NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER THESIS

Multimedia Content Analysis For The Assessment of Depression Level

Christos Smailis

Supervisor: **Perantonis Stavros,** Research Director at NCSR-Demokritos

ATHENS

AUGUST 2015

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Πολυμεσικού Περιεχομένου για την Εκτίμηση Επιπέδου Κατάθλιψης

Χρήστος Σμαΐλης

Επιβλέποντες: **Περαντώνης Σταύρος,** Διευθυντής Ερευνών ΕΚΕΦΕ-Δημόκριτος

ΑΘΗΝΑ

ΑΥΓΟΥΣΤΟΣ 2015

MASTER THESIS

Multimedia Content Analysis For The Assessment of Depression Level

Christos V. Smailis

Registry Number: ΠΙΒ0107

Supervisor: **Perantonis Stavros,** Research Director, NCSR-Demokritos

EXAMINATION BOARD: **Perantonis Stavros**, Research Director, NCSR “Demokritos"

Giannakopoulos Theodore, Research Associate, NCSR-Demokritos

Spyrou George, Staff Research Scientist - Professor Level, Biomedical Research Foundation of the Academy of Athens

August 2015

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση Πολυμεσικού Περιεχομένου για την Εκτίμηση Επιπέδου Κατάθλιψης

Χρήστος Β. Σμαΐλης
Α.Μ.: ΠΙΒ0107

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Περαντώνης Σταύρος,** Διευθυντής Ερευνών ΕΚΕΦΕ-Δημόκριτος

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Περαντώνης Σταύρος,** Διευθυντής Ερευνών, ΕΚΕΦΕ-Δημόκριτος

Γιαννακόπουλος Θεόδωρος, Επιστημονικός Συνεργάτης, ΕΚΕΦΕ-Δημόκριτος

Σπύρου Γεώργιος, Ειδικός Λειτουργικός Επιστήμονας (βαθμίδα Α΄), Ίδρυμα Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών

Αύγουστος 2015

ABSTRACT

A considerable portion of the world population is affected by depression. According to numbers from the World Health Organization, until 2012 there were at least 350 million people living with some form of depression. Mental illness often affects people of working age, a fact that has strong social implications causing significant losses and burdens to the economic system, as well as the social, educational, and justice systems. For the assessment of Depression level, psychologists and psychiatrists base their evaluation of a patient's condition on questionnaires as well as expressive facial and vocal cues.

Within the context of this thesis we use regression to predict an individual's depression level in the BDI-II scale, utilizing features of his/hers face and voice , using machine learning techniques.

In particular, the videos we used as input data stem from the Depression Recognition Sub-Challenge of the 2014 Audio-Visual Emotion Challenge and Workshop (AVEC 2014). The estimation of depression level was performed, using Support Vector Regressors (SVR) models, based on features extracted from AAMs (Active Appearance Models) as well as a variety of features from the audio modality. Results indicated that the fusion of AAMs and Audio features leads to better performance compared to individual modalities. What is more the performance of the system remained stable under subject-independent random sampling validation. Additionally the performance was improved by 6% by fusing AAM and LGBP-TOP Features compared to the performance of Baseline LGBP-TOP features of AVEC 2014.

SUBJECT AREA: Multimedia Content Analysis

KEYWORDS: Depression, Video Analysis, Audio Analysis

ΠΕΡΙΛΗΨΗ

Ένα σημαντικό ποσοστό του παγκόσμιου πληθυσμού, επηρεάζεται από τη κατάθλιψη. Σύμφωνα με στοιχεία του Παγκόσμιου Οργανισμού Υγείας ως το 2012 υπήρχαν τουλάχιστον 350 εκατομμύρια άνθρωποι που υπέφεραν από κάποια μορφή κατάθλιψης. Η κακή ψυχική υγεία συχνά επηρεάζει άτομα σε ηλικία εργασίας, γεγονός που έχει έντονες κοινωνικές επιπτώσεις. Για την αξιολόγηση του επιπέδου κατάθλιψης ενός ασθενούς οι ψυχολόγοι και οι ψυχίατροι χρησιμοποιούν ερωτηματολόγια (π.χ Beck Depression Index-II - BDI-II) καθώς και στοιχεία των εκφράσεων του προσώπου και της φωνής.

Στα πλαίσια αυτής της διπλωματικής η εκτίμηση του επιπέδου κατάθλιψης αντιμετωπίστηκε ως πρόβλημα πρόβλεψης του βαθμού κατάθλιψης ενός ατόμου στη κλίμακα BDI-II χρησιμοποιώντας τεχνικές μηχανικής μάθησης για την αναγνώριση χαρακτηριστικών του προσώπου και της φωνής.

Πιο συγκεκριμένα, τα video που χρησιμοποιήθηκαν ως δεδομένα εισόδου στο πλαίσιο αυτής της εργασίας προήλθαν από το Depression Recognition Sub-Challenge του 2014 Audio-Visual Emotion Challenge and Workshop (AVEC 2014). Η εκτίμηση του επιπέδου κατάθλιψης έγινε με χρήση μοντέλων παλινδρόμησης SVR (Support Vector Regressors) βασισμένων σε χαρακτηριστικά εικόνας που εξήχθησαν με χρήση του αλγορίθμου AAM(Active Appearance Models), καθώς και ποικιλίας χαρακτηριστικών ήχου. Τα αποτελέσματα έδειξαν πως ο συνδυασμός των χαρακτηριστικών AAMs και ήχου, οδήγησε σε βελτίωση της ακρίβειας της διάγνωσης της κατάθλιψης συγκριτικά με τις περιπτώσεις όπου χρησιμοποιείτε μόνο μία πηγή πληροφορίας. Επιπλέον η απόδοση του συστήματος παρέμεινε σταθερή σε πειράματα όπου χρησιμοποιήθηκε subject-independent random sampling validation. Τέλος υπήρξε 6% αύξηση της απόδοσης του συστήματος χρησιμοποιώντας συνδυασμό AAM χαρακτηριστικών με τα Baseline LGBP-TOP χαρακτηριστικά, συγκριτικά με τη περίπτωση αποκλειστικής χρήσης των Baseline LGBP-TOP χαρακτηριστικών του AVEC 2014 dataset .

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Ανάλυση Πολυμεσικού Περιεχομένου

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Depression, Video Analysis, Audio Analysis

AKNOWLEDGMENTS

Completing this thesis, I think I owe sincere thanks to my supervisor Stavros Peradonis and Theodore Giannakopoulos, for their guidance and support as well as the opportunities they offered me for gaining substantial knowledge.

I would also like to thank my dear friend John Sioulas and my family for their support during my postgraduate studies.

CONTENTS

LIST OF FIGURES

LIST OF IMAGES

LIST OF TABLES

PROLOGUE

This thesis was carried out at NCSR Demokritos, under the supervision of Stavros Perantonis and Theodore Giannakopoulos. This thesis resulted in the following publication:

T. Giannakopoulos, C. Smailis, S. Perantonis and C. Spyropoulos, "Realtime depression estimation using mid-term audio features," International Workshop on Artificial Intelligence and Assistive Medicine, pp. 41-46, 2014.

1. INTRODUCTION

1.1 Automatic Estimation Of Depression

A considerable portion of the world population is affected by depression. According to numbers from the World Health Organization , until 2012 there were at least 350 million people living with some form of depression. Severe cases of depression impact negatively the lives of patients. In fact depression is the leading cause of disability in the world [1]. For those patients, a strict and almost permanent monitoring is necessary in order to control the progress of the disease and to prevent undesired side effects.

A method of measuring a patient's level of depression is the Beck Depression Inventory-II test (BDI-II, [2]). It is a very popular psychological tool and is currently used in numerous clinical settings. BDI-II contains 21 questions, where each is a forced-choice question scored on a discrete scale with values ranging from 0 to 3. The final BDI-II scores range from 0 to 63: 0 to 13: indicates minimal depression, 14 to 19: indicates mild depression, 20 to 28: indicates moderate depression, 29 to 63: indicates severe depression. Although the measurement and assessment of behaviour is a central component of mental health practice it is severely constrained by individual subjective observation and lack of any real-time naturalistic measurements.

For the aforementioned reason, automatic estimation or detection of depression has gained research interest during the last years, through means of audio-visual signal analysis. The Audio-Visual Emotion recognition Challenge (AVEC) 2014 [3] and 2013 challenges [4] are efforts towards developing decision support tools that can help patients and therapists to keep track of the progress of the disease. In particular, these challenges focus on the analysis of audiovisual information recorded from patients performing a predefined task.

There are two tasks presented in the AVEC challenges: the Affect Recognition Sub-challenge (ARS), the goal of which, is to predict continuous affective dimensions (Valence, Arousal and Dominance) and the Depression Recognition Sub-challenge (DRS), the aim of which is depression analysis. This thesis focuses on the latter task by using the data offered by the AVEC 2014 challenge. The goal in the DRS is to develop methods that can automatically predict the value of a self-reported depression indicator from an audiovisual recording [3]. A decision system capable of making predictions with acceptable rates would be very useful not only for monitoring patients with diagnosed depression, but also, to identify people who may have the disease and are not aware of their situation. The scenario considered in the challenge is appealing because it is a non-invasive procedure that does not rely on specialized equipment and can be applied massively.

However, the DRS is very challenging in multiple ways: there is a small sample of training and development instances for building a predictive model; information was recorded with a standard webcam and under uncontrolled conditions; the users were not advised to maintain a fixed position/behavior with respect to the webcam and/or microphone.

1.2 Related Work

In this section, a general overview of the scientific area of automatic depression level estimation is given. The related literature has been categorized based on the modalities used.

1.2.1 Video:

In [5] the performance of head pose and movement features extracted from face videos using a 3D face model projected on a 2D Active Appearance Model (AAM), in the context of a binary classification task (depressed vs. non-depressed) is examined. This is achieved by modeling low-level and statistical functional features for an SVM classifier using real-world clinically validated data. Findings of these research, illustrated that the head pose and movement could be used as a complementary cue, since their recognition rate was 71.2% on average. The method proposed in this paper was evaluated using real world data collected from a study at the Black Dog Institute, a clinical research facility in Sydney, Australia.

Similarly in [6] the performance of eye movement features extracted from face videos using Active Appearance Models (AAMs) for a binary classification task (depressed vs. non-depressed) is analyzed. The aforementioned research concluded that eye movement low-level features gave 70% accuracy using a hybrid classifier of Gaussian Mixture Models and Support Vector Machines. A 75% accuracy was achieved when using statistical measures with SVM classifiers over the entire interview.

Additionally in [7] Active Appearance Models (AAMs) are used to measure facial expression in the scope of depression detection. SVM algorithm was adopted, using leave-one-out validation. Facial AAM derived features demonstrated moderate concurrent validity with depression by achieving an 79% accuracy in detecting depression.

1.2.2 Audio:

In [8] the performance of different audio features is explored for the task of classifying audio files contained within the AVEC 2013 dataset [4], as either depressed or nondepressed. In [9] several novel Canonical Correlation Analysis (CCA) based feature selection methods are presented, in order to reduce the massive dimensionality introduced for the AVEC 2013. Results revealed that using only 17% of original features, a relative improvement of 30% decrease of RMSE over the baseline on challenge test set was obtained, achieving a 10.22 RMSE in the test partition of the AVEC 2013 dataset.

Moreover, in [10], authors used features that reflected changes in coordination of vocal tract motion associated with depression. Specifically, they investigated changes in correlation that occur at different time scales across formant frequencies and also across channels of the delta-mel-cepstrum. With these feature sets, using the AVEC 2013 depression dataset, they designed a novel Gaussian mixture model (GMM)-based multivariate regression scheme, that provided a root-mean-squared-error (RMSE) of 7.42 in the test partition of the AVEC 2013 Dataset.

In [11] authors explored a diverse set of features based only on spoken audio to understand which features correlate with self-reported depression scores according to the Beck depression rating scale. These features, included estimated articulatory trajectories during speech production acoustic characteristics, acoustic-phonetic

characteristics and prosodic features. Features were modeled using a variety of approaches, including support vector regression, a Gaussian backend and decision trees. The aforementioned approach achieved a 7.71 RMSE score on the development partition and 11.10 RMSE score on the test partition of the AVEC 2014 Dataset.

1.2.3 Fusion:

Other studies have adopted different approaches by attempting to fuse audio as well as video modalities in order to boost their performance. For example in AVEC 2014 [3], local dynamic appearance descriptor histograms of Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) has been adopted as video features. On the other hand, the AVEC 2014 audio baseline feature set consists of various acoustic low-level descriptors (LLDs), such as relative spectral MFCCs, spectral energies, and voicing/unvoiced related features. The reported RMSE scores for the development partition of the AVEC 2014 dataset, using video features, is 9.26 while for the test partition of the dataset is 10.859.

For the AVEC 2013, the same audio features have been adopted, while for the video modality, a combination of geometric features (head location, head motion, head pose) and appearance features (Local Phase Quantisation - LPQ) is employed.

Other studies that used the AVEC 2013 dataset for evaluation purposes are the following: In [12] the authors made use of Motion History Histogram (MHH) in the video modality to capture the movement of each pixel (texture variation) within the face area. In the audio modality the authors used a set of spectral Low-Level Descriptors (LLD). MHH were once again used for extracting change information of the vocal expression. For each modality, the Partial Least Square (PLS) regression algorithm is applied and predicted values of visual and vocal clues were further combined at decision level.

In [13] a multimodal approach using a GMM-UBM system with three different kernels for the audio subsystem and Space Time Interest Points in a Bag-of-Words approach for the vision subsystem is presented. These are then fused at the feature level to form the combined audio visual system. Key results include the strong performance of acoustic audio features and the bag-of-words visual features in predicting an individuals level of depression.

A series of other studies use the AVEC 2014 dataset: For example in [14] Moore-Penrose Generalized Inverse (MPGI) and Extreme Learning Machines (ELM) are utilized for audio-visual depression recognition. The authors used the original 2014 baseline video feature set, enriched with Local Phase Quantization (LPQ) features. The reported RMSE score for the test partition of the dataset is 9.61.

In [15] the original audio AVEC 2014 feature set was extended using streams that model a different cue in the auditory spectrum that is related to human speech production such as spectral shape, spectro-temporal modulations, periodicity structure due to the presence of pitch harmonics, and the long-term spectral variability profile. The video baseline feature set was extended using static LBP features per frame, LBP features computed along three orthogonal planes (LBP-TOP), motion information captured through optical-flow-based motion vectors as well as features derived from facial landmarks fitted using mean-shift based deformable model fitting. Moreover this study also made use of text based features through the generation of affective lexicon, followed by the extraction of several session-level features. To reduce feature dimensionality authors resorted to several variations of feature selection methods such as forward selection, brute-force selection and backward selection. The reported

RMSE scores using the fusion of the three aforementioned modalities is 7.44 for the development partition and 9.85 for the test partition of the dataset.

In another study [16], researchers made use of the a affective dimensions as attributes, while extracting audio features through a voice/silence segmentation process. For the video modality, authors used a variety of general features that mainly comprise motion and velocity information such as the difference of final and initial positions of face and eyes within the video segment as well as the average, minimal and maximal coordinates of face and eyes, the average velocity of face and eyes in x and y axis and finally the motion history image, motion static image and motion average image from the segment of video that contains the face only. A feature consolidation procedure and a fusion scheme based on a meta-classifier information are also proposed. This approach achieved an 8.3 RMSE score on development partition and a 10.82 RMSE score for the test partition of the AVEC 2014 Dataset using both video and audio modalities.

In [17] an i-vector based representation for short term audio features has been adopted. In this study a classification step prior to regression was also employed, to allow having different regression models depending on the presence or absence of depression. Experiments showed that a combination of the employed audio-based model and two other models based on the LGBP-TOP video features led to RMSE score of 7.90 for the development partition and an RMSE score of 10.43 for the test partition of the AVEC 2014 dataset using both audio and video modalities. Moreover in this research two different systems for depression classification, one based on the i-vector representation and another on the LGBP-top features were presented. Both systems had a similar accuracy of 82%.

Additionally in [18] the use of a dynamic feature generation method in the extracted video feature space based on the idea of Motion History Histogram (MHH) in combination with the use of Local Binary Patterns (LBP), Edge Orientation Histogram (EOH) and Local Phase Quantization(LPQ) for 2-D video motion extraction is proposed. Partial Least Squares (PLS) and Linear regression were applied to learn the relationship between the dynamic features and depression scales using training data, and then to predict the depression scale for unseen ones. Decision level fusion was done for combining predictions from both video and audio modalities. The proposed approach was also evaluated on the AVEC 2014 dataset achieving an 9.09 RMSE for the development partition and an 10.26 RMSE for the test partition of the AVEC 2014 dataset, using both modalities.

In [19] researchers examined the use of Local Binary Patterns-Three Orthogonal Planes(LBP-TOP) and Dense Trajectories for depression assessment on the AVEC 2014 dataset. The visual information was encoded using Fisher Vector encoding. The fusion of the video and audio modalities achieved an 8.16 RMSE for the development partition and the 10.24 RMSE for the test partition of the AVEC 2014 dataset.

The best results for the AVEC 2014 dataset, were obtained by the methodology presented in [20]. In this study, authors presented a multimodal analysis pipeline that exploits complementary information in audio and video signals for estimating depression severity, by investigated how speech source, system, and prosody features, along with facial action unit features, correlate with depression. The fusion of video and audio modalities , resulted in an 8.12 RMSE for the test partition of the AVEC 2014 dataset.

1.3 Thesis Contribution

This thesis focuses on the Depression Recognition Sub-challenge (DRS), task of the AVEC 2014, by evaluating the performance of regression schemes, using features derived from Active Appearance Model Coefficients and the audio modality. The performance of the features extracted from each modality, are examined either independently or through their fusion. For the estimation of depression level, regression methodologies based on Support Vector Machines are adopted.

1.4 Thesis Outline

The rest of this thesis is organized as follows:

- Chapter 2 describes the methodology used for the extraction of features from the video modality.
- Chapter 3 describes the specifics of the features extracted from the audio modality.
- Chapter 4 presents the theory of SVM Regression that was employed in our experimental setup.
- In Chapter 5 the results obtained for all the different experiments are discussed.
- Finally in the Epilogue, we analyze the conclusions derived by our results and make some final considerations for future work.

2. VIDEO MODALITY

This chapter focuses on the extraction of features from the video modality. This process consists of four steps:

1. Detection of face and eyes.
2. Registration of the cropped facial images based on eye locations.
3. Fitting the AAM Model in each extracted facial image.
4. Computation of statistical measures derived from AAM shape and appearance coefficients.

Each section in the remainder of this chapter is dedicated in explaining how each step of the video feature extraction procedure is implemented, as well as presenting the theory related to Active Appearance Models and Active Orientation Models.

2.1 Face Detection And Registration

Since the head pose and distance to the camera vary over time in the Depression recordings of the AVEC 2014 dataset, we first detect the location of the face, and within that the locations of the eyes to help reduce the pose variance. To obtain the face position, we employ the open-source implementation of the Viola & Jones face detector that is included in OpenCV. This returns a four-valued descriptor of the face position and size. To refine the detected face region, we proceed with detection of the locations of the eyes. This is again performed using the OpenCV implementation of a Haarcascade object detector, trained for either a left or a right eye. Let us define the detected left and right eye locations as p_l respectively p_r, and the line connecting p_l and p_r as l_e.The angle between l_e and the horizontal axis is then defined as α. The registered image is now obtained by rotating it so that $\alpha = 0$ degrees, then scaled to make the distance between the eye locations $\|p_l - p_r\| = 100$ pixels, and finally cropped to be 200 by 200 pixels, with p_r at position $\{p_r^x, p_r^y\} = \{80,60\}$ to obtain the registered face image. In frames where the eyes or the face, are not detected, Linear Interpolation is employed in order to estimate the coordinates of the face bounding box.

2.2 Active Appearance Models and Active Orientation Models

Active Appearance Models (AAMs) [21], [22], [23], [24], [25], first proposed in [26], are non-linear, generative, and parametric models of a certain visual phenomenon. The most frequent application of AAMs to date has been face modeling [27]. However, AAMs may be useful for other phenomena too [28],[29]. In a typical application, the first step is to fit the AAM to an input image, i.e. model parameters are found to maximize the match between the model instance and the input image. The model parameters are then used in whatever the application is. For example, the parameters could be passed to a classifier to yield a face recognition algorithm. Many different classification tasks are possible. In [27], the same model was used for face recognition, pose estimation, and expression recognition. Fitting an AAM to an image is a non-linear optimization problem. The usual approach [21], [24], [25] is to iteratively solve for incremental additive updates to the parameters (the shape and appearance coefficients.) Given the current estimates of the shape parameters, it is possible to warp the input image onto the model coordinate frame and then compute an error image between the current

model instance and the image that the AAM is being fit to. In most previous algorithms, it is simply assumed that there is a constant linear relationship between this error image and the additive incremental updates to the parameters. The constant coefficients in this linear relationship can then be found either by linear regression [21], [30], [26] or by other numerical methods [24], [25]. AAMs are generally able to fit unseen faces however the reported fitting performances have been always outperformed [31]. For this reason in this work, we adopt Active Orientation Models (AOMs), which is a variant of AAMs that have been proved to generalize well to unseen faces and variations by performing better than state-of-the-art algorithms for the same task [31]. The main difference of AOMs with the AAM paradigm is that they use a different statistical model of Appearance.

2.2.1 Active Appearance Models

An AAM is defined by a shape, appearance and motion model. The shape model is typically learned by annotating N fiducial points $s_i = [x1, y1, x2, y2, \dots xN, yN]$ to each of a set of training images $\{I_i\}$ and then applying PCA on s_i. The resulting model $\{\bar{s}; \Phi_s \in \mathrm{R}^{\{2\mathrm{N},\mathrm{P}\}}\}$ can be used to represent a test shape s_y as

$$\hat{\mathrm{s}}_{\mathrm{y}} = \bar{\mathrm{s}} + \Phi_s \mathrm{p}$$

$$\mathrm{p} = \bar{\Phi}_s^T (\mathrm{s}_{\mathrm{y}} - \bar{\mathrm{s}}) \tag{2.1}$$

The appearance model is learned by first warping each of the training images $I_i(x) \in R^K$ to the canonical reference frame defined by the mean shape s_i using motion model $W(x; p)$ and then applying PCA on the shape-free textures. We choose piecewise affine warps as the motion model in this work. The resulting model $\{\bar{\alpha}, \Phi_A \in \mathrm{R}^{\{\mathrm{K},q\}}\}$ can be used to represent a shape-free test texture a_y as

$$\hat{\alpha}_{\mathrm{y}} = \bar{\alpha} + \Phi_A c$$

$$\mathrm{c} = \bar{\Phi}_A^T (\mathrm{s}_{\mathrm{y}} - \bar{\alpha}) \tag{2.2}$$

Given a test image I_y, inference in AAMs entails estimating p and c after initializing the fitting process. This initialization is typically performed by placing the mean shape according to the output of a face detector. Various algorithms and cost functions have been proposed to estimate p and c such as regression, classification and non-linear optimization methods. The latter approach minimizes the $l-2$ norm of the error between the model instance and the given image with respect to the model parameters as follows:

$$\{\mathrm{p}_0, \mathrm{c}_0\} = \arg\min_{\{\mathrm{p},\mathrm{c}\}} \left\| \mathrm{I}\big(\mathrm{W}(\mathrm{x}; \mathrm{p})\big) - \bar{\alpha} - \bar{\Phi}_A c \right\|^2 \tag{2.3}$$

2.2.2 Active Orientation Models

The deformable model fitting framework of the previous section has been highly criticized as difficult to optimize mainly due to the high-dimensional parameter space and the existence of numerous undesirable local minima in the cost function of (2.3). Therefore, the problem in hand is how to avoid these local minima during optimization. Active Orientation Models address this problem by using a similarity criterion robust to outliers. We define outliers to be anything that the learned appearance model cannot reconstruct because:

1. It was not seen in the training set (e.g. appearance variation due to different identity, expression or illumination)
2. It does not belong to the face space at all (e.g. glasses)
3. It was excluded from Φ_A as noise because in any case the number of principal components in Φ_A should be kept as small as possible so that the model is easier to optimize and cannot generate appearance which is unrelated to faces.

AOMs are designed to use the same shape and motion model as the ones used by AAMs but a different appearance model and a different cost function to fit this model. AOMs, use a robust similarity criterion based on image gradient orientations (IGO) [32]. In order to measure the similarity between two images $I_i, i = 1, 2$, for each image, we extract image gradients $g_{i,x}, g_{i,y}$ and the corresponding estimates of gradient orientation φ_i In the following equation z_i represents the so-called normalized gradients:

$$z_i = \frac{1}{\sqrt{K}} [\cos(\varphi_i)^T, \sin(\varphi_i)^T]^T \tag{2.4}$$

where $\cos(\varphi_i) = [\cos(\varphi_i(1)), \dots, \cos(\varphi_i(K))]^T$ and $\sin(\varphi_i)$ is similarly defined. Then, the following kernel can be used to measure image similarity

$$s = z_1^T z_2 = \frac{1}{K} \sum_{k \in \Omega} \cos\big(\varphi_1(k) - \varphi_2(k)\big) \tag{2.5}$$

Where Ω denotes the image support ($\Omega = \Omega_1 \cup \Omega_2$). Ω_1 is the image support that is outlier-free and Ω_2 the image support that is corrupted by outliers. The robust kernel of equation (2.5) can be used to define a kernel PCA [32] . The appearance model in AOMs is learned using this robust PCA. Note that the kernel can be written using the explicit mapping of (2.5) by computing the normalized image gradients, defining the data matrix Z the columns of which are the shape-free normalized gradients of the training faces and then apply standard PCA on Z. To preserve the kernel properties no subtraction of the mean normalized gradient is needed and the first eigenvector is treated as the mean where it is required. By $\{\Phi_z \in \mathrm{R}^{2\mathrm{Kq}}\}$ we denote the learned appearance model. Inference in AOMs is performed by maximizing the normalized correlation of a test image with the learned appearance model

$$\{p_0, c_0\} = \arg\max_{\{p,c\}} \frac{z[p]^T \Phi_z c}{\|\Phi_z c\|} \tag{2.6}$$

where $z[p]$ denotes the normalized gradients of $\mathrm{I}(\mathrm{W}(\mathrm{x};\mathrm{p}))$, using the inverse compositional framework presented in [30]. As an optimization method, we adopted the alternating optimization to maximize (2.6) with respect to both $\{p, c\}$.

2.3 AAM Implementation

For this thesis, an AAM implementation from the Menpo project [33] was used. Menpo is a set of Python libraries for manipulating data that is particularly useful Computer Vision, since it contains tools required to build, fit, visualize, and test deformable models like AAMs or Constrained Local Models (CLMs) [34]. The adopted AAM implementation uses IGO features extracted from each face image instead of the original images. IGO (Image Gradient Orientations) [32] features concatenate the cos and sin of the gradient orientation angles for each image pixel. Moreover a multiresolution model of appearance is adopted by utilizing a Gaussian pyramid of 3 levels in order to improve the correctable range of displacement of the AAM initialization shape. Each video frame is scaled down by a factor of 2 in each level of the pyramid. Finally 90% of the energy was retained in the PCA dimensionality reduction step, resulting in 16, 10 and 7 shape eigenvectors and 772, 604 and 306 appearance eigenvectors for each Gaussian pyramid level. The training of our AAM Model has been performed using the faces images from the Labeled Faces in the Wild database [35].

2.4 AAM Feature Extraction

For generating features from the AAM model, an extension of the technique described in [6], that includes mid-term feature extraction is employed. Each interview video, is divided into short-term segments before feature extraction. In particular, the video is broken into (non-)overlapping 30 frame (1 sec) segments. For each segment the mean, median, and standard deviation of velocities (frame to frame differences) in the coefficients corresponding to each shape and appearance eigenvector, are computed. The result of this procedure is three different sequences of feature vectors per video, computed from the shape and appearance eigenvectors separately. After this step, the feature extraction process proceeds in two ways:

1. Computation of three long-term feature vector per video.
2. Computation of three mid-term feature vector sequences per video.

For the case of the long-term feature vector computation, the short-term statistics are combined by taking their mean, median, minimum, and maximum values. Thus three feature vectors corresponding to each video are formed:

- A feature vector computed by utilizing shape eigenvectors
- A feature vector computed by utilizing appearance eigenvectors
- A feature vector created by concatenating the shape and appearance derived vectors

Since the PCA dimensionality reduction process, mentioned in the previous section, resulted in 16, 10 and 7 shape eigenvectors for each Gaussian pyramid level, the corresponding long-term feature vector consists of 396 dimensions. Similarly since the

same process resulted in 772, 604 and 306 appearance eigenvectors the corresponding feature vector consists of 20184 dimensions. The concatenated feature vector, has 20580 dimensions.

For the case of mid-term feature extraction, the mean, median, minimum, and maximum values of the short-term statistics are computed for mid-term windows of 300 frames (10 seconds). This results in the composition of 3 sequences of feature vectors:

- A feature vector sequence computed by utilizing shape eigenvectors
- A feature vector sequence computed by utilizing appearance eigenvectors
- A feature vector sequence created by the concatenation of the two previous ones.

Each feature vector within these sequences corresponds to the mid-term window from which it was derived. Similarly to the long-term case, mid-term feature vectors contained consist of 396 dimensions if they were computed using shape eigenvectors, 20184 dimensions if they were computed from appearance eigenvectors. Finally the concatenated feature vectors consist of 20580 dimensions.

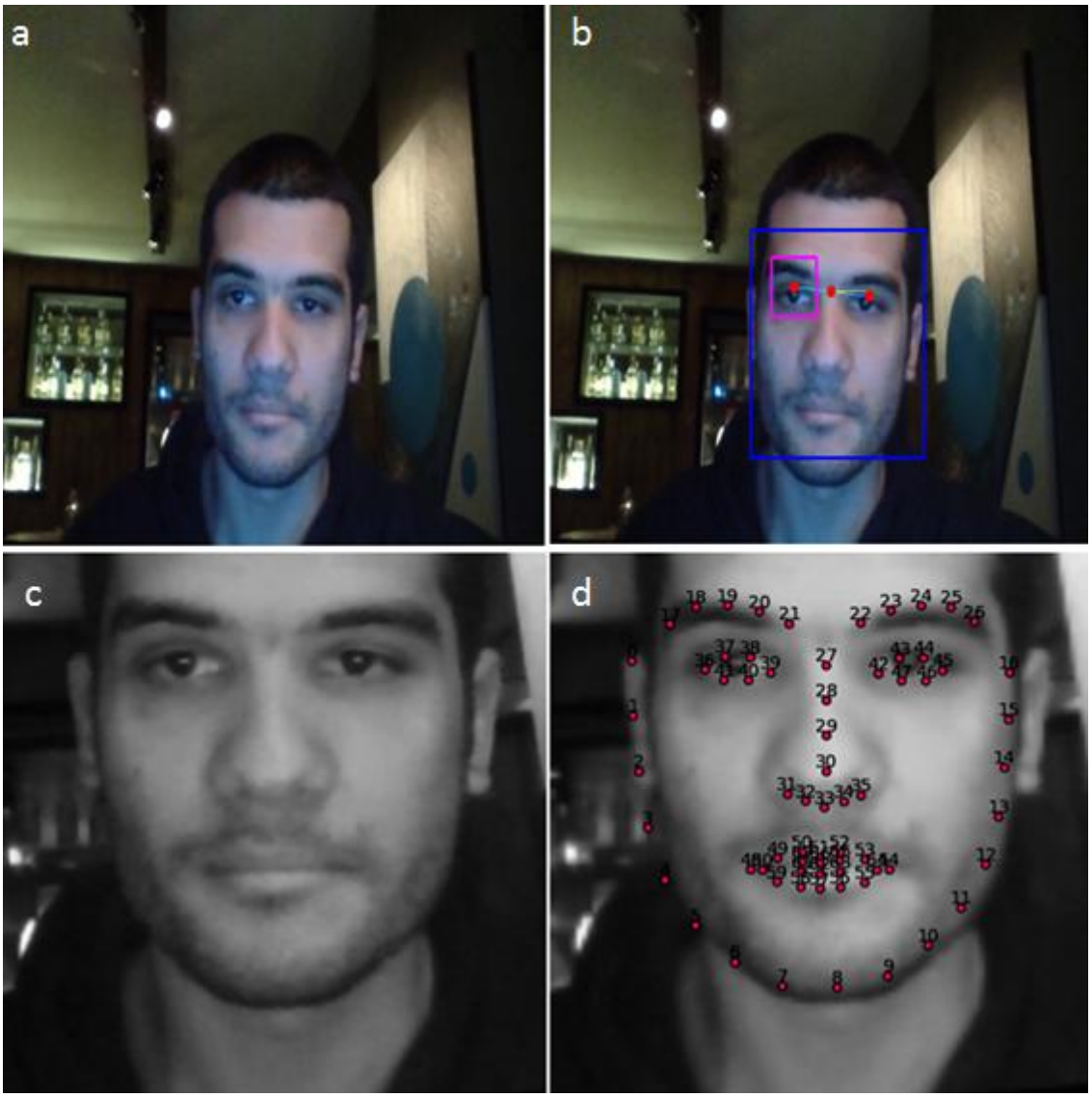

Image 2-1 Applying AAMs in Face Images: a) Original Frame b) Face and Eye detection c) Image registration d) AAM Fitting

3. AUDIO MODALITY

In most audio analysis and processing methods, it is rather common that the input audio signal is divided into short-term frames before feature extraction. In particular, the audio signal is broken into (non-)overlapping frames and a set of features is extracted for each frame. The result of this procedure is a sequence of feature vectors per audio signal. Another common technique used as a second step in audio feature extraction is the processing of the feature sequence on a midterm basis. The audio signal is first divided into mid-term segments and then, for each segment, the short-term processing stage is carried out. At a next step, the feature sequence, which has been extracted from a mid-term segment, is used for computing feature statistics. In practice, the duration of mid-term windows typically lies in the range 1 - 10 secs, depending on the application domain. In this work, extensive experimentation has led to selecting a 10 second mid-term window and a 50 msec short-term frame. In both cases, 50% overlap has been adopted.

3.1 Short-Term Audio Features

In this section we describe the adopted short-term features. These features have been used in several general audio analysis methods and speech processing applications [36], [37], [38] and they cover a wide range of audio signal properties achieving discrimination abilities in several classification and regression tasks.

- **Energy**: The energy feature is computed as a sum of squared signal values (in the time domain), normalized by the window length. Short-term energy usually exhibits high variation over successive speech frames, since speech signals contain weak phonemes and short periods of silence between words.

- **Zero Crossing Rate**: The Zero Crossing Rate (ZCR) of an audio signal is the rate of sign-changes of the signal divided by the duration of that signal. ZCR has been interpreted as measure of the noisiness of a signal, therefore it usually exhibits higher values in the case of noisy signals.

- **Entropy of Energy**: The entropy of energy is a measure of abrupt changes in the energy level of an audio signal. It is computed by firstly dividing each frame in sub-frames of fixed duration. Then, for each sub-frame j its energy is computed and divided by the total energy. Finally, the entropy of that sequence of (normalized) sub-energies e_j is computed as the final feature value. This feature has lower values if there exist abrupt changes in the energy envelop of the respective signal.

- **Spectral Centroid and Spread**: The spectral centroid and the spectral spread are two basic spectral domain features that quantify the position and shape. The spectral centroid is the center of gravity of the spectrum, while spectral spread is the second central moment of the spectrum.

- **Spectral Entropy**: Spectral entropy is computed in a similarly to the entropy of energy, however it is applied on the frequency domain.

- **Spectral Flux**: Spectral Flux is a measure of spectral change between two successive frames and it is computed as the squared difference between the normalized magnitudes of the spectra of the two successive frames.

- **Spectral Rolloff**: Spectral rolloff is the frequency below which a certain percentage of the magnitude distribution of the spectrum is concentrated. It can be treated as a spectral shape descriptor of an audio signal and it has been used for discriminating between voiced and unvoiced sounds.

- **MFCCs**: The Mel-Frequency Cepstrum Coefficients (MFCCs) have been very popular in the field of speech analysis. In practice, MFCCs are the discrete cosine transform coefficients of the mel-scaled log-power spectrum. MFCCs have been widely used in speech recognition, musical genre classification, speaker clustering and many other audio analysis applications.

- **Chroma vector**: This is a 12-dimensional representation of the spectral energy of an audio signal. This is a widely used descriptor, mostly in music related applications, however it has also been used in speech analysis. The chroma vector is computed by grouping the spectral coefficients of a frame into 12 bins representing the 12 equal-tempered pitch classes of western-type music.

3.2 Audio Feature Vector Extraction

The process described in the previous section leads to a sequence of short-term feature vectors of 21 dimensions (this is the total number of short-term features described above). As a next step, statistics are calculated in a mid-term basis as described in the beginning of this chapter. In particular, the following statistics are computed:

- Average value μ
- Standard deviation σ^2

This leads to several mid-term feature vectors of 68 elements. The number of these mid-term vectors depends on the overall duration of the audio signal. Each of these vectors are fed as input in the next regression step that produces the final depression estimate decision.

4. DEPRESSION ESTIMATION

In this thesis the problem of depression level estimation is treated as a regression task. The following sections present necessary theory for SVM classification and regression and introduce the regression schemes that were adopted.

4.1 Support Vector Machine Classification

Support Vector Machine Classification (SVM) has been used with success in many previous studies related to depression detection as well as the relevant problem of facial expression recognition [39] [40] [41] [7]. As a powerful machine learning technique for data classification, SVM [42] performs an implicit mapping of data into a higher (maybe infinite) dimensional feature space, and then finds a linear separating hyperplane with the maximal margin to separate data in this higher dimensional space.

Given training vectors $\{x_i \in \mathrm{R}^{\mathrm{n}}, \mathrm{i} = 1, \dots, \mathrm{l}\}$ in two classes, and an indicator vector $x_i \in \mathrm{R}^{\mathrm{l}}$ such that $y_i \in 1, -1$, Support Vector Machines classifiers [43] [44] solve the following primal optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{l}\xi_i$$

subject to:

$$y_i\,(w^T\varphi(x_i) + b) \geq 1 - \xi_i$$

$$\xi \geq 0\,, i = 1, \dots, l$$

where $\varphi(x_i)$ maps x_i into a higher-dimensional space and $C > 0$ is the regularization parameter. Slack variables ξ_i measure the degree of misclassification of the data x_i. Due to the possible high dimensionality of the vector variable $\boldsymbol{w}$, usually we solve the following dual problem:

$$\min_{\alpha} \frac{1}{2}\alpha^T Q a - e^T a$$

subject to:

$$y^T a = 0$$

$$0 \leq a_i \leq C\,, i = 1, \dots, l$$

where $e = [1, \dots 1]^T$ is the vector of all ones, Q is an l by l positive semi definite matrix $Q_{\iota j} = y_\iota y_j K(x_i, x_j)$ and $K(x_i, x_j) \equiv \varphi(x_i)^T\varphi(x_j)$ is the kernel function. After the primal optimization problem has been solved, using the primal-dual relationship, the optimal $\boldsymbol{w}$ satisfies:

$$\boldsymbol{w} = \sum_{i=1}^{l} y_i a_i \varphi(\boldsymbol{x}_i)$$

and the decision function is:

$$f(x) = sgn(\boldsymbol{w}^T\varphi(\boldsymbol{x}) + b) = sgn(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b)$$

It must be mentioned that α_i are the Lagrange multipliers of the dual optimization problem that describe the separating hyperplane and b is the threshold parameter of the hyperplane. The training sample x_i with $\alpha_i > 0$ is called support vectors, and SVM finds the hyperplane that maximizes the distance between the support vectors and the hyperplane. SVMs allow domain-specific selection of the kernel function. The most frequently used kernel functions are:

- Linear: $K(x_i, x) = x_i x$
- Polynomial, features are mapped with polynomials of a defined degree d: $K(x_i, x) = (\gamma x_i \cdot x + r)^d, \gamma > 0$
- Radial Basis Functions (RBF): $K(x_i, x) = e^{(-\gamma \|x_i - x\|^2)}, \gamma > 0$

4.2 Support Vector Machine Regression

In Support Vector Regression the role of the maximal separation margin is inverted. In this case, the goal is to find the optimal regression hyperplane so that most training samples lie within an ε-margin around this hyperplane. Consider a set of training points, $\{(\mathrm{x}_1, \mathrm{z}_1), \ldots, (\mathrm{x}_\mathrm{l}, \mathrm{z}_\mathrm{l})\}$ where $x_i \in R^n$ is a feature vector and $z_i \in R^l$ is the target output. Under given parameters $C > 0$ and > 0 , the standard form of support vector regression [42] is

$$\min_{w,b,\xi} \frac{1}{2} \boldsymbol{w^T w} + C \sum_{i=1}^{l} \xi_i + C \sum_{i=1}^{l} \xi_i^*$$

subject to:

$$\boldsymbol{w^T} \varphi(x_i) + b - z_i \leq \varepsilon - \xi_i$$
$$z_i - \boldsymbol{w^T} \varphi(x_i) - b \leq \varepsilon - \xi_i^*$$
$$\xi_i \xi_i^* \geq 0, i = 1, \ldots, l$$

The dual problem is:

$$\min_{\alpha,\alpha^*} \frac{1}{2} (\alpha - \alpha^*)^{\boldsymbol{T}} Q (\alpha - \alpha^*) + \varepsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + C \sum_{i=1}^{l} z_i (\alpha_i + \alpha_i^*)$$

subject to:

$$\boldsymbol{e^T} (\alpha - \alpha^*) = 0$$
$$0 \leq a_i, \alpha_i^* \leq C, i = 1, \ldots, l$$

where $Q_{ij} \equiv K(x_i, x_j) \equiv \varphi(x_i)^T \varphi(x_j)$. $K(x_i, x_j) \equiv \varphi(x_i)^T \varphi(x_j)$ represents the kernel function as mentioned in section 4.1. Slack variables ξ_i, ξ_i^* measure the deviation of training samples outside ε - insensitive zone. After solving the dual problem the approximate function is:

$$f(x) = \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + K(x_i, x) + b$$

SVM Regression has been employed for the purpose of estimating a BDI score for each video. To accomplish this task, we made use of the epsilon-SVM implementation contained in LIBSVM with Linear Kernel.

4.3 Regression Schemes

In our experiments, we employ six different Regression schemes for the purpose of BDI estimation:

- Regression scheme using Monomodal long-term features
- Regression scheme using Monomodal mid-term features
- Regression scheme using long-term features with Early Fusion
- Regression scheme using mid-term features with Early Fusion
- Regression scheme using long-term features with Late Fusion
- Regression scheme using mid-term-term features with Late Fusion

The details of the aforementioned schemes, are examined in the next subsections:

4.3.1 Regression using long-term features

For the purpose of regression using long-term monomodal features, each recording is represented using only one feature vector extracted from the respective modality.

In the case of early fusion, each video is represented by a concatenated vector composed of the feature vectors derived from the respective video and audio modalities.

During late fusion, SVM regression is performed separately for the two modalities and the final decision is computed as the mean of the two BDI score estimates.

4.3.2 Regression using mid-term features

For the purpose of regression using mid-term monomodal features, each recording is represented by a feature matrix of T rows, where T= duration of recording/10 (10 sec is the mid-term window size). Note that since one decision is calculated per feature vector, the final decision (per recording) is extracted by averaging the mid-term BDIs. This rationale helps in generating a sufficient number of samples for the SVM regression model training phase. In the case of early fusion, each video is represented by a concatenated matrix composed of the feature matrices derived from the respective video and audio modalities. On the other hand, during late fusion, SVM regression is performed separately for the two modalities and the final decision is computed as the mean of the BDI score estimates.

5. EXPERIMENTATION

5.1 Avec 2014 Dataset Details

Audio-Visual Emotion Challenge and Workshop 2014(AVEC 2014) [3] was the fourth competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, video and audio-visual emotion analysis.

The Depression dataset is formed of 150 videos of task oriented depression data, recorded in a human-computer interaction scenario. It includes recordings of subjects performing a Human-Computer Interaction task while being recorded by a webcam and a microphone.

There is only one person in every recording and the total number of subjects is in the dataset is 84.The subjects were recorded between one and four times, with a period of two weeks between the measurements. 18 subjects appear in three recordings, 31 in 2, and 34 in only one recording. The length of the full recordings is between 50 minutes and 20 minutes (mean = 25 minutes). The total duration of all clips is 240 hours. The mean age of subjects was 31.5 years, with a standard deviation of 12.3 years and a range of 18 to 63 years. The recordings took place in a number of quiet settings.

The behavior within the clips consisted of different human computer interaction tasks which were Power Point guided. The recordings in the AVEC 2014 subset consist of only 2 tasks. Both tasks are supplied as separate recordings, resulting in a total of 300 videos (ranging in duration from 6 seconds to 4 minutes 8 seconds). The two tasks are:

1. **Northwind**: Participants read aloud an excerpt of the fable "Die Sonne und der Wind" (The North Wind and the Sun), spoken in the German language.
2. **Freeform**: Participants respond to one of a number of questions such as: "What is your favorite dish? ","What was your best gift, and why? ", "Discuss a sad childhood memory " again in the German language.

The original audio was recorded using a headset connected to the built-in sound card of a laptop at a variable sampling rate, and was resampled to a uniform audio bitrate of 128kbps using the AAC codec. The original video was recorded using a variety of codecs and frame rates, and was resampled to a uniform 30 frames per second at 640 x 480 pixels. The codec used was H.264, and the videos were embedded in an mp4 container.

For the organization of the challenge, the recordings were split into three partitions: a training, development, and test set of 150 Northwind-Freeform pairs, totaling 300 task recordings.

5.2 Experiments

We conducted two main experiments using the AVEC 2014 Dataset. In the first experiment we trained the Regression Schemes presented in the previous chapter using the Training partition and evaluated it using the Development Partition of the AVEC 2014 Dataset. This experiment allows us to compare the performance of our methodology with other approaches utilizing AVEC 2014 Dataset, in the literature. The measurements are conducted separately for the recordings of the Freeform and Northwind tasks. In order to conduct measurements on the merged dataset, we concatenated the corresponding Freeform and Northwind recordings in both partitions before performing feature extraction.

In the second experiment, we used a subset of the videos contained in the Training and Development Partitions AVEC 2014 Dataset. In this custom dataset, each subject appears only in one video. This policy has been utilized in order to reduce bias in the experiment.

In all experiments, the Root mean square error (RMSE) has been adopted as a performance measure. The RMSE refers to the mean value of the squared deviations of the predictions from the true values. It can be computed as shown in the following formula:

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \widehat{y_k})^2}$$

The results of the experiments are summarized in the following sections.

5.3 Experiment 1: AVEC Training - Development Partitions

As mentioned in the previous section, in the first experiment we trained the Regression Schemes presented in using the 50 recordings of the Training partition and evaluated it using the 50 recordings of the Development Partition of the AVEC 2014 Dataset. The measurements are conducted separately for the recording parts of the Freeform and Northwind tasks. In order to conduct measurements on the merged dataset, we concatenated the corresponding Freeform and Northwind part of the recordings in both partitions before performing feature extraction. A grid search over the hyper parameters C of the SVR algorithm has been performed for the following values: [0.001, 0.01, 0.1, 0.25, 0.5, 1.0, 2.5, 5.0, 10.0, 100]. Only the best RMSE scores from the grid search were included in the result tables presented below.

Table 5-1: RMSE results for the AVEC Training - Development Partitions using recordings from the Freeform task.

Freeform	Modalities	Long Term	Mid Term
Monomodal	Audio	10.78	10.78
	Shape	11.07	13.07
	Appearance	13.4	12.05
	Shape + Appearance	**10.67**	12.05
Early Fusion	Audio + Shape	11.26	11.7
	Audio + Appearance	**10.143**	11.16
	Audio + Shape + Appearance	10.37	11.19
Late Fusion	Audio + Shape	**12.667**	13.284
	Audio + Appearance	12.718	13.028
	Audio + Shape + Appearance	12.987	13.053

Table 5-2: RMSE results for the AVEC Training - Development Partitions using recordings from the Northwind task.

Northwind	Modalities	Long Term	Mid Term
Monomodal	Audio	11.45	11.45
	Shape	11.95	11.5
	Appearance	12.044	11.53
	Shape + Appearance	**10.87**	11.19
Early Fusion	Audio + Shape	11.96	10.57
	Audio + Appearance	10.88	10.86
	Audio + Shape + Appearance	**10.84**	10.84
Late Fusion	Audio + Shape	**10.84**	10.901
	Audio + Appearance	11.378	10.875
	Audio + Shape + Appearance	11.377	10.875

Table 5-3: RMSE results for the AVEC Training - Development Partitions using the concatenated Freeform and Northwind recordings.

Both Tasks	Modalities	Long Term	Mid Term
Monomodal	Audio	11.62	11.62
	Shape	12.41	11.5
	Appearance	13.09	11.53
	Shape + Appearance	**10.84**	11.19
Early Fusion	Audio + Shape	12.08	10.57
	Audio + Appearance	10.83	10.51
	Audio + Shape + Appearance	10.86	**10.44**
Late Fusion	Audio + Shape	11.527	11.555
	Audio + Appearance	12.155	**11.452**
	Audio + Shape + Appearance	12.631	11.452

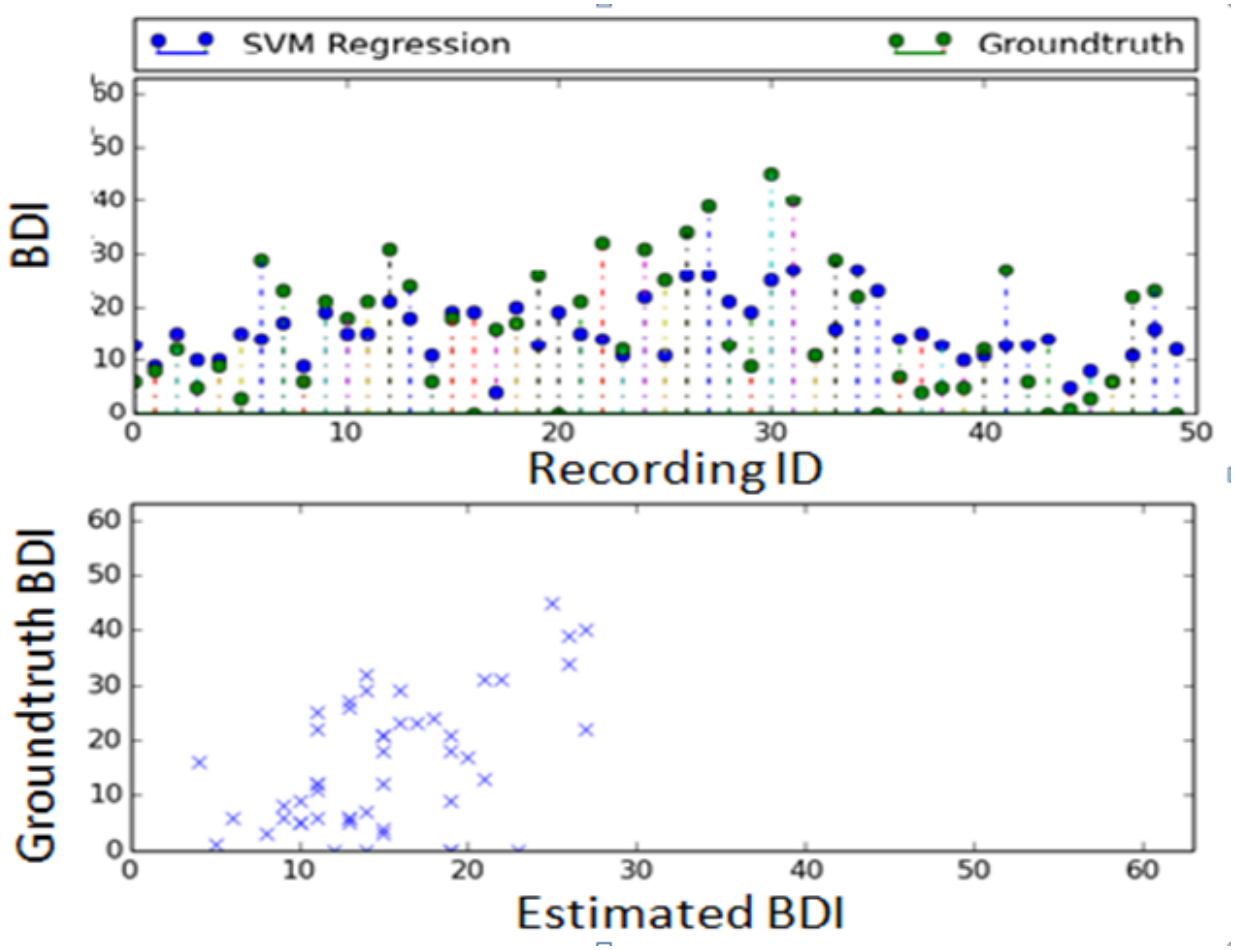

Figure 5-1: Comparison between estimated BDIs and Groundtruth BDIs for the concatenated Freeform recordings of the Development Partition. The regression schemed used in this measurement, utilized mid-term features from the early fusion of audio and Appearance modalities. The obtained RMSE was 10.143. The SVR C parameter value was 0.001.

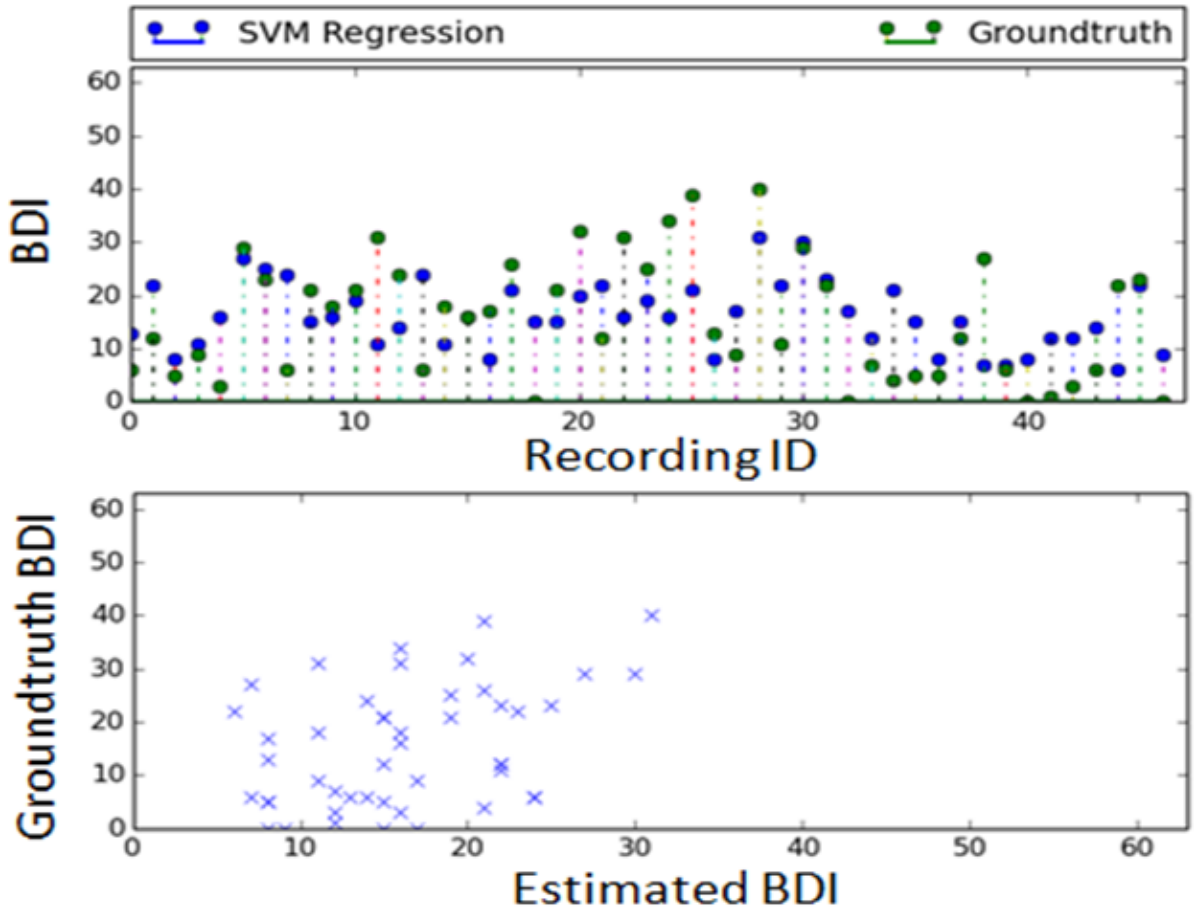

Figure 5-2: Comparison between estimated BDIs and Groundtruth BDIs for the concatenated Northwind recordings of the Development Partition. The regression schemed used in this measurement, utilized mid-term features from the early fusion of audio and shape modalities. The obtained RMSE was 10.57. The SVR C parameter value was 0.25.

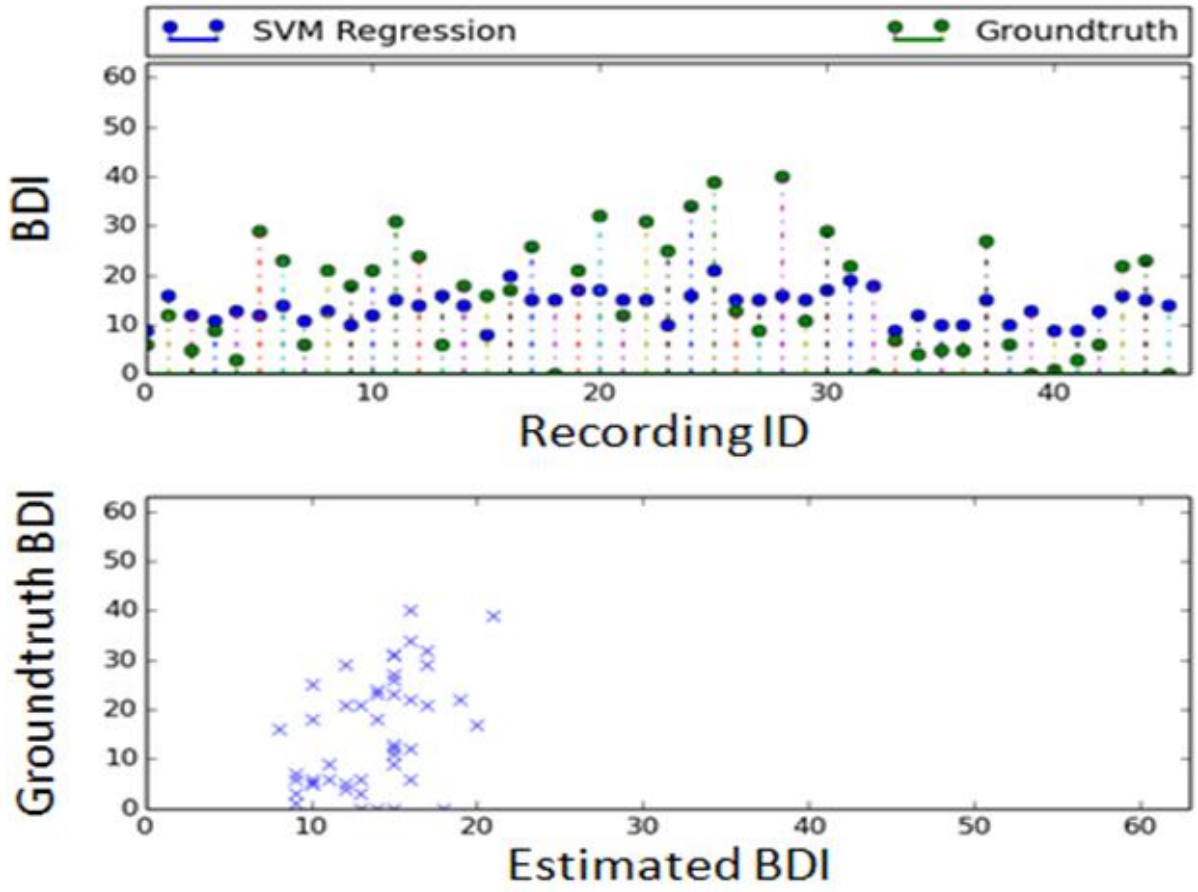

Figure 5-3: Comparison between estimated BDIs and Groundtruth BDIs for the concatenated Freeform and Northwind recordings of the Development Partition. The regression schemed used in this measurement, utilized mid-term features from the early fusion of audio, shape and Appearance modalities. The obtained RMSE for this measurement was 10.44. The SVR C parameter value was 0.01.

5.4 Experiment 2: Custom AVEC 2014 Subset with unique subjects

In the second experiment, we used a subset of 59 videos contained in the Training and Development Partitons AVEC 2014 Dataset. In this custom dataset, each subject appears only in one video. This policy has been utilized in order to reduce bias in the experiment. For the evaluation of this experiment a repeated random sampling procedure has been adopted and tested for 1000 samples. A grid search over the hyperparameters C of the SVR algorithm has been performed for the following values: [0.001, 0.01, 0.1, 0.25, 0.5, 1.0, 2.5, 5.0, 10.0, 100]. Only the best RMSE scores from the grid search were included in the result tables presented here.

Table 5-4: RMSE results for the AVEC Subset containing unique subjects using the concatenated Freeform recordings.

Freeform	Modalities	Long Term	Mid Term
Monomodal	Audio	11.73	11.73
	Shape	11.32	12.15
	Appearance	10.83	11.28
	Shape + Appearance	**10.7**	11.4
Early Fusion	Audio + Shape	11.37	11.93
	Audio + Appearance	10.87	11.22
	Audio + Shape + Appearance	10.92	**10.07**
Late Fusion	Audio + Shape	11.444	12.062
	Audio + Appearance	**11.262**	11.405
	Audio + Shape + Appearance	11.403	11.944

Table 5-5: RMSE results for the AVEC Subset containing unique subjects using the concatenated Northwind recordings.

Northwind	Modalities	Long Term	Mid Term
Monomodal	Audio	11.44	11.44
	Shape	12.62	11.91
	Appearance	11.91	11.37
	Shape + Appearance	11.81	**11.31**
Early Fusion	Audio + Shape	11.58	10.622
	Audio + Appearance	11.82	10.16
	Audio + Shape + Appearance	11.8	**10.05**
Late Fusion	Audio + Shape	12.053	11.66
	Audio + Appearance	**11.164**	11.601
	Audio + Shape + Appearance	11.474	11.366

5.5 Additional Experiments

Except of the experiments presented in the previous sections we performed a series of additional experiments on the dataset. These experiments include:

1. The evaluation of the system using early fusion between Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) video features adopted by the baseline system AVEC 2014 system for the case of the merged dataset.
2. Calibration of the output produced by the best combination of features from the previous experiment.
3. The use of Principant Component Analysis as a feature dimensionality step before performing evaluation on the merged dataset.

5.5.1 A few words about LGBP-TOP features

LGBP-TOP takes a block of consecutive input video frames which are first convolved with a number of Gabor filters to obtain Gabor magnitude response images for each individual frame. This is followed by LBP feature extraction from the orthogonal XY, XT and YT slices through the set of Gabor magnitude response images. The resulting binary patterns are histogrammed for the three orthogonal slices separately, and concatenated into a single feature histogram.

5.5.2 Experiment 1: Early Fusion with the merged dataset.

In this experiment, the evaluation of the system was performed, using early fusion between LGBP-TOP video features adopted by the baseline system AVEC system for the case of the merged dataset. Only one long term feature vector was used to represent each recording for each modality.

Table 5-6: RMSE results for the AVEC Training - Development Partitions using the concatenated Freeform and Northwind recordings.

Both Tasks	Modalities	Long Term
Monomodal	Audio	11.62
	Shape	12.41
	Appearance	13.09
	Shape + Appearance	10.84
	LGBP-TOP	**9.69**
Early Fusion	Audio + Shape	12.08
	Audio + Appearance	10.83
	Audio + Shape + Appearance	10.86
	LGBP-TOP + Audio	10.58
	LGBP-TOP + Shape	9.89
	LGBP-TOP + Appearance	10.76
	LGBP-TOP + Shape + Appearance	**9.07**
	Audio + Shape + LGBP-TOP	12.13
	Audio + Appearance + LGBP-TOP	10.74
	Audio + Shape + Appearance + LGBP-TOP	9.16

5.5.3 Experiment 2: Calibration

Calibration was performed as a linear regression step performed in the BDIs estimated from the test partition of the merged dataset, using the early fusion of LGBP-TOP, Shape and Appearance features. Only one long term feature vector was used to represent each recording for each modality. The RMSE was reduced to 9.00 after the use of the calibration procedure.

5.5.4 Experiment 3: Use of Principant Component Analysis

The third experiment examined the use of Principant Component Analysis (PCA) as a method of dimensionality reduction. We performed PCA for the video long term feature vectors of the merged dataset. The appropriate number of principant components was empirically determined to be equal to 35 for all video feature types. As a second step a Random Projection procedure was performed prior to PCA reducing the initial dimensions to 500. Results are summarized in the table below:

Table 5-7: RMSE results for the AVEC Training - Development Partitions using the concatenated Freeform and Northwind recordings.

Both Tasks	Modalities	Long Term	Random Projection + PCA
Monomodal	Shape	12.33	12.34
	Appearance	11.78	**12.08**
	Shape + Appearance	11.63	12.19
	LGBP-TOP	**9.03**	12.21
Early Fusion	Audio + Shape	12.42	13.36
	Audio + Appearance	10.5	12.48
	Audio + Shape + Appearance	11.79	12.18
	LGBP-TOP + Audio	**9.92**	12.72
	LGBP-TOP + Shape	12.92	13.36
	LGBP-TOP + Appearance	11.45	12.48
	LGBP-TOP + Shape + Appearance	10.96	12.33
	Audio + Shape + LGBP-TOP	11.97	11.82
	Audio + Appearance + LGBP-TOP	10.5	11.12
	Audio + Shape + Appearance + LGBP-TOP	11.41	12.21

6. CONCLUSION AND FUTURE DIRECTIONS

We have presented a method for detecting a subject's clinical depression score using audiovisual information. The performance of video features derived from Active Appearance Model shape and appearance coefficients and several Audio features has been examined individually, as well as through the fusion of their respective modalities by using SVM based regression schemes. Results demonstrated that the fusion of features extracted from the AAM with the audio features leads to better performance compared to individual modalities. This means that the AAM and audio features act in a complementary manner. Moreover in the experiment with the custom version of the AVEC Dataset that contained unique subjects, performance remained stable. In addition to the aforementioned conclusions, the fusion of the baseline Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) features with the AAM derived features let to an RMSE improvement of 6%. Future research directions of this thesis could possibly include temporal analysis techniques to enhance the signal representation process by selecting representative areas of a recording with high discrimination ability in terms of depression analysis.

ACRONYMS

BDI-II	Beck Depression Inventory-II test
AVEC	Audio-Visual Emotion recognition Challenge
ARS	Affect Recognition Sub-challenge
DRS	Depression Recognition Sub-challenge
AAM	Active Appearance Model
GMM	Gaussian Mixture Models
SVM	Support Vector Machines
RMSE	Root Mean Square Error
CCA	Canonical Correlation Analysis
LGBP-TOP	Local Gabor Binary Patterns from Three Orthogonal Planes
LLD	Low-Level Descriptors
MFCC	Mel-Frequency Cepstrum Coefficients
LPQ	Local Phase Quantisation
MHH	Motion History Histogram
PLS	Partial Least Square regression
MPGI	Moore-Penrose Generalized Inverse
ELM	Extreme Learning Machines
LBP	Local Binary Patterns
LBP-TOP	Local Binary Pattern features computed along three orthogonal planes
AOM	Active Orientation Model
PCA	Principant Component Analysis
SVR	Support Vector Regression

APPENDIX I

A. Installation and Technical Details

The software framework developed for the task of depression level estimation within the context of this thesis is analyzed within this chapter. The software framework has been developed using Python 2.7. It has been tested in an Ubuntu 14.04 LTS machine.

B. Dependencies

Before using the software framework, several dependencies must be installed first. The full list of dependencies and the terminal commands used to install them is given in this section:

Table 2: Dependencies and Terminal Commands

Library	Commands
NUMPY	sudo apt-get install python-numpy
MATPLOTLIB	sudo apt-get install python-matplotlib
SCIPY	sudo apt-get install python-scipy
GSL	sudo apt-get install libgsl0-dev
MLPY	wget http://sourceforge.net/projects/mlpy/_les/mlpy%203.5.0/mlpy-3.5.0.tar.gz tar xvf mlpy-3.5.0.tar.gz cd mlpy-3.5.0 sudo python setup.py install
SKLEARN	sudo apt-get install libgsl0-dev
Simplejson	sudo easy install simplejson
OpenCV	sudo apt-get install python-opencv
eyeD3	sudo apt-get install python-eyed3

An additional dependency than needs to be installed is the Menpo library. The software framework described in this thesis strictly requires the 0.3.0 version of the library. However since the installation process of the library is a bit more complex than the previous ones, we only reference a webpage with instructions on how to install the library: http://www.menpo.org/installation/linux/index.html

C. General Source Code Implementation Details

The core framework code is structured within the following Python files:

- **Main_App_No_PostProcessing_Linear_Interpolation_ExternalFile**: This file takes as input a directory of video _les. It crops and registers face images of the subjects and stores them in an output directory that is specified as an argument and stored within the registered face directory.

- **AAM_Menpo.py**: This _le implements the AAM Model described in the thesis. It takes as input the directory of the cropped face images and outputs two .npz file containing the frame to frame differences of the AAM shape and appearance coefficients for each video as described in Chapter 3. The .npz files produced are stored in the AAMOutput directory.

- **video_longterm_features.py**: This file implements the long-term feature extraction process as described in Chapter 3. It takes as input the .npz files stored in the AAMOutput directory. Long-term feature vectors are stored in the form of .npy files in the feature vectors directory

- **video_midterm_features_+meta.py**: This file implements the mid-term feature extraction process described in Chapter 3. It takes as input the .npz files stored in the AAMOutput directory. Mid-term feature vectors are stored in the form of .npy files in the feature vectors directory

- **joinShapeAndAppearence.py**: This file concatenates long-term monomodal features of the audio and video modalities and stores the result in the feature vectors directory.

- **joinShapeAndAppearenceMT.py**: This file concatenates mid-term monomodal features of the audio and video modalities and stores the result in the feature vectors directory.

- **depressionAnalysis_traditional_experiment.py**: This file implements the regression schemes proposed in chapter 4) that are used in the Experiment 1 (as described in chapter 5).

- **depressionAnalysis.py**: This file implements the regression schemes (proposed in chapter 4) that are used in Experiment 2 (as described in chapter 5).

The aforementioned files are executed from a series of bash shell scripts that are responsible of initializing them. The bash shell file responsible for executing the whole framework is the "create feature vectors.sh" file.

BIBLIOGRAPHY

[1] "World health organization. Depression - a hidden burden. WHO flyer.." http://www.who.int/mediacentre/factsheets/fs369/en/. Link verified on: 24-08-2015.

[2] A. T. Beck, R. A. Steer, R. Ball, and W. Ranieri, “Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients," J Pers Assess, vol. 67, pp. 588-597, Dec 1996.

[3] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional afect and depression recognition challenge," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, (New York, NY, USA), pp. 3- 10, ACM, 2014.

[4] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: The continuous au- dio/visual emotion and depression recognition challenge," in Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC '13, (New York, NY, USA), pp. 3-10, ACM, 2013.

[5] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear, "Head pose and movement analysis as an indicator of depression," in Afec- tive Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, pp. 283-288, IEEE, 2013.

[6] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection.," in ICIP, pp. 4220-4224, 2013.

[7] 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 Afective Computing and Intelligent Interaction and Work- shops, 2009. ACII 2009. 3rd International Conference on, pp. 1-7, IEEE, 2009.

[8] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for the classification of depressed speech,"

[9] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "Cca based feature selection with application to continuous depression recognition from acoustic speech features,"

[10] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 41-48, ACM, 2013.

[11] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, and M. Graciarena, "The sri avec-2014 evaluation system," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 93-101, ACM, 2014.

[12] 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 Proceedings of the 3rd ACM interna- tional workshop on Audio/visual emotion challenge, pp. 21-30, ACM, 2013.

[13] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: a multimodal approach," in Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 11-20, ACM, 2013.

[14] H. Kaya, "Speaker-and corpus-independent methods for affect classification in computational paralinguistics," in Proceedings of the 16th International Conference on Multimodal Interaction, pp. 359-363, ACM, 2014.

[15] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal prediction of afective dimen- sions and depression in human-computer interactions," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 33-40, ACM, 2014.

[16]] H. P´ Espinosa, H. J. Escalante, L. Villase˜ Pineda, M. Montes-y Gomez, D. Pinto-Aveda˜ and V. Reyez-Meza, "Fusing affective dimensions and no, audio-visual features from segmented video for depression recognition: Inaoe- buap's participation at avec'14 challenge," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, (New York, NY, USA), pp. 49-55, ACM, 2014.

[17] M. Senoussaoui, M. Sarria-Paja, J. a. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, (New York, NY, USA), pp. 57-63, ACM, 2014.

[18] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14, (New York, NY, USA), pp. 73-80, ACM, 2014.

[19]]V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in Proceedings of the 4th In- ternational Workshop on Audio/Visual Emotion Challenge, AVEC '14, (New York, NY, USA), pp. 87-91, ACM, 2014.

[20] 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 Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 65-72, ACM, 2014.

[21] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, pp. 681-685, June 2001.

[22] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "A comparative evaluation of active appearance model algorithms.," in BMVC, vol. 98, pp. 680-689, 1998

[23] T. F. Cootes and P. Kittipanya-ngam, "Comparing variations on the active appearance model algorithm.," in BMVC, pp. 1-10, Citeseer, 2002.

[24] T. F. Cootes, C. J. Taylor, et al., "Statistical models of appearance for computer vision," 2004.

[25] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Transactions on pattern analysis and machine intelligence, vol. 23, no. 6, pp. 681-685, 2001.

[26] G. J. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, pp. 300-305, IEEE, 1998.

[27] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 19, no. 7, pp. 743-756, 1997.

[28] M. J. Jones and T. Poggio, "Multidimensional morphable models: A framework for representing and matching object classes," International Journal of Computer Vision, vol. 29, no. 2, pp. 107-131, 1998

[29] S. Sclarof and J. Isidoro, "Active blobs: region-based, deformable appearance models," Computer Vision and Image Understanding, vol. 89, no. 2, pp. 197- 225, 2003

[30] G. J. Edwards, Learning to identify faces in images and video sequences. University of Manchester, 1999.

[31] G. Tzimiropoulos, J. Alabort-i Medina, S. Zafeiriou, and M. Pantic, "Generic active appearance models revisited," in Computer Vision-ACCV 2012, pp. 650-663, Springer, 2013.

[32] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Principal component analysis of image gradient orientations for face recognition," in Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp. 553-558, IEEE, 2011.

[33] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou, "Menpo: A comprehensive platform for parametric image alignment and vi- sual deformable models," in Proceedings of the ACM International Conference on Multimedia, pp. 679-682, ACM, 2014.

[34] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," International Journal of Computer Vision, vol. 91, no. 2, pp. 200-215, 2011.

[35] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained envi- ronments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007

[36] T. Giannakopoulos and A. Pikrakis, Introduction to Audio Analysis: A MATLAB Approach. Academic Press, 2014.

[37] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Fourth Edition. Academic Press, Inc., 2008.

[38] K. Hyoung-Gook, M. Nicolas, and T. Sikora, MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval. John Wiley & Sons, 2005.

[39] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," Journal of Multimedia, vol. 1, no. 6, pp. 22-35, 2006.

[40] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spon- taneous behavior," in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 2, pp. 568-573, IEEE, 2005.

[41] M. F. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on, pp. 76-76, IEEE, 2005.

[42] V. N. Vapnik and V. Vapnik, Statistical learning theory, vol. 2. Wiley New York, 1998.

[43] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computa- tional learning theory, pp. 144-152, ACM, 1992.

[44] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273-297, 1995.