Depression and Other Common Mental Disorders. Technical report, Global Health Estimates, 2017. 2

[230] X. Xiong and F. de la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 41

[231] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli. Decision Tree Based Depression Classification from Audio Video and Language Information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, pages 89–96, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4516-3. doi: 10.1145/2988257.2988269. URL `http://doi.acm.org/10.1145/2988257.2988269`. 21, 23, 25, 31

[232] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 53–59, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5502-5. doi: 10.1145/3133944.3133948. URL `http://doi.acm.org/10.1145/3133944.3133948`. 21, 25

[233] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. Jiang. Hybrid depression classification and estimation from audio video and text information. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 45–51, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5502-5. doi: 10.1145/3133944.3133950. URL `http://doi.acm.org/10.1145/3133944.3133950`. 21, 25

[234] T.-H. Yang, C.-H. Wu, K.-Y. Huang, and M.-H. Su. Coupled HMM-based Multimodal Fusion for Mood Disorder Detection Through Elicited Audio–Visual Signals. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2016. doi: 10.1007/s12652-016-0395-y. 13, 21, 23, 30, 34, 125

[235] J. Zhang, Z. Zhang, W. Huang, Y. Lu, and Y. Wang. Face recognition based on curvefaces. In *Proc. of 3rd International Conference on Natural Computation-Volume 02*, pages 627–631. IEEE Computer Society, 2007. 46

[236] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, June 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1110. 43, 44

[237] D. Zhou, J. Luo, V. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. Kautz. Tackling Mental Health by Integrating Unobtrusive Multimodal Sensing. In *29th AAAI Conference on Artificial Intelligence*, pages 1401–1408. AAAI, 2015. 4, 21, 23, 29, 34

[238] X. Zhou, K. Jin, Y. Shang, and G. Guo. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*, pages 1–1, 2018. ISSN 1949-3045. doi: 10.1109/TAFFC.2018.2828819. 21, 35, 36

[239] Y. Zhou, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell. Multimodal Prediction of Psychological Disorders: Learning Verbal and Nonverbal Commonalities in Adjacency Pairs. In *17th Workshop on the Semantics and Pragmatics of Dialogue*, pages 160–169. SEMDIAL, 2013. 4, 12, 21, 22, 23, 31

[240] Y. Zhu, Y. Shang, Z. Shao, and G. Guo. Automated Depression Diagnosis based on Deep Networks to Encode Facial Appearance and Dynamics. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2017. ISSN 1949-3045. doi: 10.1109/TAFFC.2017.2650899. 21, 36