



Artificial Intelligence System for Continuous Affect
Estimation from Naturalistic Human Expressions

A thesis submitted as partial fulfilment of the requirement of
Doctor of Philosophy (Ph.D.)

by

Yona Falinie Abd Gaus

Department of Electronic and Computer Engineering

Brunel University London

January 2018

Abstract

The analysis and automatic affect estimation system from human expression has been acknowledged as an active research topic in computer vision community. Most reported affect recognition systems, however, only consider subjects performing well-defined acted expression, in a very controlled condition, so they are not robust enough for real-life recognition tasks with subject variation, acoustic surrounding and illumination change. In this thesis, an artificial intelligence system is proposed to continuously (represented along a continuum e.g., from -1 to +1) estimate affect behaviour in terms of latent dimensions (e.g., arousal and valence) from naturalistic human expressions. To tackle the issues, feature representation and machine learning strategies are addressed. In feature representation, human expression is represented by modalities such as audio, video, physiological signal and text modality. Hand-crafted features is extracted from each modality per frame, in order to match with consecutive affect label. However, the features extracted maybe missing information due to several factors such as background noise or lighting condition. Haar Wavelet Transform is employed to determine if noise cancellation mechanism in feature space should be considered in the design of affect estimation system. Other than hand-crafted features, deep learning features are also analysed in terms of the layer-wise; convolutional and fully connected layer. Convolutional Neural Network such as AlexNet, VGGFace and ResNet has been selected as deep learning architecture to do feature extraction on top of facial expression images. Then, multimodal fusion scheme is applied by fusing deep learning feature and hand-crafted feature together to improve the performance. In machine learning strategies, two-stage regression approach is introduced. In the first stage, baseline regression methods such as Support Vector Regression are applied to estimate each affect per time. Then in the second stage, subsequent model such as Time Delay Neural Network, Long Short-Term Memory and Kalman Filter is proposed to model the temporal relationships between consecutive estimation of each affect. In doing so, the temporal information employed by a subsequent model is not biased by high variability present in consecutive frame and at the same time, it allows the network to exploit the slow changing dynamic between emotional dynamic more efficiently. Following of two-stage regression approach for unimodal affect analysis, fusion information from different modalities is elaborated. Continuous emotion recognition in-the-wild is leveraged by investigating mathematical modelling for each emotion dimension. Linear Regression, Exponent Weighted Decision Fusion

and Multi-Gene Genetic Programming are implemented to quantify the relationship between each modality. In summary, the research work presented in this thesis reveals a fundamental approach to automatically estimate affect value continuously from naturalistic human expression. The proposed system, which consists of feature smoothing, deep learning feature, two-stage regression framework and fusion using mathematical equation between modalities is demonstrated. It offers strong basis towards the development artificial intelligent system on estimation continuous affect estimation, and more broadly towards building a real-time emotion recognition system for human-computer interaction.

Declaration

I, Yona Falinie Abd Gaus, here declare that the work presented in this thesis was carried out by myself at Brunel University London, and no part of this work has been previously submitted to Brunel University London, nor any other academic institution, for admission to a higher degree. Some of the work has appeared in the forms of publications, and those are listed on the List of Publications section.

Yona Falinie Abd Gaus
January 2018

Acknowledgements

In the name of Allah, the Most Gracious and the Most Merciful

Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis. Firstly, I would like to express my sincere gratitude to my supervisor Dr Hongying Meng for his guidance, useful advice, and encouragement. His invaluable help of constructive comments and suggestions throughout the experimental and thesis works have contributed to the success of this research.

I am also deeply thankful to the Brunel Graduate School for their support and help towards my postgraduate affairs. My acknowledgement also goes to all the technicians and office staffs of Department of Electronic and Computer Engineering, Brunel University London for their co-operations.

I would like to express my appreciation to my team member, Asim Jan, Jingxin Liu, Rui Qin, Fan Zhang and Yi Liu for their expertise and productive critic, especially during seminars, discussions thus provided new ideas for the work.

Sincere thanks to my sisters by heart; Dr Noorhasyimah Ismail, Dr Akhma Adlin Khalid, Dr Anisah Hambali, Dr Liyana Azmi, Dr Nurashikin Suhaili, Dr Sumayyah Dzulkifly. *~ kerana Murobbi itu lebih dari seorang guru yang mengajar hingga faham, tapi betul betul menyentuh hati kita ~*

I wish to express my unqualified thanks to my parents, Abd Gaus Hj Mansor and Masmun Bakri. I could never have accomplished this dissertation without your love, support and understanding. I also wish to thank my brothers; Mohd Hanif, Hadi Nazrin, Nor Halim and Mohd Harraz for doing their best to understand a sister who had to be confined to her study for such a long time. Nur Faziera Said, thank you for the friendship!

Last but not least, special appreciation and gratefulness to my home country, Malaysia for their sponsorship through Majlis Amanah Rakyat (MARA) for its financial support throughout my PhD studies.

Thank you so much!

Acronyms

ASM Active Shape Model

AVEC Audio-Visual Emotion recognition Challenge

AlexNet Alex Network

BOW Bag-of-Words

BOAW Bag-of-Audio-Words

BOTW Bag-of-Text-Words

BOVW Bag-of-Video-Words

BLSTM Bidirectional Long Short Term Neural Networks

CCC Concordance Correlation Coefficient

CNN Convolutional Neural Network

DBN Deep Belief Network

EEG Electroencephalogram

EMFACS Emotional Facial Action Coding System

EOH Edge Orientation Histogram

EMG Electromyography

eGeMAPS Geneva Minimalistic Acoustic Parameter Set

EW Exponent weighted

EMOTIW Emotion-in-the-Wild

FAU Facial Action Unit

FAC Facial Action Coding System

FFT Fast Fourier Transform

fc fully-connected

GP Genetic Programming

GPU Graphics Processing Unit

HCI Human Computer Interaction

HOG Histogram of Oriented Gradients

HR Heart Rate

HRV Heart Rate Variability

ILSVRC ImageNet Large Scale Visual Recognition Challenge

KF Kalman Filter

LBP Local Binary Pattern

LPQ Local Phase Quantization

LLD Low-Level Descriptor

LPC Linear Prediction Cepstral

LSTM Long Short Term Memory

LGBP-TOP Local Gabor Binary Patterns from Three Orthogonal Plane

MNIST Modified National Institute of Standards and Technology

MGGP Multi-Gene Genetic Programming

MFCC Mel-Frequency Cepstral Coefficients

OSS Open-Source Software

OA Output-Associative

PLS Partial Least Squares

PCC Pearson Correlation Coefficients

ReLU Rectified Linear Unit

RECOLA Remote Collaborative and Affective Interaction

RMSE Root Mean Square Error

RBM Restricted Boltzmann Machine

RNN Recurrent Neural Network

RGB Red Green Blue

ResNet Residual Network

RVM Relevance Vector Machines

SVM Support Vector Machine

SEMAINE Sustained Emotionally coloured Machine-Human Interaction using Nonverbal Expression

SEWA The Automatic Sentiment Analysis in the Wild

SVR Support Vector Regression

SCL Skin Conductance Level

SCR Skin Conductance Response

SDM Supervised Descent Method

TDNN Time Delay Neural Network

VAD Voice-Activity Detector

VGG Visual Geometric Group

WT Wavelet Transform

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Research Gap	4
1.3	Aim and Objectives	4
1.4	Contribution	5
1.5	List of Publications	6
1.6	Thesis Outline	7
2	Literature Review	9
2.1	Affective Computing	9
2.2	Modalities	11
2.2.1	Visual Modality	12
2.2.1.1	Facial Action Unit	12
2.2.1.2	Geometric Feature	12
2.2.1.3	Appearance Feature	15
2.2.1.4	Deep Learning Feature	16
2.2.2	Audio Modality	19
2.2.3	Physiological Signal Modality	20
2.2.4	Text Modality	21
2.3	Database	21
2.3.1	SEMAINE Dataset (AVEC 2014 Challenge)	22
2.3.2	Remote Collaborative and Affective Interaction Dataset-RECOLA (AVEC 2016 Challenge)	23

2.3.3	The Automatic Sentiment Analysis in the Wild-SEWA (AVEC 2017 Challenge)	24
2.4	Regression Approach	24
2.4.1	Support Vector Regression	26
2.4.2	Linear Regression	27
2.4.3	Recurrent Neural Network-Long Short Term Memory	28
2.4.4	Time Delay Neural Network	31
2.4.5	Kalman Filter	32
2.4.6	Genetic Programming	34
2.5	Chapter Summary	37
3	Continuous Affect Estimation based on Wavelet Filtering and PLS Regression	38
3.1	Introduction	38
3.2	Related Database and Regression Technique	40
3.3	Methodology	41
3.3.1	Stage 1: Feature Extraction	41
3.3.1.1	Edge Orientation Histogram	42
3.3.1.2	Local Binary Pattern	42
3.3.1.3	Local Phase Quantization	42
3.3.1.4	Audio Feature Extraction	43
3.3.2	Stage 2: Wavelet Filtering	43
3.3.2.1	Haar Wavelet Transform	44
3.3.2.2	Haar Wavelet Filtering	44
3.3.3	Stage 3: Modeling Approach	45
3.3.3.1	Partial Least Square Regression	45
3.3.3.2	Filtering on Initial Estimation	46
3.3.4	Stage 4: Final Estimation using Decision Fusion	46
3.4	Experimental Evaluation	47
3.4.1	Dataset of AVEC 2014	47
3.4.2	Experimental Results and Discussion	48
3.4.2.1	Results on Development Set	49

3.4.2.2	Results on Test Set	51
3.5	Comparison on the best performer of the Challenge	52
3.6	Chapter Summary	53
4	Continuous Affect Estimation based on Deep Learning Features	55
4.1	Introduction	55
4.2	Related Feature Extraction Technique	57
4.3	Methodology	59
4.3.1	Stage 1: Feature Extraction	59
4.3.1.1	Deep Learning Feature using Convolutional Neural Network	60
4.3.1.2	Hand Crafted Features	67
4.3.2	Stage 2: Support Vector Regression as Modelling Approach	68
4.3.3	Stage 3: Post-processing	68
4.3.4	Stage 4: Fusion Approach for Final Estimation	69
4.4	Experimental Evaluation	69
4.4.1	Dataset of AVEC 2016	70
4.5	Experimental Results and Discussion	71
4.6	Comparison on the best performer of the Challenge	74
4.7	Chapter Summary	75
5	Continuous Affect Estimation in Two Stage Regression Framework	77
5.1	Introduction	77
5.2	Related Features and Modeling Technique	79
5.3	Methodology	82
5.3.1	First-Stage Regression	83
5.3.2	Second-Stage regression	84
5.4	Dataset and Features	89
5.4.1	Audio Features	90
5.4.2	Video Features	90
5.4.3	Physiological Signal	92
5.5	Experimental Evaluation	93
5.6	Results and Discussion	95

5.7	Comparison on the best performer of the Challenge	98
5.8	Chapter Summary	99
6	Linear and Non-linear Multimodal Fusion for Continuous Affect Estimation	
	in-the-Wild	101
6.1	Introduction	101
6.2	Related Works	103
6.3	Dataset and Features	107
6.3.1	Audio	107
6.3.2	Video	108
6.3.3	Text	108
6.3.4	Regression models	108
6.4	Decision Level Fusion	109
6.4.1	Linear Regression (LR)	109
6.4.2	Exponent Weighted Decision Fusion	109
6.4.3	Genetic Programming (GP)	110
6.4.4	Kalman Filter (KF)	112
6.5	Experimental Results	114
6.5.1	Experimental Set-ups and Evaluation Metrics	114
6.5.2	Affect Estimation in Unimodal Modality	115
6.5.3	Affect Estimation in Mutimodal Modality	116
6.5.3.1	Linear Regression	116
6.5.3.2	EW	116
6.5.3.3	GP Modelling	117
6.5.3.4	Performance Comparison	118
6.6	Comparison on the best performer of the Challenge	119
6.7	Chapter Summary	120
7	Conclusion and Future Works	122
7.1	Conclusion	122
7.2	Future Works	124
7.2.1	Wide Variety of Datasets	125

7.2.2	Detecting Mental Health Disorders	125
7.2.3	Reinforcement Learning	125

List of Figures

1.1	Illustration of the commonly utilised pipeline in emotion recognition. (1) Given a set of observations (input) possibly from multiple modality, such as video, audio physiological signal and text. (2) Pre-processing is needed to filter out noisy information in features provided. (3) Refers to the feature extraction method to facilitate the task at hand. (4) Machine learning takes place to give the the features learned without being explicitly programmed. (5) Prediction analysis usually in terms of classification (into discrete classes) or regression (into continuous values).	3
2.1	Facial expressions of the six basic emotions - Anger, Disgust, Neutral, Surprise, Happiness (twice), Fear, and Sadness - taken from [29]	10
2.2	Circumplex of affect with the seven basic emotions displayed as in [45]	10
2.3	Geometric feature points.	14
2.4	Edge Orientation Histogram feature extractor	15
2.5	Local Binary Pattern feature extractor	16
2.6	LGBP-TOP feature extraction procedure: a) original block of frames, b) Gabor magnitude responses, c) XY, XT and YT mean slices of each response and d) LBP histograms concatenated into LGBP-TOP histogram	16
2.7	Typical example of deep learning diagram network.	17
2.8	A typical neural network. The variable x_i is the input value, o_l is the output, q_j and r_k are the hidden variables, and w is the weight for each connection . . .	28
2.9	(a) is a traditional RNN. (b) is this RNN unrolled into a chain.	29
2.10	A simple LSTM block	30

2.11	: (a) Overall architecture of the TDNN. (b) Single TDNN with M inputs and N delays for each input at time t . D_d^i are the registers that store the values of delayed input $I^i(t-d)$	31
2.12	Example of a tree structure representing the model term $\sin(x_1) + \cos(3x_1)$. . .	36
2.13	Example of a tree structure representing the model term $\sin(x_1) + \cos(3x_1)$. . .	36
3.1	Overview of the proposed automatic affective dimension recognition system . . .	41
3.2	Edge Orientation Histogram feature extractor	42
3.3	Local Binary Pattern feature extractor	43
3.4	Haar Wavelet Transform filtering on selected feature space. 1840 frames, selected components and decomposition level=4. (a). Original feature vectors (b). Low and high frequency parts of the feature vectors in wavelet space (c). Filtered feature vectors.	45
3.5	Bar chart comparison with state-of-the-art in AVEC 2014 in terms of average CORR and RMSE values	53
4.1	: A depiction of the pipeline of the proposed system. This depiction is specific to the combination of features from video modality and audio modality. In deep learning feature, $fc-6$ to $fc-8$ of VGGFace, $conv-7$ to $fc-8$ of ResNet and $fc-6$ to $fc-8$ of AlexNet is extracted from each CNN network. This is done for all frames of the video which produces a single feature vector in respective layer. Along with hand-crafted features and LLD descriptor, these feature is then feed into SVR for regression purpose.	60
4.2	CNN architecture for VGG-Face	64
4.3	CNN architecture for AlexNet	65
4.4	: CNN architecture for ResNet. (a) Residual network with 34 parameter layer. Noted that shortcut connection is added by skipping layer.	66
4.5	:Performance value in term of correlation as an exponent q is scanned in the exponentially weighted decision fusion. Noted that when proper q is selected, it gives maximum performance in development sets	70
4.6	Bar chart comparison with state-of-the-art in AVEC 2016 in terms of concordance correlation in Arousal and Valence dimension.	75

5.1	Architecture of Two-stage Regression Modeling for Continuous Emotion Recognition	83
5.2	(a) Feedforward network (b) feedforward network with delay	85
5.3	Each LSTM cell remembers a single floating point value c_t (Equation 5.6). This value may be diminished or erased through a multiplicative interaction with the forget gate f_t (Equation 5.5) or additively modified by the current input x_t multiplied by the activation of the input gate i_t (Equation 5.4). The output gate o_t controls the emission of h_t , the stored memory c_t transformed by the hyperbolic tangent nonlinearity (Equation 5.7 5.8). Images are reproduced from [43]	86
5.4	:At each time step, the input features are fed into the LSTM to compute the final estimation of arousal and valence.	87
5.5	Illustration of the facial landmark features extraction from RECOLA dataset .	92
5.6	:Architecture of Two-stage Regression Modeling of Continuous Affect Recognition by using RECOLA dataset, Video provides two set of features, LBPTOP and video geometric features. Bio-signal such as ECG and EDA reflected by HRHRV and SCR with SCL respectively [165]	95
5.7	Bar chart comparison with state-of-the-art in AVEC 2016 in terms of concordance correlation in Arousal dimension. Results reported based on baseline features.	99
5.8	Bar chart comparison with state-of-the-art in AVEC 2016 in terms of concordance correlation in Valence dimension. Results reported based on baseline features.	99
6.1	:Overview of the proposed system. Fusion of the predictions of the three modalities: audio, video and text.	106
6.2	:Mutation process in GP. The dashed circle part of tree is replaced with a random generated tree	110
6.3	Recombination process in GP. Dashed circle parts in parents are exchange. . .	111
6.4	Graphical formula with three input variable.	111
6.5	:Comparison of C_{corr} and P_{corr} between two time series. Dashed line denoted as prediction label and clear line is gold standard label	115

6.6	Bar chart comparison with state-of-the-art in AVEC 2017 in terms of concordance correlation in multimodal fusion approach.	120
-----	--	-----

List of Tables

2.1	10 example Facial Action Units (FAUs) [30]	13
2.2	Comparison of each database associated in this thesis.	25
3.1	Pearson correlation coefficients of six sub-systems for the AVEC 2014 development set. Bold indicate the highest correlation at each affect dimension across each features and modality.	48
3.2	Pearson correlation coefficients of video final systems for the AVEC 2014 development set without wavelet filtering	50
3.3	Pearson correlation coefficients of final systems for the AVEC 2014 development set with wavelet filtering. Bold indicate the highest correlation at each affect dimension across each features and modality.	50
3.4	Pearson correlation coefficients of six sub-systems for the AVEC 2014 test set. Bold indicate the highest correlation at each affect dimension across each features and modality.	51
3.5	Pearson correlation coefficients of final systems for the AVEC 2014 test set. Bold indicate the highest correlation at each affect dimension across each features and modality.	52
3.6	Performance comparison with state-of-the-art in AVEC 2014 in terms of average CORR and RMSE values.	53
4.1	Delay in seconds applied to the gold-standard, according to the emotional dimension (A= arousal, V=valence). The delay were obtained by maximising the results in development partition while applying the delay in training partition.	69
4.2	CCC obtained in development partition after SVR and post-processing method. Features are taken from deep-learned features.	72

4.3	CCC obtained in development partition after SVR and post-processing method. Features are taken from hand-crafted features.	73
4.4	Comparison of CCC on fusion from baseline and proposed approach. Noted that proposed approach results are obtained on 2 fold cross validation.	74
4.5	Comparison of CCC on fusion of AVEC 2016 state-of-the-art and proposed approach.	75
5.1	32 Acoustic Low-Level Descriptor (LLDs)	91
5.2	Comparison of baseline results with two-stage regression approach of unimodal performance on development sets	96
5.3	Comparison of unimodal performance by Somandepalli et al [153] and SVR-KF on the development set.	98
5.4	Comparison of unimodal performance by Brady et al. and SVR-LSTM on the development set.	98
5.5	Comparison of CCC on fusion of AVEC 2016 state-of-the-art and proposed approach.	100
6.1	Unimodal performance on the development set	118
6.2	Multimodal performance on 2-fold cross validation	119
6.3	Comparison of CCC on fusion of AVEC 2017 state-of-the-art and proposed approach.	120
7.1	Final performance in terms of CCC taken from AVEC 2016 development set. .	123
7.2	Comparison of CCC based on mathematical modeling in Chapter 6. Noted that the proposed approach are obtained on 2 fold cross validation.	124

Chapter 1

Introduction

1.1 Introduction

Since the emergence of Human Computer Interaction (HCI) in the 1980s, researchers have concentrated on finding novel ways to design technologies that let humans interact with computers [177]. One simple way to make human and computer understand each other is to input human emotions into the computer system. For example, in medical imaging, pain detection systems automatically recognise the pain intensity of patients; in the education field, e-learning in schools boosts academic skills; and in brand advertising, it provides better entertainment experience for users by measuring their engagement.

The study of human emotion inside computer system falls into category of affective computing field, a cross-discipline of computer science, psychology, and cognitive science [120] [135]. In affective computing, researchers aim to bridge the gap between human emotions and computational technology by developing intelligent systems that can recognize, interpret, process, and simulate human affect [158]. To recognise emotions, three types of modality can be used: visual modality, audio modality, and physiological signal modality. The visual modality typically involves facial expressions, hand movements and body posture of the users. The most popular one is facial expression because it is the main way for humans to express their emotions [29]. Audio modality such as pitch range, and vowel duration, speech rate can also be used for detecting user emotions. Physiological signal modality refers to heart rate, heartbeat or skin conductance level among others. Each of the modalities described above on their own is referred as uni-modal processing, where the combination of each modality can be

referred as multi-modal processing.

In order to bridge the gap between human emotions and computer system, an automatic affect recognition system is proposed to detect, process, and analyse human emotions in real-time. Emotion recognition is typically done by having the computer acquire information from sensors that capture a variety of modalities. Facial expressions from visual modality and acoustic channel from audio modality in particular, have been shown to be strong indicators of emotion [3]. For example, sadness and fear are more noticeable in a human voice (audio) while happiness and joy are more easily detected in human face (video) [148].

There are two streams of emotion recognition modelling approach: categorical and dimensional approach. The categorical approach classifies human emotion to a limited number of emotions. Dimensional approach represents emotions in a multi-dimensional space so more emotions and subtle changes can be detected. Early researchers were mostly focusing on categorical approach pioneered by Darwin [23], interpreted by Tomkins [160] [161] and supported by findings of Ekman et al. [31]. Ekman and his team claim that there exists a set of six basic emotions (anger, fear, disgust, happiness, sadness and surprise) which are biologically hard-wired to humans and are common across different cultures. As the performance of recognising basic emotions improved, researchers gradually realised that six basic emotion corresponds only to a small subset of the human emotion [92] [93]. This led to the adoption of dimensional approach, which is based on *continuous and dimensional* emotion descriptions. A number of researchers [129] [134] [133] defined the human affect in terms of small latent dimensions. The most commonly used latent dimensions are Valence and Arousal, with Valence indicating how positive (happiness, optimism) or negative (unhappy, depressed) the emotional state is, and Arousal describing how active or passive the emotional state is. It essentially transformed the problem from a classification task to learning continuous real-valued functions, that is regression task. Developing an automated algorithm for continuous emotion recognition will be a central part of this thesis, which will be heavily discussed in this work.

A typical system aimed towards automatic continuous emotion recognition consists of four parts, as shown in Figure 1.1. The first part is the input channel, usually known as modality. Usually, the input channel or modality is in the form of video (facial expression images), audio (speech), physiological signal (ECG signal) and text (verb) modality. The second part is feature extraction. It is done depending on what modalities are used. In the case of facial images

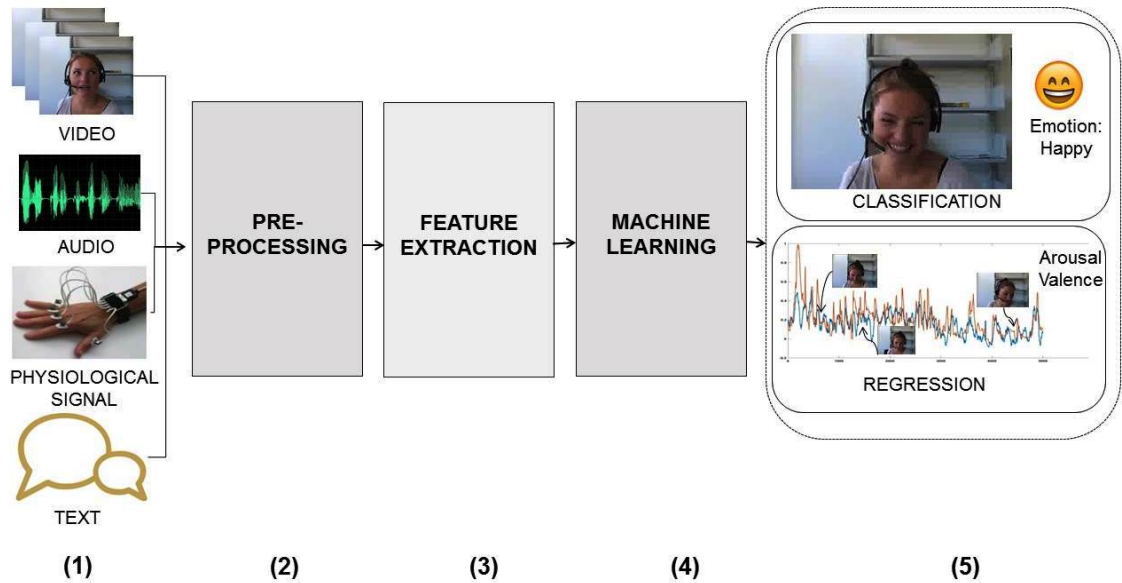


Figure 1.1 Illustration of the commonly utilised pipeline in emotion recognition. (1) Given a set of observations (input) possibly from multiple modality, such as video, audio physiological signal and text. (2) Pre-processing is needed to filter out noisy information in features provided. (3) Refers to the feature extraction method to facilitate the task at hand. (4) Machine learning takes place to give the the features learned without being explicitly programmed. (5) Prediction analysis usually in terms of classification (into discrete classes) or regression (into continuous values).

from video modality, feature representation typically in the form of collection of coordinates essential to the location of various interest points of face, such as the corners of the eyes, the lips and the nose. Features derived from a collection of points are called *geometric features*, while features based on the image pixels are defined as *appearance-based features*. These features are called low-level features, because they only responds to minor details such as edge, gradient or corner of the face. Also, low-level features do not carry specific knowledge about the face to detect. In contrast, high level features concerns with finding face shapes in the images by modelling the face, such as eye position, mouth, hair and so on. In case of audio modality, it can be prosody features such as pitch or energy. In the case of physiological signal modality, it can be electroencephalogram (EEG) signal or galvanic skin response or any type of peripheral signals. The third part is pre-processing step, where raw features obtained from each modality is pre-processed by applying dimensionality reduction technique. The goal of pre-processing is to remove irrelevant components of the raw features, such as noise from background, and amplify relevant components which can be deemed beneficial for next stage. The fourth and fifth step is typically correlated, which involve machine learning technique, be it classifying into discrete emotions, or regression, in order to learn continuous values emotions. Fusion

between modalities, also known as multi-modality is also common approach in the final step of emotion recognition. The advantages of applying decision fusion is when one modality is weak or absent. For example, when the subject does not look at the camera (weak visual modality) or does not speak (weak audio modality). In such cases, a single modality system would fail, whereas a multi-modality system can rely on the other modalities to do the emotion recognition instead.

1.2 Research Gap

Most of the research related to automatic emotion recognition focuses on giving computers the ability to recognize discrete basic emotions. In addition, most of the database used are using static facial images or acted facial expressions. However, emotions are not discrete: they continuously change over time due to their natural progression. Unlike discrete emotion, few studies have investigated continuous emotion dynamic from natural human behaviour. It is not obvious which database or which features that can mapped facial expression or audio speech to emotions space. Therefore, the goal of this thesis is to develop an automatic continuous affect recognition from natural human expressions. Specifically, it adapts uni-modal and multi-modal approach, where affect information taken from single modality or combined to arrive at an emotion label that is represented in Valence-Arousal space.

1.3 Aim and Objectives

The aim of this research is to develop a artificial intelligence system that, when given a set of features from multiple modality, can map them to some appropriate emotion space, say Valence-Arousal space. This thesis focuses on the continuous emotions recognition rather than their discrete posed displays. In order to achieve this goal, a set of objectives has been formulated. The objectives are:

- To develop feature representation from each modality in continuous affect estimation.
- To investigate the problem of noisy feature vector and best possibility to solve it.
- To demonstrate the effectiveness on convolutional neural network features by analysing in terms of layer wise; convolutional and fully connected layer.

- To propose a two-stage regression framework by separating emotional state dynamic modelling from an individual emotional state prediction step based on input features
- To examined the possibility of constructing affect estimation equation by employing proper fusion approach.

1.4 Contribution

This work makes the following four following contributions:

- A direct regression approach to map feature representation to emotion space is introduced. Haar Wavelet Transform is implemented as smoothing effect in feature space is employed.
- New features based on deep learning such as convolutional neural network is introduced. The proposed framework is built, firstly by analysing the features in terms of layer wise: convolutional and fully connected layer, then fuse it together with hand crafted features, in decision label.
- A two stage regression framework is introduced. These approach is achieved by firstly concatenating the strength of an initial SVR model, then using it as a basis for another regression analysis in a subsequent model. Kalman Filter (KF), Long Short Term Memory (LSTM) and Time Delay Neural Network (TDNN) is employed as subsequent model. For the first time, physiological signal is employed as new modality in continuous affect estimation system.
- The possibility of constructing an equation from each modality is developed. Linear Regression (LR), Exponent Weighted Decsion Fusion (EW) and Multi-Gene Genetic Programming (MGGP) is leveraged to quantify the relationship of each modality in each affect dimension.

The primary contributions of this thesis can be divided into two parts; feature level and decision level. In feature level, the demonstration of feature smoothing (Chapter 3) and investigation of deep learning feature in terms of layer-wise (Chapter 4) make up as the first contribution. The secondary contribution is in decision level, where two-stage regression framework

is proposed to model emotional dynamic (Chapter 5) and finally building continuous emotion equation based on multiple modalities (Chapter 6).

1.5 List of Publications

- [A] **Y.F.A.Gaus**, H. Meng, and A. Jan, “Exploring continuous emotion recognition in feature space based deep learning features,” *Journal of Electrical and Computer Engineering*. (*submitted*)
- [B] **Y.F.A.Gaus** and H. Meng, “Linear and non-linear multimodal fusion for continuous affect estimation in-the-wild,” in 2018 13th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2018 (*accepted*)
- [C] **Y.F.A.Gaus**, H. Meng, and A. Jan, “Decoupling temporal dynamics for naturalistic affect recognition in a two-stage regression framework,” 2017 3rd IEEE International Conference on Cybernetics (CYBCONF), Exeter, 2017, pp. 1-6.
- [D] 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*, 2017.
- [E] A. Jan, **Y.F.A.Gaus**, F. Zhang, and H. Meng, “BUL in MediaEval 2016 emotional impact of movies task,” in *CEUR Workshop Proceedings*, 2016, vol. 1739.
- [F] **Y.F.A.Gaus**, H. Meng, A. Jan, F. Zhang, and S. Turabzadeh, “Automatic affective dimension recognition from naturalistic facial expressions based on wavelet filtering and PLS regression,” 2015 11th IEEE FG, Ljubljana, 2015, pp. 1-6.
- [G] **Y.F.A.Gaus**, T. Olugbade, A. Jan, R. Qin, J. Liu, F. Zhang, H. Meng, and N. B.-Berthouze, “Social touch gesture recognition using random forest and boosting on distinct feature sets,” In *Proceedings of ICMI 2015*, ACM, New York, 399-406
- [H] 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 AVEC 2014*, ACM, New York, NY, USA, 73-80.

1.6 Thesis Outline

This thesis is organised in the following Chapters:

Chapter 2 gives a literature review regarding continuous affect recognition in the aspects of modalities, feature extraction, modelling approaches, databases, and applications. This Chapter begins with the definition of affective computing and how does it relate with discrete emotion and continuous emotion. It also includes basic concepts of dimensional emotional space. Subsequently, a comprehensive review is provided on each modality selected and database used throughout the thesis. Then, each of regression approach is detailed to recognise then estimate the performance of each affect dimension.

Chapter 3 introduces the work on automatic continuous affect recognition by introducing different type feature descriptors and investigation of Haar Wavelet Transform . Noisy hand crafted features such as EOH, LBP and LPQ features is investigated using Wavelet Transform to determine if noise-cancellation should be considered in continuous emotion recognition design. Then, PLS regression is adopted as machine learning approach to estimate each of emotion dimension.

Chapter 4 of this thesis investigates the effectiveness of deep learning features in term of layer-wise, as well as employing simple fusion with hand-crafted features. An architecture based on CNN features is investigated, namely AlexNet, VGGFace and ResNet. Each of the network is applied directly to frames, then produce features in terms of layer wise: convolutional and fully connected layer. Following this, experimental analysis is performed by combining hand-crafted features with deep learning features which yields competitive performance over the baseline results.

Chapter 5 dedicated on two-stage regression approach, where SVR model, which captures the strength of the features represents the initial estimation in the first stage regression. Then, it become an input in subsequent model for regression analysis to produce final estimation of affect. By using this approach, it allows the network to exploit the slow changing dynamic between emotional state. Comparison of this approach with baseline and other previous results is also conducted, to show where they provide the best efficiency for continuous emotion recognition.

Chapter 6 leverages the continuous affect recognition 'in-the-wild' setting, by investigating mathematical modeling for each emotion dimensions. Mathematical modeling such as linear

regression, exponent weighted decision fusion and genetic programming are implemented to quantify the relationship between each affect, by employing multimodal fusion approach. The experimental results are shown respectively to show the effect of linear and non linearity of each modality towards emotion dimension.

Chapter 7 concludes the work with a summary of each contributions and presents directions for future works.

Chapter 2

Literature Review

Automatic continuous emotion recognition researchers have recently started exploring how to model, analyze and interpret the continuity of affective behaviour in terms of latent dimensions, such as arousal, dominance, valence and so on. This Chapter reviews the related work regarding the process of automatic continuous affect in the aspects of modality, feature extraction, modelling approaches, databases, and applications.

2.1 Affective Computing

Affective Computing term was first popularised by Rosalind Picard's book defined as '*computing that relates to, arises from, or deliberately influences emotion or other affective phenomena*' [120]. According to Picard, "if we want computers to be genuinely intelligent and to interact naturally with us, we must give computers the ability to recognise, understand, even to have and express emotions". To be exact, it is an inter-disciplinary study for developing intelligent interactive systems that can recognise, interpret, process, and simulate human affects or emotions [158]. While emotion is fundamental in human communication, most computer systems fail to leverage emotion in human-computer interaction. One of the goals of affective computing is to integrate emotion into a computer system to empower them with the ability of providing more accurate, sensitive and respectful response to the user.

Emotion has been represented over the years by dividing it into several categories. The most common categories are *categorical labels* originally proposed by Paul Ekman [32] [28] as shown in Figure 2.1. There are six basic emotion: 1. *anger*, 2. *disgust*, 3. *fear*, 4.



Figure 2.1 Facial expressions of the six basic emotions - Anger, Disgust, Neutral, Surprise, Happiness (twice), Fear, and Sadness - taken from [29]

happiness, 5. *sadness*, and 6. *surprise*. These emotions were selected because it was common and unambiguous across different cultures [32]. The second category is *dimensional labels*, where a person's emotions can be described using a low-dimensional signal that varies with time. Typically, a low dimensional signal can be represented by two or three dimensions. The two most commonly used dimensions are Arousal and Valence. Dimensional labels have two advantages over categorical labels. The first one is dimensional labels can describe a potentially broader set of emotions. Russell [129] describes how the Arousal and Valence dimension defines in a circle called the *circumplex model of affect* and the six basic emotions are represented as specific regions in the circle, as in Figure 2.2. From Figure 2.2, it can be

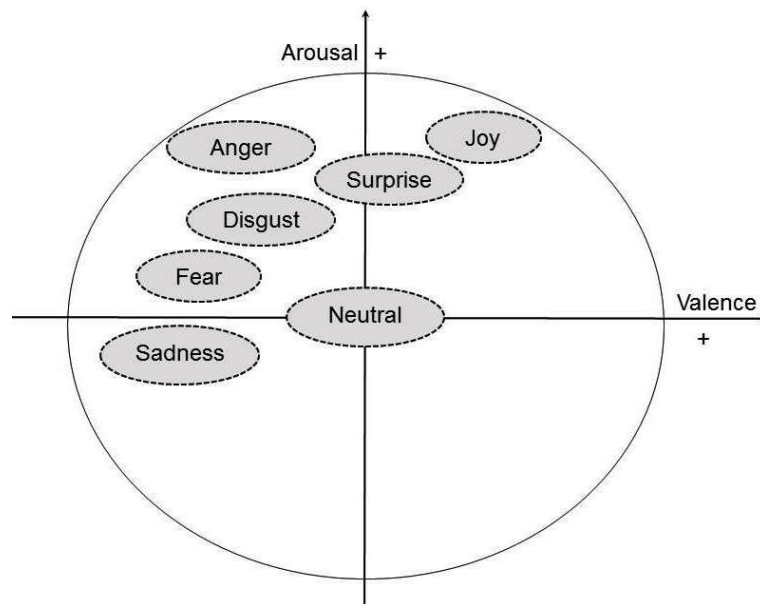


Figure 2.2: Circumplex of affect with the seven basic emotions displayed as in [45]

seen that six basic emotion only cover the upper part of the circumplex. There are no emotion categories from the categorical label that cover the lower-right portion of the circumplex. This region of low Arousal and high Valence corresponds to feelings of tiredness and relaxation, which is rarely discussed in emotion perspective. The second advantage is, having dimensional

label can output time-continuous labels which allow more realistic modelling of emotion over time. This could be particularly useful for representing emotion in video data. The third advantage is, having continuous label can describe the intensity of emotion, which can be used for recognising dynamics and allows for adaptation to individual moods and personalities [138]

While the main basic dimensions of emotion, *Valence* and *Arousal*, are deemed to capture most affective variability encountered in human-computer interaction, other dimensions such as *Dominance* and *Liking* dimension can be represented in the emotional state. In this thesis, it adopted affective annotations in terms of four total emotion dimensions, summarised in what follows.

- Valence refers to the positive or negative feeling of the person's emotional state. The valence scale ranges from unhappy or sad like emotions (negative emotions) to happy or joyful like emotion (positive emotions).
- Arousal points to the person's feeling of dynamism or lethargy. It determines how passive or active the emotion state of the person is.
- Dominance refers to the scale ranges from submissive (or 'without control') to dominant (or 'in control').
- Liking dimension measure inquires about the participants' tastes, not their feelings. For example, it is possible to like videos that make one feel sad or angry. It is introduced in DEAP [85] then adopted in AVEC 2017 [126]. In AVEC 2017, liking dimension is used as the indication of person's preference to the commercial product.

2.2 Modalities

An individual's emotional states usually can be delivered in the forms of three modalities: visual modality, audio modality, and physiological signal modality. Visual modality often refers to as facial expressions, but the body and hand movements can also be part of it. Audio modality includes speech rate, pitch range, and vowel duration. Physiological signal modality refers to data such as blood pressure, heart rate, and skin conductance. Then feature extraction is undergone to produce feature representation for each modality.

2.2.1 Visual Modality

Visual modality usually can be represented as video data. A video can be seen as a frame-by-frame image sequence, and these images are encoded to low or high-level information as feature representations. From video data, the most interesting part of analysing a person affect is facial features. A significant amount of research over the last decade has been devoted to finding and extracting better facial features in order to improve classification accuracy. According to the survey by Zeng et al. [184], many previous emotion recognition techniques either used geometric features, appearance-based features, or both. The next subsection will discuss each category of features which use facial action units (FAUs) and convolutional neural network (CNN).

2.2.1.1 Facial Action Unit











The oldest method that is used to measure emotion is Facial Action Unit (FAU). It is based on an older system proposed by Hjortsjö [61]. Then, Ekman and Friesen enhance it by proposing encoding facial expressions using the Facial Action Coding System (FACS) [30]. Their proposed system explains how facial expressions can be decomposed into contractions of specific muscle groups in the face. Each of these muscle groups corresponds to a “facial action” and is called an action unit/facial action unit (AU/FAU). Once the AUs have been detected, they can be used to determine the emotion of the subject using the Emotional Facial Action Coding System (EMFACS). Common examples of FAUs are listed in Table 2.1

2.2.1.2 Geometric Feature

In this thesis, geometric features is used to represent emotions by explicitly detecting distinctive features in human faces automatically such as corners of the eyes, the tip of the nose, and the edges of the mouth. Pantic and Rothkrantz [116] use landmark points extracted from two views (frontal and profile). Chang et al. [65] modelled the motion of 58 facial landmarks using an Active Shape Model (ASM). Valstar et al. [168] tracked 12 facial points using a particle filter then combined it with head and shoulder motion information to determine the emotion recognition.

Chapter 4 and 5 of this thesis explores the way on how geometric feature can be used to detect continuous affect recognition. It is based on 49 landmarks detected and subsequently

Table 2.1: 10 example Facial Action Units (FAUs) [30]

FAU Number	FAU Name	Example Image
1	Inner Brow Raiser	
2	Outer Brow Raiser	
4	Brow Lowerer	
5	Upper Lid Raiser	
9	Nose Wrinkler	
12	Lip Corner Puller	
15	Lip Corner Depressor	
17	Chin Raiser	
24	Lip Pressor	
25	Lips Part	

tracked with the Cascaded Regression facial point detector/tracker proposed by Xiong and De la Torre [180]. In this method, overall there are 316 features extracted, from four sets of sub-features. The first sets of sub features are from 49 facial landmarks, of every video frame after aligned it with a mean shape using a set of stable point. Stable points are defined as those not affected by AU activations. In the 49 landmarks, the notation points are given as below: 20, 23, 26, 29 (nose region) and 11 - 19 (nose region). These points are considered to be stable. The mean facial landmarks shape by taking mean of 10% randomly selected video frames from every session. The alignment is performed by computing a non-reflective affine transformation, which minimizes the difference between stable point coordinates of the two shapes. All mean shape landmark coordinates are then subtracted from the corresponding aligned shape points resulting in a set of aligned facial points which form the first subset of $49 \times 2 = 98$ geometric features.

The second subsets is composed by subtracting the aligned facial point locations of the previous frame from that of the current one. This applies to all frames except the very first one of every session, for which these features are the same as the first 98 geometric features.

The third subsets, the notation points of facial landmarks are given as below: 20 - 25

(left eye), 1 - 5 (left eyebrow), 26 - 31 (right eye), 6 - 10 (right eyebrow) and 32 - 49 (mouth region). For each notation, a set of features representing Euclidean distances as well as angles in radians between points within the groups is extracted. Distances between points within a group are computed by taking the squared L2-norm between points of:

$$F(i) = \|\tilde{p}_i - \tilde{p}_{i+1}\| \quad (2.1)$$

$$i = \{1 \dots N_{p-1}\} \quad (2.2)$$

where N_p is the total number of points within the region, \tilde{p} is the point coordinates vector and F is the feature array in the region. The angles between two lines are defined by two pairs of points at a time within a group, where the two pairs share one common point. For each consecutive triplet of points Euclidean distances between them are computed first, which are then used to calculate the angle between the points:

$$F(i) = \arccos\left(\frac{\tilde{p}_{12}^2 + \tilde{p}_{13} - \tilde{p}_{23}}{2 \times \tilde{p}_{12} \times \tilde{p}_{13}}\right) \quad (2.3)$$

where \tilde{p}_{ij} is defined as Euclidean distance between points i and j . Using these equation, 71 features in total extracted from the above face regions.

The fourth subsets are computed by getting median of stable points of the aligned shape. Then, Euclidean distance is calculated between each of the aligned shape and the median. In total there are 79 features extracted from fourth subsets. Figure 2.3 shows the final geometric features indicate by red dots.

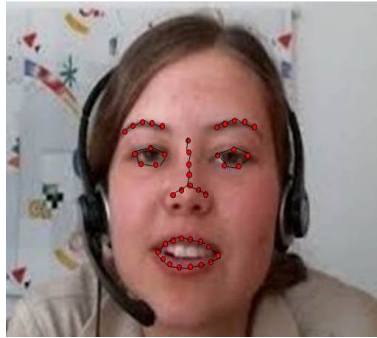


Figure 2.3: Geometric feature points.

2.2.1.3 Appearance Feature

While methods that use geometric features focus on specific facial points and FAU method focus on specific muscles, Chapter 3, 4, 5 and 6 of this thesis also examines the relationship between appearance feature and emotion, because it modelled the overall texture and general face shape/configuration.

One of the examples of appearance based features are Edge Orientation Histogram (EOH) [41]. EOH is an efficient and powerful operator, is regarded as a simpler version of Histogram of Oriented Gradients (HOG) [21] that captures the edge or the local shape information of an image. Firstly, the edge image is captured using Sobel edge detection algorithm from each frame. Secondly, the angle and intensity of the gradient function on each pixel is calculated and arranged into a polar coordinate system. Finally, the histogram from each block is normalized and concatenated into a feature vector. The process is visualized in the Figure 2.4. The division of the whole image to 4×4 and each polar coordinate system has 24 bins, resulting 384 feature vector for each image frames.

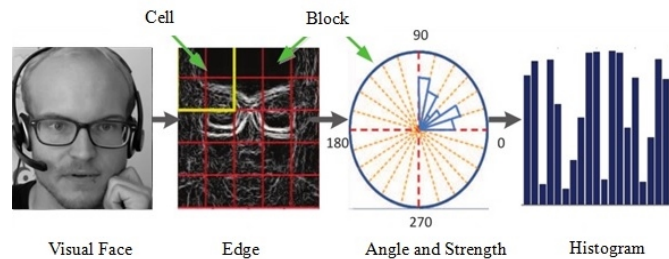


Figure 2.4: Edge Orientation Histogram feature extractor

The second common example is Local Binary Pattern (LBP), a non-parametric descriptor summarizes local texture structures of images into a set of patterns. Shan et al. [147] and Gaus et al. [41] used Local Binary Pattern (LBP) features coupled with Support Vector Machine (SVM) and Haar Wavelet Transform, respectively in their approach. The basic LBP operator labels the pixels of an image with decimal numbers, called LBP codes, which encode the local structure around each pixel, as shown in Figure 2.5

The third common example is Local Gabor Binary Pattern from Three Orthogonal Plane (LGBP-TOP) [2]. Unlike two appearance method mentioned before, LGBP-TOP is based on dynamic texture descriptors, where it applied to consecutive of frames and not standalone frames. The choice of the dynamic descriptor is based on the fact that facial appearance

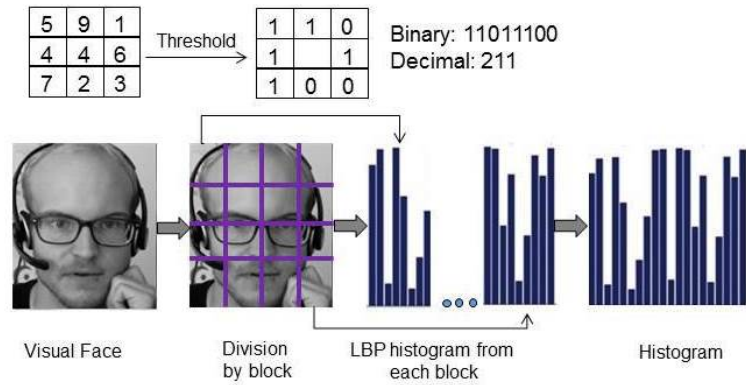


Figure 2.5: Local Binary Pattern feature extractor

changes over time and thus dynamic descriptor is more suitable than a static descriptor. Initially, video modality is split into spatio-temporal video volumes. Then, each slice of the

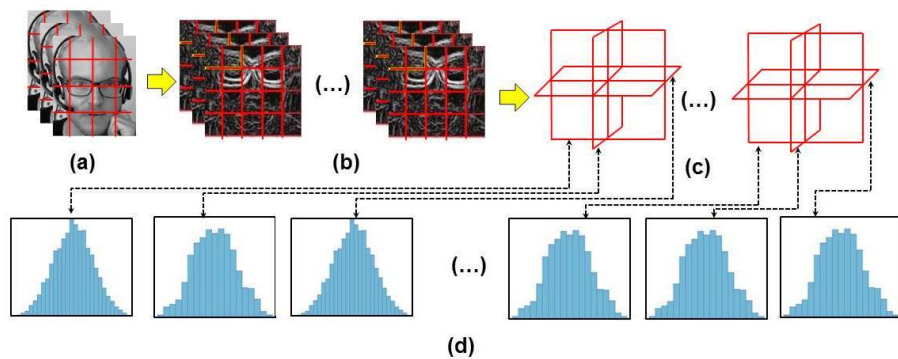


Figure 2.6 LGBP-TOP feature extraction procedure: a) original block of frames, b) Gabor magnitude responses, c) XY, XT and YT mean slices of each response and d) LBP histograms concatenated into LGBP-TOP histogram

video volume is first convolved with a bank of 2D Gabor filters, then extracted along three orthogonal planes (XY, XT and YT). The resulting Gabor pictures in the direction of XY plane is divided into 4x4 blocks. As for XT and YT plane, they are divided into 4x1 blocks. LBP feature extraction method is applied on the resulting blocks then followed by the concatenation of the resulting LBP histograms from all the blocks. Each of the step described above is shown in Figure 2.6.

2.2.1.4 Deep Learning Feature

In the last two sub-sections, hand-crafted feature extraction (geometric feature and appearance feature) from visual modality and mathematical models for facial expression is analysed. The

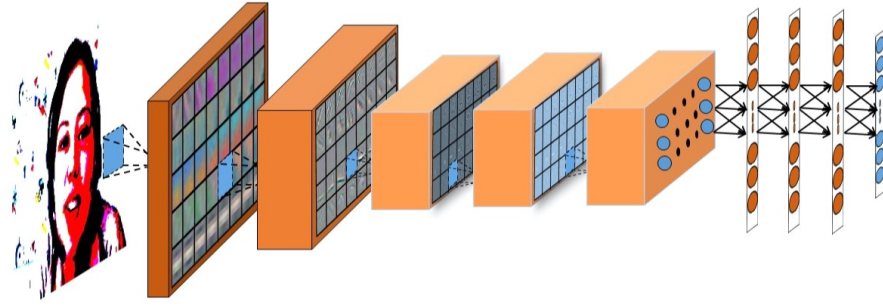


Figure 2.7: Typical example of deep learning diagram network.

recent success of deep learning algorithm enables robust and accurate feature learning on a range of computer vision applications, such as visual object recognition [88], human pose estimation [162], face verification [156] and many more. It is also due to the availability of computing power and existing big databases that allow deep learning to extract highly discriminative features from respective images. Inspired by the recent success of deep learning, emotion tasks have also been enhanced by the adoption of deep learning algorithm, e.g., convolutional neural network (CNN) [105] [104] [58] [39] [157]. However, the drawback of CNNs is that they require very large amounts of data. Therefore, in Chapter 4, transfer learning method is used to adapt pre-trained CNN models to the emotion recognition Figure 2.7 shows the overall architecture of deep learning.

To understand the architecture going on in each stage of a deep convolutional neural network, Chapter 4 of this thesis is processed based on mathematical principle is briefly explained as below:

- **Convolutional Layer:** Convolutional layers employ learnable filters which are each convolved with the layer's input to produce feature maps $Z^l(x, y, i)$ for neuron i from each convolutional layer l is computed as:

$$Z^l(x, y, i) = X^{l-1}(x, y, c) * K_i^l(x, y, c) + B_i^l \quad (2.4)$$

The input to the convolutional neural network can be represented as a X^{l-1} from the previous layer with elements $X(x, y, c)$ to indicate the value of the input unit within channel c at row x and column y . The input to the convolution is convolved with the kernel using filters K_i^l for the current layer with the same number of channels present in

X^{l-1} . Each convolved feature map in a given layer gets its corresponding bias B_i^l

- **Stride:** Strides is defined as how much the filter moves in the convolution. For normal convolution, stride = 1, which denotes that the filter is dragged across every part of the input. Consider a 7×7 with a 3×3 filter. If convolution is only applied where the kernel overlaps the input, the output is 5×5 . With stride = 2, the output is 3×3 .
- **Channel:** Channel is a conventional term used to refer to a certain component of an image. A colour image usually will have three channels – red, green and blue, where 2d-matrices stacked over each other (one for each color), each having pixel values in the range 0 to 255. A grayscale image, on the other hand, has just one channel. It is formed of 2d matrix representing an image. The value of each pixel in the matrix will range from 0 to 255 – zero indicating black and 255 indicating white.
- **Rectified Linear Unit:** A Rectified Linear Unit (ReLU) is a cell of a neural network which uses the following nonlinear activation function to calculate all convolved extracted features. ReLU is often assigned to the output of each hidden unit in a convolutional layer and the fully connected layers. The output of the ReLU $P^l(x, y, i)$ is computed as:

$$P^l(x, y, i) = \max(0, Z^l(x, y, i)) \quad (2.5)$$

- **Normalization layer:** In this process, local response normalization is used for normalizing the output of the ReLU. This step is needed to yield better generalization and introduces non-linearity that is absent in the right hand side of the ReLU responses. It can be computed as:

$$Q^l(x, y, i) = P^l(x, y, i) (\gamma + \alpha \sum_{j \in M^l} (P^l(x, y, j))^2)^{-\beta} \quad (2.6)$$

where $Q^l(x, y, i)$ computes the response of the normalized activity from the ReLU output $P^l(x, y, i)$. This is done by multiplying the output with an inverse sum of squares plus an offset γ for all ReLU outputs within a layer l

- **Max pooling layer:** The max-pooling operator computes the maximum response of

each feature channel obtained from the normalized output. It can be computed as:

$$R^l(\bar{x}, \bar{y}, i) = \max_{x, y \in M(\bar{x}, \bar{y}, i)} Q^l(x, y, i) \quad (2.7)$$

where (\bar{x}, \bar{y}) is the mean image position of the positions (x, y) inside $M(\bar{x}, \bar{y}, l)$ that denotes the shape of the pooling layer, and $R^l(x, y, i)$ is the result of the spatial pooling of the convolutional layers. Noted that, max pooling reduces the dimensionality by applying the maximum function over the input R

- **Average pooling layer:** The average pooling operator computes the mean response of each feature channel obtained from the normalised output. It can be computed as:

$$R^l(\bar{x}, \bar{y}, i) = \frac{\sum_{x, y \in M(\bar{x}, \bar{y}, i)} Q^l(x, y, i)}{|M(\bar{x}, \bar{y}, i)|} \quad (2.8)$$

- **Classification layer:** The probability of the class labels from the output of the fully connected layer is computed using the softmax activation function. It computes the probabilities of the multi-class labels using the sum of weighted inputs from the previous layer and is used in the learning process [114]:

$$y_d = \frac{\exp(x_d)}{\sum_{d=1}^D \exp(x_d)} \quad (2.9)$$

where y_d is the output of the softmax activation function for class d , x_d is the summed input of output unit d in the final output layer of the fully connected network and D is the total number of classes.

Since continuous emotion recognition has been defined as regression problem, the straight forward way is treating CNN as off-the-shelf tool, by passing each of video frame into on CNN model, such as AlexNet [88], then take the extracted feature from the last fully-connected layer, that is right before classification layer.

2.2.2 Audio Modality

Similar to visual modality, the audio modality conveys affective information through explicit (linguistic) messages and implicit (acoustic and prosodic) messages that reflect the way the

words are spoken. Therefore, Chapter 3, 4, 5, 6 is dedicated for audio modality. The most popular features from audio modality are the acoustic feature such as prosodic features (e.g., pitch-related feature, energy-related features and speech rate) and spectral features (e.g., MFCC and cepstral features) in speech-based emotion recognition [184]. Speech energy and pitch feature has been regarded as feature that contribute the most to affect recognition [90] [24].

The researchers not only analyse the message the user is saying but also how the user convey the messages. Therefore, speech-prosody analysers ignore the messages and focus on the acoustic feature that reflects emotions. Firstly, the tonal feature was extracted, then preprocessed to enhance and denoise [176]. From the features, low- level descriptor (LLDs) were extracted usually at 100 frames per second with segment sizes between 10 and 30 ms by using windowing functions. Numerous LLDs can be extracted, such as pitch (fundamental frequency F_0), energy (e.g., maximum, minimum, and root mean square), linear prediction cepstral (LPC) coefficients, perceptual linear prediction coefficients, cepstral coefficients (e.g., mel-frequency cepstral coefficients, MFCCs), formants (e.g., amplitude, position, and width), and spectrum (mel-frequency and FFT bands) [176] [140] [7] [72]

With all the acoustic feature presented above, detecting which optimal feature that contributes more to affect recognition is still an open question. However, findings confirm that acoustic feature such as mean of the fundamental frequency (F_0), mean intensity, speech rate, as well as pitch range and high-frequency energy are positively correlated with the Arousal dimension [49]. There is relatively less evidence on the connection between acoustic feature with Valence, Dominance and Liking dimension.

2.2.3 Physiological Signal Modality

Physiological signals modality can be used for affect recognition through the detection of biological patterns that are reflective of emotional expressions. These signals are collected through sensors that are attached to the body of the subject/person. Chapter 5 will explain the relationship of physiological signal modality and emotion.

Koelstra et al. [85] introduce DEAP dataset, where it consists of collection of emotion captured from physiological signal analysis such as electroencephalogram (EEG) and peripheral physiological signals. Typical physiological signals used for affect detection are electrocardiography (ECG), electromyography (EMG), electroencephalograph (EEG), skin conductance

(galvanic skin response, and electrodermal activity), respiration rate, and skin temperature. From the ECG signal, which records the electrical activity of the heart, the heart rate (HR) and heart rate variability (HRV) can be extracted, where HRV is used widely to assess mental stress [115] [60] [62] [73]. Skin conductance measures the resistance of the skin by passing a negligible current through the body. The resulting signal is reflective to Arousal dimension as it corresponds to the activity of the sweat glands [106].

EMG measures muscle activity and is known as negatively on Valence emotions dimension [106]. EEG is the electrical activity of the brain measured by electrodes connected to the scalp and possibly forehead. EEG is widely used to classify emotional dimensions of Arousal [62] [82] [64] [52], Valence [82] [64] and Dominance dimension [52] [20].

2.2.4 Text Modality

The task of automatically identifying seven basic emotions expressed in the text has been addressed by several researchers [164] [1]. It is because people can perceive emotion straight away from hearing and vision. Also, the semantics of the words can provide valuable information in emotion recognition. For example, some particular words, such as laughter can reflect the current emotion state of the person. Lexicon-based approach and word embedding are applied to take advantage of text modality. The lexicon-based approach obtains semantic information from the lexicon of emotional words to estimate emotion. On the other hand, word embedding maps the words to real number vectors in a lower dimensional space. In this thesis, a set of Bag-of-Word (BOW) representation was applied on transcripts to form frame-level 521-dimensional features. Each of text modality described above can be found in Chapter 6.

2.3 Database

Audio-Visual Emotion recognition Challenge (AVEC) is a series of competition that provide datasets and benchmark on continuous affect recognition from multiple modality [144] [143] [167] [166] [127] [165]. Earlier version of AVEC uses SEMAINE dataset [98] while latter version uses RECOLA dataset [128]. While audio and visual modality is frequently used in the earlier version of AVEC, physiological signal modality is only introduced in AVEC 2015, then used

till AVEC 2016. The recent AVEC 2017 challenge recorded the dataset 'in-the-wild' setting, and at the same time introduce new modality, that is text modality for continuous affect recognition. Various continuous affect recognition dimensions were explored in each challenge year such as Valence, Arousal, Dominance, and Liking, where the estimation of Valence and Arousal are studied in all challenges. Three different datasets discussed above will be explained thoroughly in the next subsection.

2.3.1 SEMAINE Dataset (AVEC 2014 Challenge)

In AVEC 2014 challenge dataset, video and audio modality of the subject are recorded by a webcam and microphone while performing a Human-Computer Interaction. For each recording, there are two tasks. The first task is *Northwind*, where subjects read aloud an excerpt of the fable. The length of this task recordings is between 33 seconds and 133 seconds. The other is *Freeform*, where subjects respond to one of a number of questions. The length ranges from 7 seconds to 248 seconds, which is a wide range. The recordings are split into three partitions: training, development, and a test set of 150 Northwind- Freeform pairs totalling 300 task recordings. Tasks are split equally over the three partitions.

For these challenges, the dataset is annotated in three emotion dimension, Arousal, Valence and Dominance by 3 to 5 raters. For each recording, the *gold-standard* label is defined as the average of each recording. In each emotion dimension, the range of each *gold-standard* is scaled to $[-1, 1]$. The average Pearson Correlation Coefficients (PCC) of the three affect dimensions is used as an objective function, as detailed in [166].

In this thesis, the architecture was trained on training set and tested on development and testing sets. Pearson's correlation coefficients (CORR), and Root Mean Square Error (RMSE) over all M sessions are both used as an objective function as shown in Equation 3.9 and 3.10, respectively

$$CORR = \frac{1}{M} \sum_{i=1}^M \frac{\sum_{j=1}^{N_i} (y_i^j - \bar{y}_i)(\hat{y}_i^j - \bar{\hat{y}}_i)}{\sqrt{\sum_{j=1}^{N_i} (y_i^j - \bar{y}_i)^2} \sqrt{\sum_{j=1}^{N_i} (\hat{y}_i^j - \bar{\hat{y}}_i)^2}} \quad (2.10)$$

$$RMSE = \frac{1}{M} \sum_{i=1}^M \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (y_i^j - \hat{y}_i^j)^2} \quad (2.11)$$

where N_i is the number of frames in session i ($i = 1, 2, \dots, M$). y_i^j and \hat{y}_i^j are the ground truth and predicted values for the frame j ($j = 1, 2, \dots, N_i$) in session i . \bar{y}_i and $\bar{\hat{y}}_i$ are the mean values of y_i^j and \hat{y}_i^j for session i ($i = 1, 2, \dots, M$).

2.3.2 Remote Collaborative and Affective Interaction Dataset-RECOLA (AVEC 2016 Challenge)

RECOLA database [128] provides the corpus contains audio, video and physiological signals modality (electrocardiogram - ECG, and electrodermal activity -EDA) — recorded synchronously from 27 French-speaking subjects. For the AVEC 2016 challenge, additional physiological channels derived from the ECG and EDA sensors have been added to the dataset. They are heart rate and its variability (HRHRV), skin conductance level (SCL), and skin conductance response (SCR). The subject is taken from different nationalities, such as French, Italian and German, in order to provide some diversity in the expression of emotion. The 27 subjects were divided into three groups of nine different subjects: training, development, and test set. *Gold-standard* label of the corpus has been performed by six gender balanced French-speaking assistants. Time-continuous ratings of emotion dimension of Arousal and Valence measures are recorded using 40-msec frame.

Unlike previous version of AVEC, Concordance Correlation Coefficient (CCC) as well as root-mean-square error (RMSE) is being employed as the performance measure of the latter version. It is because PCC is insensitive to scaling and shifting [175]. To alleviate this problem, CCC is applied to unites both correlation and mean squared error, and can be thought of as a CC that enforces the correct scale and offset of the outputs. As a result, CCC takes into account the effects of shifting and scaling the prediction when computing the performance [175].

The proposed framework was trained on the training set and tested on development sets for each of affect recognition. It is measured as shown in Equation 4.7

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (2.12)$$

where ρ is the Pearson Correlation Coefficient between two-time series (eg., estimation and gold-standard), σ_x^2 and σ_y^2 is the variance of each time series, and μ_x and μ_y are the mean

value of each time series.

2.3.3 The Automatic Sentiment Analysis in the Wild-SEWA (AVEC 2017 Challenge)

AVEC 2017 challenge is based on SEWA dataset 'in-the-wild' setting, where it collects spontaneous and naturalistic interactions through human-human interactions. Subjects participated in pairs and were asked to discuss the commercial product they had just viewed. However, only the behaviours of one person are recorded, which makes the recording of audio can record the sound of another interlocutor, which would influence the effectiveness of acoustic features. There are 64 German subjects in the dataset and are divided into training with 36 subjects, validation with 14 subjects and testing with 16 subjects. Apart from audio and video modality, text modality is also being introduced in these challenges. Besides the common affect emotional dimensions: Arousal and Valence, these challenge introduces another emotion dimension of likability, which presents the user's preference for the commercial product. All three emotion dimensions are annotated every 100ms and scaled into $[-1, +1]$. Concordance Correlation Coefficient (CCC) works as the evaluation metric for this challenge, same as in the previous challenge.

The performance of proposed architecture is reported based on C_{corr} [126] metric:

$$C_{corr} = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \quad (2.13)$$

where ρ is the P_{corr} between two time series (e.g: prediction and gold-standard); $\mu_{\hat{y}}$ and μ_y are the means of each time series; and $\sigma_{\hat{y}}^2$ and σ_y^2 are the corresponding variance. In contrast to the largely used P_{corr} , C_{corr} also take the bias and variance, e.g., $(\mu_{\hat{y}} - \mu_y)^2$ between the two compared series into account. Hence, the value of C_{corr} is within the range of $[-1, 1]$, where ± 1 represents perfect concordance and discordance while 0 means no concordance between two-time series.

2.4 Regression Approach

Continuous emotion recognition is usually evaluated in terms of the correlation between the learner's outputs and the target values (such as correlation) or average deviation of outputs and

Table 2.2: Comparison of each database associated in this thesis.

	SEMAINE	RECOLA	SEWA
Modality			
Video	✓	✓	✓
Audio	✓	✓	✓
Physio. Signal		✓	
Text			✓
Objective Function			
Pearson CC	✓		
Concordance CC		✓	✓
Experiment Settings			
Naturalistic	✓	✓	
in-the-wild			✓

target values (such as mean square error) [143] [71]. Therefore, continuous emotion recognition has been formulated as regression problem, to predict the value of Arousal and Valence. In this section, a set of modelling approach which is closely related to the content of this thesis is discussed.

In machine learning, a formal introduction on regression are from the simple linear equation, that is:

$$\mathbf{x} = \mathbf{w}^T \boldsymbol{\psi} \quad (2.14)$$

where \mathbf{x} and $\boldsymbol{\psi}$ is random variable, and \mathbf{w} is desired behaviour which needs to be specified. An observation or input data will represent any information with regards to the problem. After training a system, it will be deployed into target application domain, using testing data.

An input data is usually in the form of features, which are usually extracted via the procedures detailed in Chapter 2.2. Apart from features, an output labels, which usually in the form of continuous time or continuous annotation, essentially represent the targets of the linear function presented in Equation 2.14. In other words, the primary aim of regression is to learn a function mapping from *features* to the *labels*. Once this function is learned, the inputs should enable the accurate prediction/estimation of the outputs.

Given the features \mathbf{x}_i and target values for each features, \mathbf{y}_i . In regression setting, Equation 2.14 become:

$$\mathbf{y}_i = \mathbf{w}^T \mathbf{x}_i \quad (2.15)$$

In Equation 2.15, the aim is to obtain the best \mathbf{w} which map the input \mathbf{y}^* as close as possible

to the given outputs \mathbf{y}_i . Having learnt the correct \mathbf{w} , the next step is to predict/estimate the \mathbf{y}^* in a given test input \mathbf{x}_i^* . Most of the state-of-the-art regression techniques employed are based on optimising this simple function, ranging from simple Linear Regression to Support Vector Machine [27].

2.4.1 Support Vector Regression

Support Vector Regression (SVR) is an application of Support Vector Machine (SVM) where SVMs, in a classification problem, provide good generalisation to generate a hyperplane that separates two datasets whereas SVR finds a hyperplane that accurately predicts the distribution of the original data.

Suppose that a training data set is given by

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \subset x^p \times \mathfrak{R} \quad (2.16)$$

where x_i denotes the time index, y_i denotes the label and l is the feature length. The goal of ϵ -SVR is to find a function $f(x)$ such that the ϵ -deviation between the actual target value y_j and the predicted target y is as small as possible [171]. The function f is described as follows:

$$f(x) = wx + b \forall w \in X, b \in \mathfrak{R} \quad (2.17)$$

where w is the hyperplane solved by SVR. By solving following quadratic optimization problem:

$$\begin{aligned} & \min \frac{1}{2} \|w\|^2 \\ & \text{subject to } \|y_i - (wx_i - b)\| \leq \epsilon \end{aligned} \quad (2.18)$$

where $\epsilon \geq 0$ denotes the maximum deviation between the actual and predicted target. However, in most application, there exists noise in the system. Thus, slack variable ξ_i, ξ_i^* is introduced. Finally, SVR is formulated as the minimisation of the following functional:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \text{ subject to } \begin{cases} y_i - wx_i - b \leq \epsilon + \xi_i \\ wx_i + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (2.19)$$

In Equation 2.19, each feature has its own corresponding ξ and ξ^* values which are used to determine whether the training instance falls outside the scope of ϵ . The penalty parameter $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ϵ are tolerated.

The optimization problem in Equation 2.19 can be solved by the Lagrange multiplier technique. Finally, the regression function of $f(x)$ is given by;

$$f(x) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) LM(x_i, x) + b \quad (2.20)$$

where $LM(x_i, x) = \varphi^T(x_i)\varphi(x)$ is known as the kernel function. A number of coefficients, $\alpha_i - \alpha_i^*$ have nonzero values and the corresponding training instances are known as support vectors that have approximation errors equal to or larger than the error level ϵ . Chapter 3, 4, 5, 6 will mainly use SVR as machine learning approach to map features with affect ratings.

2.4.2 Linear Regression

With linear regression, it is assumed that dependent and explanatory variables have a linear relationship, which can be expressed as a linear combination of random variables, i.e.:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_i x_i + \epsilon \quad (2.21)$$

where x_i is the given features, y_i is the target value for each feature and ϵ is the residual or noisy factors. The parameter β is called regression coefficient and it can be estimated using least squares regression.

Even though its called *linear* regression, it does not mean that it is necessarily a straight line. It can be polynomial or curve lines. For example, the line:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon \quad (2.22)$$

can be rewritten in polynomial form as:

$$x_2 = x_1^2$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

In higher dimension, linear regression is actually dealt with planes, hyperplanes, and so on. In this thesis, linear regression is employed as a method for the fusion of multiple modality.

2.4.3 Recurrent Neural Network-Long Short Term Memory

Neural Networks Primer by Maureen Caudill [13] defines a neural network as “a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs”. A neural network is organised in layers, which consists of an input layer, number of hidden layers, and an output layer. Layers are grouped by a number of interconnected nodes and nodes contain activation functions, as shown in Figure 2.8. From this Figure, all inputs are independent of each other. However, in Recurrent Neural Network (RNN), it addresses this issue by predicting every next value based on the previous information. RNN uses loops to capture information into its memory and passes prior knowledge to predict the future value, as shown in Figure 2.9

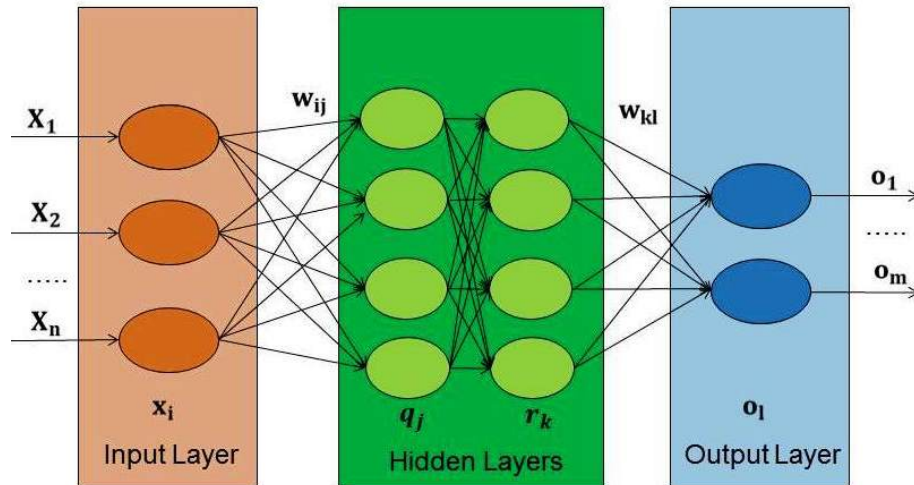


Figure 2.8 A typical neural network. The variable x_i is the input value, o_l is the output, q_j and r_k are the hidden variables, and w is the weight for each connection

However, the main problem in RNN is that it is not capable of learning long-term dependencies. In continuous emotion recognition, capturing long-term dependencies are necessary because emotion is typically evolved over time [178]. In RNN, only recent information is needed to perform the prediction. As the length of sequence grows, it will be more difficult for RNN to capture and connect the information. The introduction of Long Short Term Memory (LSTM) by Hochreiter and Schmidhuber [63] and improved by Felix Gers in 2000 [42] over-

come the long-range dependencies problem. LSTM has special properties such as input gates and forget gates, which allow for a better control over the gradient flow and enable better preservation of “long-range dependencies”. LSTM also used heavily in Chapter 5.

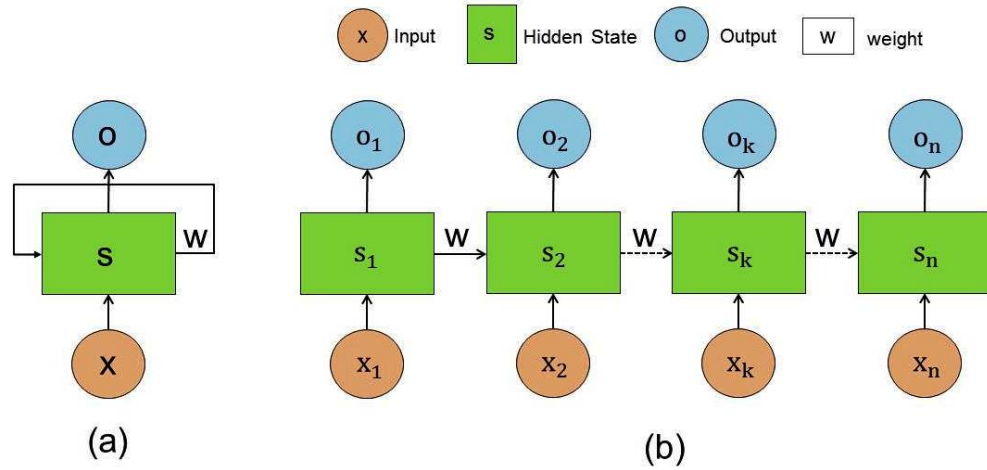


Figure 2.9: (a) is a traditional RNN. (b) is this RNN unrolled into a chain.

Figure 2.10 shows the structure and function of LSTM. x_t denotes the input vector at time t and h_t is the output of current block. LSTM uses a cell state c_t to denote the memory from current block and makes three gates: *input*, *forget* and *output* to control the cell state. There are two operations: $+$ denotes the concatenation and \times is the element-wise multiplication.

The first step of LSTM is to decide which previous information needs to be erased or to keep. To erased the previous information, the *forget gate* is enabled. A *gate* is a filter to add or remove information to the cell state. It is in the form of a sigmoid layer, and the output is between zero and one. If the output value is zero, it means no information passes, if the value is one means the gate keeps all the information. The formula to update forget gate f at time t is shown in Equation 2.23, where W_{xf} , W_{hf} , and W_{cf} are the weights for input x , output h and cell state c and σ for sigmoid function.

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_i) \quad (2.23)$$

The second step is to choose which new information is added to the cell. LSTM combines two values to update the cell state. One value uses the input gate layer to decide how much scale needed to update the state, as in Equation 2.24 and the other creates a tanh layer to generate a new candidate value to be added to the cell state, as in Equation 2.25. The cell

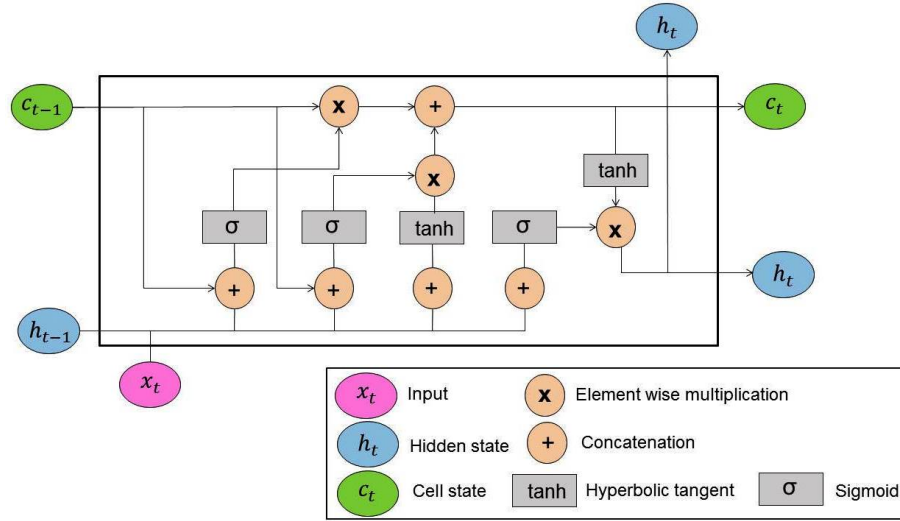


Figure 2.10: A simple LSTM block

state is put through a \tanh layer so that only a final value between -1 to 1 will be generated.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (2.24)$$

$$c_{temp} = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (2.25)$$

To calculate a new cell state, the previous state is multiplied with the forget gate in order to remove the memory and then add the candidate value, as in Equation 2.26

$$c_t = f_t * c_{t-1} + i_t * c_{temp} \quad (2.26)$$

Simultaneously, the output gate layer chooses which part of cell state to output using a sigmoid gate. Then, a \tanh layer on the cell state is implemented to push the range to be $[-1, 1]$, as in Equation 2.27, then multiply it by the output of the sigmoid layer, as in Equation 2.28.

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (2.27)$$

$$h_t = o_t * \tanh(c_t) \quad (2.28)$$

2.4.4 Time Delay Neural Network

Time Delay Neural Network (TDNN) is a special type of neural network whose primary purpose is to work on sequential data. TDNN, like other neural networks, consists of nodes organised into three layers of clusters including input layer, the output layer, and the hidden layer which handles the manipulation of the input through filters as in Figure 2.11(a). As

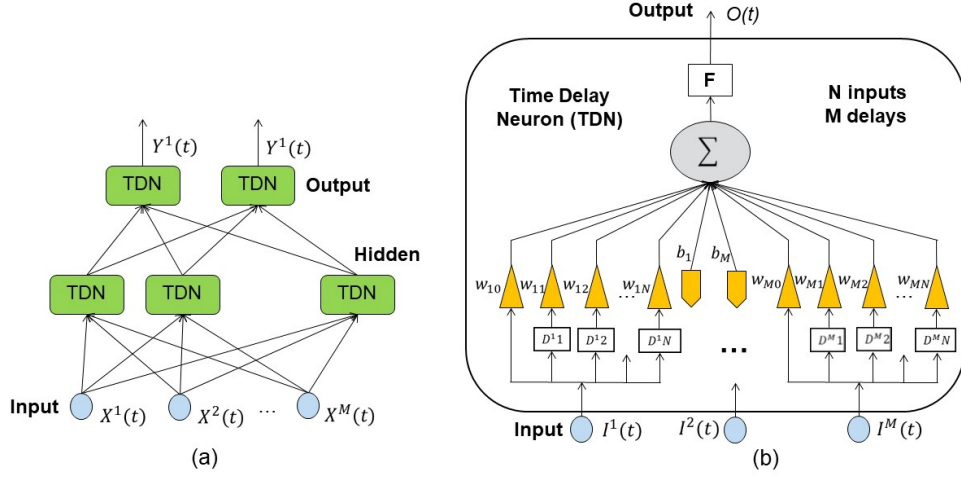


Figure 2.11 : (a) Overall architecture of the TDNN. (b) Single TDNN with M inputs and N delays for each input at time t . D^i_d are the registers that store the values of delayed input $I^i(t-d)$

in Figure 2.11(b), the hidden layer is responsible for the processing of the input signal by calculating the weighted sum of input signals with the help of transfer function.

From Figure 2.11(b), a single TDNN has M inputs $I^1(t), I^2(t), \dots, I^M(t)$ and one output $O(t)$ where these inputs are time series with time step t . At each input $I^i(t)$ and $i = 1, 2, \dots, M$, there are two properties: bias value b_i and delay N . The delay N indicate as $D^i_1, D^i_2, \dots, D^i_N$ storing the previous inputs $I^i(t-d)$ with $d = 1, \dots, N$ and also N unknown weight $w_{i1}, w_{i2}, \dots, w_{iN}$. F indicate the transfer function, which is in this case, nonlinear transfer function. A single TDNN node can be represent as in Equation 2.29:

$$O(t) = f\left(\sum_{i=1}^M \left[\sum_{d=0}^N I^i(t-d) * w_{id} + b_i\right]\right) \quad (2.29)$$

From Equation 2.29, both the inputs at current time step t and previous time step $t-d$ with $d = 1, \dots, N$ contribute to the overall outcome of the neuron. A single TDN can be used to model the dynamic nonlinear behaviour that characterises series inputs.

The network is classified by comparing the value of the weighted sum in the input signal

and the threshold value while using the activation function to convert a neuron's weighted input to its output activation. To achieve sequential nature, a set of delays are added to the input so that the data are represented at different points in time such as audio files or sequences of video frames

An essential feature of TDNN is the ability to express relations between inputs in time, which can be used to recognise patterns between the delayed inputs. The main difference between TDNN and LSTM is the flow of the data. TDNN is implemented as feed forward neural network, which means the flow of the data is in only one direction, if forward from the input nodes through the hidden nodes and to the output nodes. Unlike LSTM, there are no cycles or loops in the TDNN network. Alongside LSTM, TDNN also implemented in Chapter 5.

2.4.5 Kalman Filter

One approach of recognising continuous emotion recognition automatically is by considering emotion trajectories as time series and apply methods from time series analysis. For a real-time emotion recognition, Kalman Filter is exploited. In this thesis, it leverages Kalman Filter by estimating the emotional state x as a function of time from the information z from the respective modality using the standard state space framework. The state transition equation models in Equation 2.30 is the time-varying nature of the emotional states, where A is the transition matrix and $w(k)$ is the zero-mean process noise in the system

$$x(k+1) = Ax(k) + w(k) \quad (2.30)$$

The measurement equation relates how the measures (z) from the individual measurement channels relate to the underlying emotional state (x):

$$z(k) = Cx(k) + \beta + v(k) = \begin{bmatrix} z_{audio} \\ z_{video} \\ z_{physiological} \\ \cdot \\ \cdot \end{bmatrix} \quad (2.31)$$

The measurement matrix C is defined as the underlying emotional states to the measurements and $v(k)$ is the zero-mean measurement noise term. In practice, it has been decided that measurement noise was nonzero. Therefore bias term β has been added to the model, as in Equation 2.31.

To determine system matrices (A , C) and noise terms (w , v), firstly the gold standard is defined as x , and z is defined as the corresponding measurement from the individual modality:

$$\begin{aligned} X_{1,N} &= [x_1, \dots, x_N] \\ Z_{1,N} &= [z_1, \dots, z_N] \end{aligned} \quad (2.32)$$

In many cases, the different z 's may correspond to the different type of features, even though it comes from the same modality, such as geometric features and appearance features, which comes from video modality. For example, x_m would correspond to a scalar value representing the emotional state (Arousal or Valence), while z_m would correspond to a vector of emotional state measurements corresponding to each of the applied feature extraction methods. From here, the state transition matrix (A) and the variance of the process noise (Q) is found by using Equation 2.33:

$$\begin{aligned} A &= (X_{2,N}, X_{1,N-1}^T)(X_{1,N-1}, X_{1,N-1}^T)^{-1} \\ Q &= \text{cov}(w, w) = \text{cov}(X_{2,N} - AX_{1,N-1}) \end{aligned} \quad (2.33)$$

Using this following substitution:

$$\begin{aligned} \bar{X}_{1,N} &= \begin{bmatrix} X_{1,N} \\ 1_{1 \times N} \end{bmatrix} \\ \bar{C} &= \begin{bmatrix} C & \beta \end{bmatrix} \end{aligned} \quad (2.34)$$

The measurement equation can be written as follows:

$$Z_{1,N} = \bar{C}\bar{X}_{1,N} + v(k) \quad (2.35)$$

It enables a convenient factor for deriving the measurement matrix (C), the bias term (β) and

the variance of the measurement noise (R):

$$\begin{aligned}\bar{C} &= (Z_{1,N}, \bar{X}_{1,N}^T)(\bar{X}_{1,N}, \bar{X}_{1,N}^T)^{-1} \\ R &= \text{cov}(v, v) = \text{cov}(Z_{1,N} - CX_{1,N} - \beta)\end{aligned}\tag{2.36}$$

Using the matrices computed in the system identification phase (A, C, Q, R) as initialisation, the Kalman filter performs two operations at each time step: (i) the time update and (ii) the measurement update. The time update takes the current estimate of the states (x_{k-1}) and the error covariance matrix (P_{k-1}) and projects them forward in time by one step using the following Equations 2.37 and 2.38:

$$x_k^- = Ax_{k-1}\tag{2.37}$$

$$P_k^- = AP_{k-1}A^T + Q\tag{2.38}$$

The new estimates of the state and error covariance are then updated using the newly observed measurements (z_k) using the measurement update Equations 2.40 and 2.41:

$$K_k = P_k^- C^T (CP_k^- C^T + R)^{-1}\tag{2.39}$$

$$x_k = x_k^- + K_k(z_k - Cx_k^-)\tag{2.40}$$

$$P_k = (I - K_k C)P_k^-\tag{2.41}$$

The Kalman gain matrix (K_k) determines how much the new measurements contribute to the state estimate. This process is repeated until all of the measurements have been observed. The Kalman filter provides a useful way for fusing each modality per time step, and also models the time-varying nature of the emotions to further improve system performance. Chapter 6 will implement Kalman Filter as fusion approach.

2.4.6 Genetic Programming

The Genetic Programming (GP) [86] is one of the most influential machine learning methods inspired by real-world biological systems. By mimicking Darwinian evolution, it does this by randomly generating a population of tree structures and then breeding together the best performing trees to create a new population. The same process is iterated until the population

contains programs that solve the task well. In this thesis, GP is used heavily in Chapter 6.

In GP, symbolic regression is employed to discover mathematical expressions of functions that can fit the given data based on the rules of accuracy, simplicity, and generalisation [95]. Unlike Kalman Filter, where the user must estimate the parameters from the data, symbolic regression automatically evolves both the structure and find the best parameters of the mathematical model from the data. This allows it to both select the features of the model and capture non-linear behaviour. In continuous emotion recognition, GP is employed to investigate the non-linear behaviour in affect dimension.

Symbolic regression models are typically in the form of:

$$y^* = f(x_1, \dots, x_M) \quad (2.42)$$

where y is an output variable, y^* is the model prediction of y and x_1, \dots, x_M is the features related to y . f is a symbolic non-linear function or a collection of non-linear functions. An example of symbolic regression model is:

$$\bar{y} = 0.32x_1 + 0.34(x_1 - x_4) + 1.43x_3^2 - 3.31 \cos(x_2) + 0.22 \quad (2.43)$$

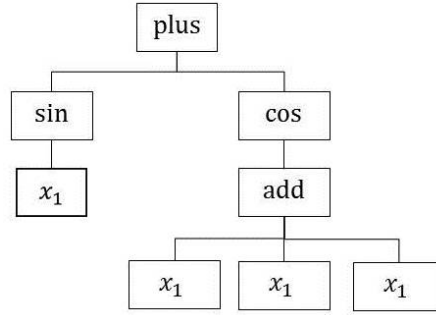
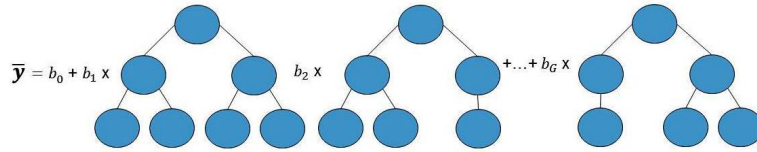
The model presented contains both linear and non-linear terms. Also, the structure and parameter of these terms are automatically determined by the symbolic regression algorithm. Hence, it can be seen that symbolic regression provides a simple approach to non-linear predictive modelling. In the case of continuous emotion recognition, y is the estimation of affect, and the x_1, \dots, x_M is the features related to the modality [146].

A special type of GP called multi-gene genetic programming (MGGP) which uses a modified GP algorithm to evolve data structures that contain multiple trees (genes). An example of a tree representing a gene is shown in Figure 2.12 The structure of MGGP models is illustrated in Figure 2.13

The prediction of \bar{y} training data is given in

$$\bar{y} = b_0 + b_1 \mathbf{t}_1 + \dots + b_G \mathbf{t}_G \quad (2.44)$$

where \mathbf{t}_i is the $N \times 1$ vector of output from the i th /gene comprising a multigene individual.

Figure 2.12: Example of a tree structure representing the model term $\sin(x_1) + \cos(3x_1)$ Figure 2.13: Example of a tree structure representing the model term $\sin(x_1) + \cos(3x_1)$

\mathbf{G} is defined as $N \times (\mathbf{G} + 1)$ gene response matrix as in Equation 2.45

$$\mathbf{G} = [\mathbf{1}t_1 \dots t_G] \quad (2.45)$$

where $\mathbf{1}$ refers to a matrix $N \times 1$ columns of ones used as bias input. Then, Equation 2.44 can be written as:

$$\bar{\mathbf{y}} = \mathbf{G}\mathbf{b} \quad (2.46)$$

The least square computation of the coefficients $b_0, b_1, b_2, \dots, b_G$ from Equation 2.44 can be formulated as $(G + 1) \times 1$ vector and can be computed from the training data as:

$$\mathbf{b} = (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{y} \quad (2.47)$$

In practice, the columns of the gene response matrix \mathbf{G} may be collinear, due to duplicate genes in an individual. Therefore, instead of standard matrix inverse, Moore-Penrose pseudo-inverse is used to solve the collinear problem. The drawbacks of GP that it tends to produce complex mathematical expression. It is mainly because the way the input communicate to the desired system is through the fitness function. The mathematical expression could be enormously complex, inefficient or incomprehensible from an emotion recognition point of view. Therefore, designing proper fitness function is need to be carefully taken to alleviate

this problem.

2.5 Chapter Summary

This following Chapter provides a thorough examination of the background which relates to continuous affect recognition covering the general directions employed by this thesis, as well as referring to various related work. To summarize research gap in Chapter 2, it is important to address each features that represent each modality, in order to mapped the feature into emotion space. Also, the key question is, which machin learning models that suitable in order to learn the feature and can mapped it to affect scales.

In more details, Section 2.2 presents a set of methodologies aiming at recognising continuous emotion dimension by further utilising multiple modalities, such as audio, video and so on. In this Section, feature extraction methodology is investigated then evaluated in terms of predicting continuous interest. The database used to apply novel methodologies is presented in Section 2.3. Two type of datasets is employed: naturalistic setting and *in-the-wild* setting. The other focus on the problem of continuous emotion is regression approach, which is presented in Section 2.4. With the aim of estimating continuous affect value (Arousal and Valence) from observable features, viewing the problem as regression is a promising approach. SVR and Linear Regression is adopted for direct estimation of the Arousal and Valence values. Temporal information is also taken into consideration. Therefore, LSTM, Kalman Filter and TDNN are also applied because it can handle sequential information. Lastly, GP is employed to investigate potential non-linearities occur in continuous emotion recognition.

Chapter 3

Continuous Affect Estimation based on Wavelet Filtering and PLS Regression

This Chapter explores human emotional states in terms of arousal and valence. Two modalities have been considered, audio and video modality. Initially, features extraction is employed from respective modality. Then each of the features is smoothed by Wavelet Filtering method. It will be feed into Partial Least Square regression to estimate arousal and valence. The proposed approach is compared with the state-of-the-art and produce a good performance.

3.1 Introduction

A persons emotional state can be presented in the forms of three modalities, audio, video and biosignals modality. Video modality often refers to facial expressions, body or hand movement. Audio modality includes speech rate, pitch range and vocal duration. Physiological data such as blood pressure, heart rate and skin conductance. This Chapter focuses mainly on using audio modality and video modality for emotion recognition.

In order to analyse the emotional state of a person, the first step needed is to make feature representation of each modality. A feature representation is encoded information generated from raw data that can be directly used in machine learning tasks. A video can be seen as a facial expression image sequences, while an audio can be seen as raw acoustic file sampled by

time rate.

When extracting features from respective modality, sometimes the feature vector is corrupted by several factors, such as background noise, different lighting condition and so on. Although most of the audiovisual recordings used are made under quiet, laboratory conditions, in the real-world application, it may be deployed in environments with cluttered backgrounds and various levels of noise. It is therefore important to remove such interference and to produce enhanced noise-free feature vector in the design of emotion recognition system. The temporal noise on feature sequence is the one to remove in the feature space. However, very little research has been conducted in this area. Only a few works have analysed noise effects in emotion recognition, but it has been limited to the audio modality. Schuller et al [139] investigated an additive white noise in speech signals, while You et al. [183] employed Lipschitz embedding in corrupted noisy speech.

Unlike the aforementioned works, in this Chapter, noisy audio and video features are investigated to determine if noise-cancellation mechanisms in feature space should be considered in the design of emotion recognition system. As for the video, three type of feature extractor have been adopted: EOH, LBP and LPQ technique. In audio modality, acoustic features such as mel-frequency cepstrum are employed. Then, each of the features selected is feed into digital filtering technique, in order to remove irrelevant noise inside the features. The motivation is the temporal noise (on feature sequence) is the one to remove in the feature space. Lastly, Partial Least Square regression was adopted as machine learning approach to estimate each of emotion dimension. In summary, the main contributions of Chapter 3 are as follows:

- Showing the effectiveness of Wavelet Transform (WT) based digital filtering method for removing the extraneous noise in feature level.
- Comprehensively analysing the robustness of WT and Partial Least Squares (PLS) regression to build a better automatic affective dimension recognition system.

A reminder of Chapter 3 is organized as follows. Section 3.2 present early literature review in the context of continuous emotion recognition. Section 3.3 outline basic methodology employed, while in Section 3.4 the experimental results is briefly explained in the system. Finally, Section 3.6 conclude Chapter 3, along with its limitation and future works.

3.2 Related Database and Regression Technique

There has been a lot of research efforts lately on identifying the continuous emotion. Many challenges have been organized, such as Audio-Visual Emotion Challenge (AVEC) to attract researchers in developing an automatic audio-visual depression and emotion analysis [143] [167] [166].

However, such approaches mention above have failed to address noisy feature exists in continuous emotion recognition. These noisy features occur mainly due to the large variance in emotional expressions. Also, since feature extracted from multiple modalities, noise also present in the system during audiovisual recording. Background noise, recording equipment, transmission channel during recording may effect on the features. The performance of machine learning system is generally negatively affected by these effects, and cause a mismatch in feature statistic during the recognition process.

In continuous emotion recognition, noise cancellation often evaluated in decision level, where the estimation of affect is already generated. Gupta et al., [51] presents two layer noise smoothing system. In the first layer, initial predictions are being smoothed by a moving average filter, then fused via linear regression. In the second layer, fusion output is further processed using temporal regression. Chao et al., [14] utilizes Deep Belief Network (DBN), where each of the features become an input to Restricted Boltzmann Machine (RBM) with one hidden layer. Temporal pooling function is added after hidden layer to undergone feature pooling. The pooled features are fed into Linear Regression to estimate affect dimension. Initial emotion estimation from different modality as well as emotion temporal context information simultaneously is combined for final estimation of affect. Few works touch on noise reduction on the feature level, where feature selection is undergone in [111]. By employing correlation-based measure on top of the features, the robustness of the system increase to time delayed labels. The system achieved the best performance on AVEC 2012 edition. In a same AVEC edition [132], they use statistical method to make initial affect estimation at each moment, then uses particle filter to make final affect estimation.

The synthesis of noise cancellation in feature level remains a major challenge in building robust continuous emotion system. Therefore, in this Chapter, a digital filtering method is designed by employing Wavelet Transform (WT) based digital filtering technique. WT is employed on top of the feature selected to remove an unnecessary noise component in feature

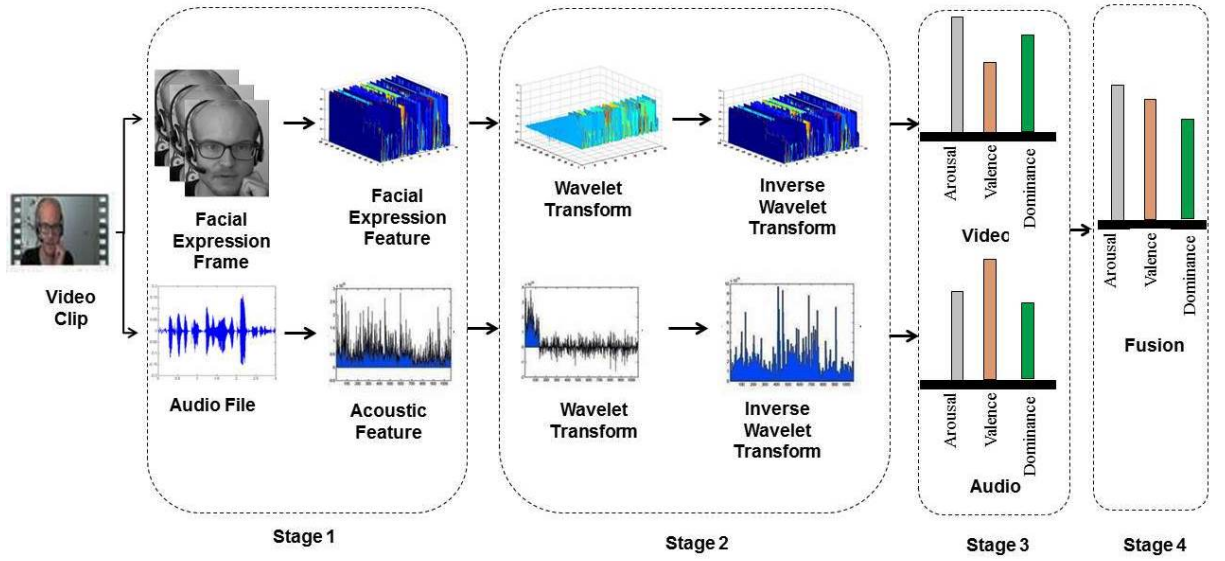


Figure 3.1: Overview of the proposed automatic affective dimension recognition system

space then integrated it in automatic continuous emotion recognition system.

3.3 Methodology

The methodology proposed consists of four stage, as illustrated in Figure 3.1. For each video clip, video and audio modality is dealt independently. Feature extraction is adopted in Stage 1. In this stage, feature extraction is implemented towards each of the facial expression frame and acoustic file. In Stage 2, wavelet transform based digital filtering technique is implied towards each feature. The motivation is driven by the fact that affect related features should also change slowly as in the affective dimensions. The high-frequency components might be noise that is irrelevant to the affect label. Stage 3 is undergone by feeding to features onto Partial Least Square regression. Then, the initial estimation of affect from video and audio is enhanced by applying low pass filtering. Finally, in Stage 4, the initial estimation from video and audio modality was combined to improve the overall performance using decision fusion approach. Each of the Stage will be discussed thoroughly in the following section.

3.3.1 Stage 1: Feature Extraction

Each of the video clip provided is converted to image frames, in order to capture the facial expression recorded in video data. Then, three dynamic features are extracted respectively,

which are detailed in the following subsection.

3.3.1.1 Edge Orientation Histogram

EOH, an efficient and robust operator, is simpler version of HOG [21] that captures the edge or the local shape information of an image. Firstly, the edge image is captured using Sobel edge detection algorithm from each frame. Secondly, the angle and intensity of the gradient function on each pixel are calculated and arranged into a polar coordinate system. Finally, the histogram from each block is normalised and concatenated into a feature vector. The process is visualised in the Figure 3.2.

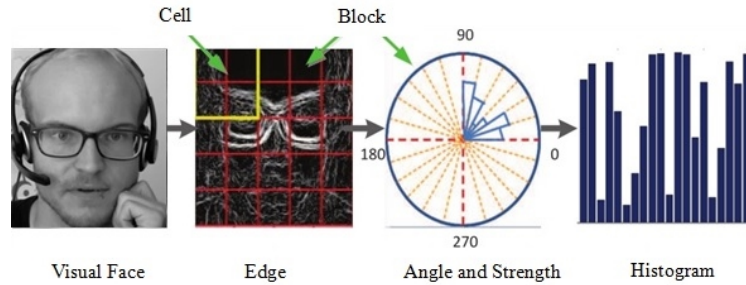


Figure 3.2: Edge Orientation Histogram feature extractor

3.3.1.2 Local Binary Pattern

LBP, a non-parametric descriptor summarises local texture structures of images into a set of patterns. The basic LBP operator started by labelling the pixels of an image with decimal numbers, namely LBP codes. It is encoded with the local structure around each pixel, as shown in Figure 3.3

3.3.1.3 Local Phase Quantization

LPQ operator, proposed by Ojansivu and Heikkila [113] initially as a texture descriptor. Then, it is further used in emotion recognition [26] for encoding the shape and appearance information. Based on the blur invariance property of the Fourier phase spectrum, it uses the local phase information extracted using the 2-D short-term Fourier transform (STFT) computed over a rectangular neighbourhood at each pixel position of the image. Only four complex coefficients are considered, corresponding to 2-D frequencies of an image.

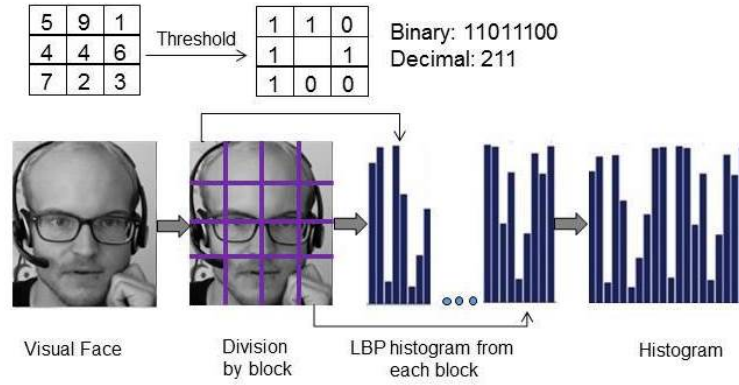


Figure 3.3: Local Binary Pattern feature extractor

3.3.1.4 Audio Feature Extraction

For audio modality, a set of 2 268-length openSMILE [36] feature has been fully utilized. The acoustic feature sets are arranged in three segmentation settings: short (3-second overlapping frames with 1-second shifts), long (20-second overlapping frames with 1-second shifts) and voice-activity detected (VAD) segments [35]. VAD segment is employed to split the clip when there is a pause for more than 200 ms. Each of the segmentation setting consists of various acoustic LLDs, such as relative spectral (RASTA) MFCCs, spectral energies, and voicing/unvoiced related features [166].

3.3.2 Stage 2: Wavelet Filtering

The methods proposed in this Chapter is based on obtaining the noise by separating the features from each time series into two: high and low-frequency components. At high frequency, the WT can capture discontinuities, ruptures and singularities in the original feature. The target is to suppress the high-frequency part, due to faulty information during audiovisual recording.

To perform WT, the statistical features extracted from Stage 1 is treated as signal per frame. Suppose a features, such as EOH is composed of a signals with a small discontinuity. WT performs as a filter when decomposing the raw signals to approximation coefficients and detailed coefficients where the high frequency signals clustered in detailed coefficients. Since the high frequency signal is contaminated with noises, approximation coefficients are used for further analysis.

3.3.2.1 Haar Wavelet Transform

For features, such as EOH, $X = (x_1, x_2, x_3, \dots, x_N)$ with N values, it can be decomposed into two parts s and d with the length of $N/2$ each based on Haar wavelet transform as the following equations:

$$s_k = \frac{x_{2k-1} + x_{2k}}{2}, k = 1, 2, \dots, N/2 \quad (3.1)$$

$$d_k = \frac{x_{2k-1} - x_{2k}}{2}, k = 1, 2, \dots, N/2 \quad (3.2)$$

s_k is called approximation of the signal that represents the low frequency part of the signal while d_k is called details of the signal that represents the high frequency of the signal. After this first level Haar wavelet transform decomposition, the signal $X = (x_1, x_2, x_3, \dots, x_N)$ will be transformed to:

$$(s_1, s_2, \dots, s_{N/2} | d_1, d_2, \dots, d_{N/2}) \quad (3.3)$$

The original signal X can be fully reconstructed from s_k and d_k by using:

$$s_k + d_k = \frac{x_{2k-1} + x_{2k}}{2} + \frac{x_{2k-1} - x_{2k}}{2} = x_{2k-1}, k = 1, 2, \dots, N/2 \quad (3.4)$$

and

$$s_k - d_k = \frac{x_{2k-1} + x_{2k}}{2} - \frac{x_{2k-1} - x_{2k}}{2} = x_{2k}, k = 1, 2, \dots, N/2 \quad (3.5)$$

3.3.2.2 Haar Wavelet Filtering

To remove the high-frequency components of the signal, after performing wavelet transform, low-frequency element s will be kept and high-frequency element d will be replaced by zeros. In this way, the reconstructed signal will lose the high-frequency components. The low-frequency part s can be decomposed further by Haar wavelet transform to remove more high frequencies. It was called level 2 wavelet transform. It can be carried this way further for level 3, level 4 decomposition. Figure 3.4 shows the wavelet transform process on EOH feature with 4 level decomposition. It can be seen, through reconstruction step by step, the final reconstructed signal generated is smoother along the timeline. In the case of emotion recognition, smooth and simple features are matching the slow change property of the real affective dimensions.

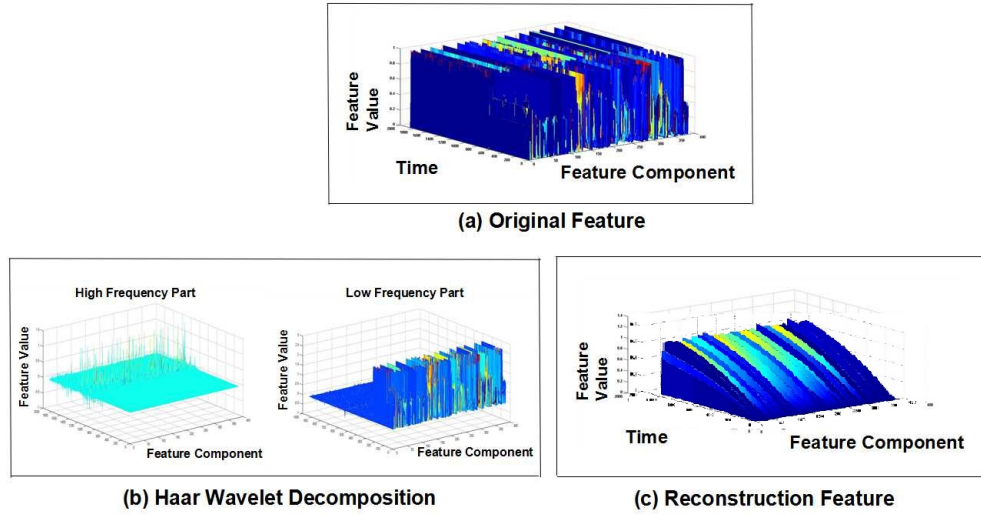


Figure 3.4 Haar Wavelet Transform filtering on selected feature space. 1840 frames, selected components and decomposition level=4. (a). Original feature vectors (b). Low and high frequency parts of the feature vectors in wavelet space (c). Filtered feature vectors.

3.3.3 Stage 3: Modeling Approach

The first step in this experiments consists of initial estimation based on each and every modality. Let $R = \{Valence, Arousal, Dominance\}$, represent each of the affect dimensions, F represent the feature vector of each audio or video. Given a set of input features $x_F = [x_{1_f}, x_{2_f}, \dots, x_{n_f}]$ where n is the training sequence with length n and $f \in F$, machine learning technique L_r is trained, in order to predict the relevant dimension output, $y_d = [y_1, \dots, y_n]$ where $d \in R$ such that:

$$f_r : x \rightarrow y_r \quad (3.6)$$

The goal of this step is to provide a set of initial estimation from machine learning point of view. As for machine learning, Partial Least Square Regression is employed to estimate each and every affect dimension.

3.3.3.1 Partial Least Square Regression

Partial Least Squares (PLS) regression is a statistical the algorithm that bears some relation to principal components regression. By projecting the response and independent variables to another common space, it builds a linear regression model. More specifically, PLS tries to

seek fundamental relations between two matrices (response and independent variables), i.e. a latent variable way to model the covariance structures in these two spaces. It is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among independent variable values.

PLS regression works as follows: consider a PLS algorithm is to model the relationship between two data sets (blocks of variables). As in this experiments, the first block is feature vector training and the second block is ground truth label training. Both of the blocks are taken from the training set. Denoted by $S \subset R^N$ an N -dimensional space of variables representing the first block and similarly by $B \subset R^M$ a space representing the second block of variables. PLS models the relationship between each two blocks utilising score vectors. After observing n data samples from each block of variables, PLS decomposes the $(n \times N)$ matrix of zero-mean variables \mathbf{S} and the $(n \times M)$ matrix of zero-mean variables \mathbf{B} in the form of Equation:

$$\begin{aligned} S &= TJ^T + E \\ B &= UQ^T + F \end{aligned} \tag{3.7}$$

where S is an $(n \times j)$ matrix of predictors and B is an $(n \times q)$ matrix of responses. T and U are two $n \times l$ matrices that are projections of R (scores, components or the factor matrix) and projections of B (scores); J and Q are $(j \times l)$ and $(q \times l)$ orthogonal loading matrices; matrices E and F are the error terms, assuming to be independent and identical normal distribution. Decompositions of S and B are made so as to maximize the covariance of T and U .

3.3.3.2 Filtering on Initial Estimation

After performing Haar wavelet transform on the features, and use PLS regression as machine learning to estimate the continuous affect, it will give estimation label as the output. Since it is a regression problem, and it requires the estimation of continuous affect labels per frame, the smoothing on initial estimation label has been carried out using simple low pass filtering, in order to further enhance the results of initial estimation.

3.3.4 Stage 4: Final Estimation using Decision Fusion

The goal of decision fusion stage aims to combine multiple decisions into a single and consensus one [150]. Here, linear opinion pool method is being employed due to its simplicity and

straightforward algorithm [8].

$$Fuse_{linear}(\hat{x}) = \sum_{i=1}^l \alpha(i) D_i(\hat{x}) \quad (3.8)$$

where \hat{x} is a testing sample and $Fuse_i(\hat{x})$ is the i_{th} decision value ($i = 1, 2, \dots, l$) while $\alpha(i)$ is its corresponding K weight which should satisfy $\sum_{i=1}^K \alpha(i) = 1$

3.4 Experimental Evaluation

The proposed system was trained on training set and tested on development and testing sets from dataset of AVEC 2014, where the level of affect has to be estimated for each frame of the recording. The Pearson's correlation coefficients (CORR), and Root Mean Square Error (RMSE) over all M sessions are both used as an objective function as shown in Equation 3.9 and 3.10, respectively

$$CORR = \frac{1}{M} \sum_{i=1}^M \frac{\sum_{j=1}^{N_i} (y_i^j - \bar{y}_i)(\hat{y}_i^j - \bar{\hat{y}}_i)}{\sqrt{\sum_{j=1}^{N_i} (y_i^j - \bar{y}_i)^2} \sqrt{\sum_{j=1}^{N_i} (\hat{y}_i^j - \bar{\hat{y}}_i)^2}} \quad (3.9)$$

$$RMSE = \frac{1}{M} \sum_{i=1}^M \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (y_i^j - \hat{y}_i^j)^2} \quad (3.10)$$

where N_i is the number of frames in session i ($i = 1, 2, \dots, M$). y_i^j and \hat{y}_i^j are the ground truth and predicted values for the frame j ($j = 1, 2, \dots, N_i$) in session i . \bar{y}_i and $\bar{\hat{y}}_i$ are the mean values of y_i^j and \hat{y}_i^j for session i ($i = 1, 2, \dots, M$).

3.4.1 Dataset of AVEC 2014

The experiment is evaluated on a subset of the AVEC 2014 audio-visual depression dataset [166]. Subjects are recorded by a webcam and microphone while performing a Human-Computer Interaction. With only one person in every recording, two tasks are evaluated in each recording. The first task is *Freeform*, where subjects respond to one of a number of questions, and the length ranges from 7 seconds to 248 seconds. The second task is *Northwind* where subjects read aloud an excerpt of the fable. The length of this task recordings is between

Table 3.1: Pearson correlation coefficients of six sub-systems for the AVEC 2014 development set. Bold indicate the highest correlation at each affect dimension across each features and modality.

	Features	Affect Dimension		
Modality		Valence	Arousal	Dominance
Video	EOH	.552	.567	.602
	LBP	.548	.587	.514
	LPQ	.521	.562	.602
Audio	<i>long</i>	.577	.614	.587
	<i>short</i>	.551	.591	.590
	<i>voice detector</i>	.579	.595	.605
Baseline (Video)	LGBP -TOP	.355	.412	.319
Baseline (Audio)	LLD +MFCC	.347	.517	.439

33 seconds and 133 seconds. The recordings are split into three partitions: a training, development, and test set of 150 *Northwind-Freeform* pairs, totalling 300 task recordings. Tasks are split equally over the three partitions. This dataset is annotated in three emotion dimensions (valence, arousal, dominance) by 3 to 5 raters. For each recording, the gold-standard labels are the average of all the raters. There are at least three raters for each recording. In each emotion dimension, the range of gold-standard label is scaled to $[-1, 1]$.

3.4.2 Experimental Results and Discussion

To evaluate the performance of proposed approach, the experiments are carried out in the unimodal and multimodal setting. In unimodal setting, dynamic features such as EOH, LBP and LPQ from video modality as well as LLD descriptor from audio modality is dealt independently. By design, the dynamic features of facial expression images are synchronised with the gold-standard annotations. In the case of the acoustic feature, they were computed at a higher frame rate and need to be synchronized with the gold-standard label. Thus, interpolation technique has been performed at initial estimation in order to match it with the gold-standard label. In multimodal setting, initial estimation from respective features is fused together using linear opinion pool. The results are reported in Pearson correlation coefficient, where *Northwind* and *Freeform* task is averaged, and the results are indicated in the following section.

3.4.2.1 Results on Development Set

The empirical analysis is initiated with arousal dimension. From Table 3.1, it was evident that audio modality appears more informative than visual modality. The acoustic feature with *long* segmentation setting provides the highest correlation with gold-standard (CORR=.614) compared to LBP features from facial expression (CORR=.587). It is consistent with previous literature reported that prediction of arousal dimension from audio modality appears to be superior than visual modality [163] [14] [79].

On the other hand, previous literature mentions that visual modality appears to perform better in detecting valence dimension [50]. However, in this analysis, the results are not in agreement with such findings, for prediction of the valence dimension appears to be more superior to audio modality than video modality. *Voice detector* segmentation setting provides the highest correlation in valence with gold-standard (CORR=.579) compared to EOH features (CORR=.552). As for dominance dimension, the same setting also provides the highest correlation (CORR=.605) compared to LBP features (CORR=.602). Noted that disagreement occurs because of different features being used, and different modelling technique is being employed as opposed to [50]. Therefore, such conclusion holds for different data and remains to be evaluated.

Table 3.1 also shows the performance of different type of features on each affect dimensions in the development dataset. It can be seen that the highest correlation value, achieved by using LBP, outperformed EOH and LPQ features for each dimensions with respect to gold-standard (CORR=.587) in arousal dimension. As for valence, features from EOH gives highest correlation value (CORR=.552), apart from LBP and LPQ. As for dominance dimension, features from LPQ gives highest correlation value (CORR=.602). It shows that visual features encoded by EOH, LBP and LPQ capture the edge and local shape information of visual face effectively and can be mapped successfully for each of affective dimension. Additionally, performing wavelet transform at Level = 4 for each video features is less computationally expensive.

WT is carried out on acoustic features with three segmentation setting (*long*, *short*, and *voice detector*). PLS is then used for regression to predict each of affective scale as does in facial expression features. In arousal dimension, the *long* segmentation outperforms *short* and *voice detector* (CORR=.614). *Voice detector* segmentation gives higher correla-

Table 3.2: Pearson correlation coefficients of video final systems for the AVEC 2014 development set without wavelet filtering

Modality	Features	Affect Dimension		
		Valence	Arousal	Dominance
Video	EOH	.249	.257	.242
	LBP	.241	.267	.228
	LPQ	.249	.229	.209

Table 3.3: Pearson correlation coefficients of final systems for the AVEC 2014 development set with wavelet filtering. Bold indicate the highest correlation at each affect dimension across each features and modality.

Modality	Features	Affect Dimension		
		Valence	Arousal	Dominance
Video + Audio	EOH+long	#	.567	#
	LBP+long		.587	
	LPQ+long		.517	
	EOH+ voice detector	#	#	.602
	LBP+ voice detector			.589
	LPQ+ voice detector			.579
	EOH+ voice detector	.553	#	#
	LBP+ voice detector	.548		
	LPQ+ voice detector	.519		
	Baseline (Video+Audio)	LGBPTOP +LLD	.282	.478

tion value (CORR=.579) on valence and also for Dominance (CORR=.605) with respect to gold-standard.

Closer inspection of Table 3.1 shows each of the sub-system mentioned above surpasses the baseline results for each modality and each of every affect dimension. On top of that, the addition of more features provides significant gain, particularly in the case of facial expression features. However, this may not be optimal in acoustic features, since the only LLD with different segmentation setting is being utilized in this system.

To investigate whether multimodal fusion leads to a better recognition performance, all evaluation were repeated on facial expression feature and acoustic features, then fuse each of

Table 3.4: Pearson correlation coefficients of six sub-systems for the AVEC 2014 test set. Bold indicate the highest correlation at each affect dimension across each features and modality.

Modality	Features	Affect Dimension		
		Valence	Arousal	Dominance
Video	EOH	.503	.571	.492
	LBP	.518	.559	.518
	LPQ	.532	.571	.484
Audio	<i>long</i>	.498	.527	.475
	<i>short</i>	.447	.491	.489
	<i>voice detector</i>	.508	.508	.536
Baseline (Video)	LGBP -TOP	.188	.206	.196
Baseline (Audio)	LLD +MFCC	.355	.540	.360

the initial estimation at decision level. Each of the additional facial expression feature (EOH, LBP, LPQ) is fused with *long* and *voice detector* segmentation setting, since each of these two gives best results in VAD dimension respectively for the acoustic feature.

From Table 3.3, LBP from video modality fuse with *long* segmentation setting from audio modality gives higher correlation value on arousal dimension. As for valence and dominance dimension, EOH on video modality fused with *voice detector* segmentation on audio modality gives higher correlation results, respectively on both affect dimension. Even though system fusion relatively gives higher performance than baseline results, however, no significant difference was found between the best performance in sub-system in Table 3.1 with system fusion in Table 3.3. The reason is that only very simple fusion rule was used here.

3.4.2.2 Results on Test Set

The results for test set can be seen in Table 3.4 for video and audio modality. First of all, training and development set were combined as new training dataset. Then, Haar Wavelet Transform is applied to the new training sets, where the number of levels is fixed to 4, identical to the previous experiment in development set. As for PLS, all parameters, such as filter window size, number of component in PLS also identical to the previous set of experiments on the development set. As opposed the results in development sets, video modality shows high correlation results (CORR=.571) when compared to audio modality in arousal

Table 3.5: Pearson correlation coefficients of final systems for the AVEC 2014 test set. Bold indicate the highest correlation at each affect dimension across each features and modality.

Modality	Features	Affect Dimension		
		Valence	Arousal	Dominance
Video + Audio	EOH+ <i>long</i>	#	.572	#
	LBP+ <i>long</i>		.559	
	LPQ+ <i>long</i>		.576	
	EOH+ <i>voice detector</i>	#	#	.491
	LBP+ <i>voice detector</i>			.518
	LPQ+ <i>voice detector</i>			.484
	EOH+ <i>voice detector</i>	.503	#	#
	LBP+ <i>voice detector</i>	.518		
	LPQ+ <i>voice detector</i>	.535		
Baseline (Video+Audio)	LGBPTOP +LLD	.282	.478	.324

dimension. Same goes to dominance dimension, video modality performs better in correlation (CORR=.518), in contrast with development set, where audio modality performs better. As for valence dimension, the best results follow the modality from development set, where the higher correlation results (CORR=.532) is from video modality. The results in Table 3.4 suggest that suggested models are clearly able to generalise on unseen test sets. The multi-modal fusion results in Table 3.5 is done identically with the setting from development sets. The results show performance gain in every affect dimension. Arousal gives higher correlation (CORR=.576) when the prediction from LPQ and *long* segmentation setting is fused using linear opinion pool. As for dominance and valence dimension, fusion prediction between LBP and LPQ with VAD segmentation setting and gives higher correlation value (CORR=.518, CORR=.535) respectively.

3.5 Comparison on the best performer of the Challenge

The final system was also compared with the state-of-the-art approach in Figure 3.5 and Table 3.5. In the final system, the dataset trained on the training and development and tested on the

testing set. Overall, each of the proposed methods in the state-of-the-art approach achieves better performance than baseline results. The best results produce by Ulm [74]. However, their method utilises an extra information on subjects and annotation process without optimising features and machine learning used, as needed by AVEC 2014. The proposed framework achieve competitive performance in correlation (CORR) while achieving best results in terms of square error (RMSE).

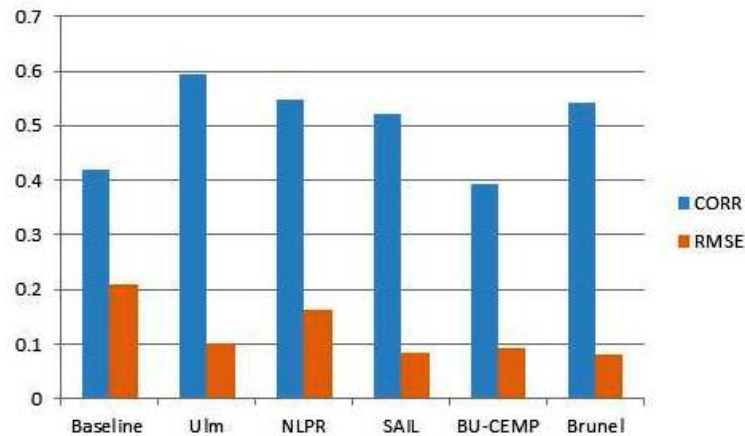


Figure 3.5 Bar chart comparison with state-of-the-art in AVEC 2014 in terms of average CORR and RMSE values

Table 3.6: Performance comparison with state-of-the-art in AVEC 2014 in terms of average CORR and RMSE values.

Team	Method	CORR	RMSE
Baseline [166]	SVR+Fusion	.419	.209
Ulm [74]	Subjects+Label Inference	.595	.101
NLPR [14]	Deep Learning + Fusion	.549	.163
SAIL [51]	Fusion + Temporal Regression	.522	.083
BU-CMPE [80]	CCA ensemble	.393	.093
Ours	Wavelet Filterng +PLS+Fusion	.543	.081

3.6 Chapter Summary

This Chapter studied and analysed human emotional state in terms of emotion dimension such as valence, arousal and dominance. Integrating wavelet transform with facial expression and acoustic features help to improve the performance significantly by removing the extraneous

noise in the feature level. One final system for continuous affect dimensional model is built, trained over multiple sub-systems involving audio-visual modality. There is a broad scope for improvements to the current approaches, both in improving the sub-systems and designing a combined system toward joint estimation. With the high complexity of human evaluation during annotation of gold-standard, the annotation delay might occur which may be degraded the performance. Also, instead of PLS, SVR as machine learning method might be taken into consideration, since it has been regarded as baseline approach in affect recognition. Last but not least, deep learning approach could usefully be explored which may be beneficial in emotion recognition domain. Each of the future work listed above is suitable to yields better performance as will be proved in the next Chapter.

Chapter 4

Continuous Affect Estimation based on Deep Learning Features

This Chapter investigates the effectiveness of deep learning in continuous emotion recognition task. The system performance is analysed in terms of layer-wise; convolutional and fully connected. In addition, multimodal fusion scheme is developed, by fusing deep learning feature and hand-crafted feature. The proposed framework is analysed and compared with the solution in the baseline, and competitive results are obtained.

4.1 Introduction

The success achieved by CNN has driven the advance in many aspects of computer vision, such as image classification, object detection, and image understanding. In face recognition, great improvement brought by CNN in solving large-scale face verification and recognition task has been witnessed [66] [81].

Like recognising the face, recognising emotion from face images has been an active research topic for the past few years. It is because facial expression of human is considered the most important way to represent the emotion which reflects the human behaviour essentially. While the work presented in Chapter 3 was based on recognizing emotion from hand crafted facial features, many researchers have demonstrated outstanding performances in recognizing emotion from CNN features [77] [76] [83]. However, the study mentioned above only recognise emotion from the categorical label. Employing deep learning features for continuous emotion

recognition is less frequently explored. In particular, no end-to-end method has been used to estimate arousal and valence directly from videos. It is because deep learning network requires large amounts of data to attain good performance. Most of the works reported their results in terms of combination from multiple modality and multiple features, be it hand crafted features or deep learning features. In [11], initial estimation is obtained with deep learning feature in multiple modalities but combined using late fusion with a linear regression. In [14], features obtained are trained using deep belief network then the results are combined using linear regression. Similarly, LSTM are used to perform multi-modal prediction using audio, video and physiological modalities [58] [9].

Even though deep learning features cannot currently work by itself for emotion recognition, this Chapter attempt to show that deep learning network has the ability to learn features on similar domain even only small sizes of a dataset is available for training. In this experiment, deep learning features for continuous emotion recognition in the face of limited data is explored. The goal is to combine deeply learned features to detect emotion with hand crafted features which will yield an improvement compared to using only the former.

Hand crafted features such as LGBP-TOP feature and facial geometric features is adopted, along with deep learning networks such as face verification model named VGGFace, image recognition model named AlexNet and Deep Residual Network. Additionally, it has been noticed that some annotation issues occur between annotator which is used to establish the gold-standard label. This delay can significantly degrade system performance when predicting emotion due to unreliable modelling caused by asynchronous ratings. This Chapter will demonstrate that by introducing a delay in continuous emotion recognition system, the performance is significantly improved.

The proposed method is evaluated on AVEC 2016 and compares with previous studies based on hand crafted features. The results show a significant improvement over the state-of-the-art and show that for a small set of database, the combination of hand crafted features and learned features obtain competitive performance.

In summary, the main contribution of this Chapter are:

- Analyzing the effectiveness of deep learning feature, where it can achieve competitive performance to the baseline model when combined with hand-crafted features.
- Exploring the complementarity of deep models with the existing visual and audio, by

applying fusion in decision label.

The remainder of the Chapter is structured as follows: Section 4.2 presents a review of previous work in continuous emotion recognition system in general. Section 4.3 describes the proposed methodology in the system. It consists of feature extraction, post-processing of annotation delay then fusion of deep-learned features with hand-crafted features. Section 4.4 describes the database used in the work and evaluation method of proposed approach. Section 4.5 shows experimental results and discussion. Finally, Section 4.7 summarises the Chapter.

4.2 Related Feature Extraction Technique

For the past few years, arousal and valence estimation from video/audio based has been defined as continuous emotion dimension [48]. Even though much of the earlier work treated arousal and valence (e.g. positive vs negative) as classification problem, more recent work treated the problem in continuous domain [48] [130] [125] [165]. The introduction AVEC challenge in 2011 marks significant progress in automatic continuous emotion research. It started with SEMAINE dataset [144] originally designed as classification problem, then in 2012 it moved to regression problem [143] using same dataset. The 2013 and 2014 AVEC edition marks an addition of audio-visual depression language corpus [167] [166], which has been previously experimented in the previous Chapter.

From 2015 edition of AVEC challenge, *Remote Collaborative and Affective Interaction* (RECOLA) dataset was used [127] [165]. It is the very first challenge that bridges across audio, video and physiological data. The main difference with the previous AVEC challenge is an adaptation of Concordance correlation coefficient (CCC) rather than Pearson's correlation coefficient to measure performance. In RECOLA database, Concordance correlation coefficient, which combines Pearson Correlation Coefficient with the square difference between the mean of the two-time series is being introduced as an evaluation measure. In this challenge, the best performer in the 2015 edition of the challenge obtained an average concordance correlation of 0.678 [58], while the best performer in 2016 obtained an average concordance correlation of 0.729 [9].

AVEC challenges employ traditional approach by providing features such as acoustic features, facial expression features, and physiological signal features. Acoustic features from audio

modality, typically referred to as acoustic LLD, include a wide range of features that cover spectral, cepstral, prosodic and voice quality information. Facial expression features from video modality mainly focused on capturing the change and intensity of the detected face over the duration of a task. The common examples are video appearance and video geometric based features. In video appearance feature, HOG, LBP, EOH is mainly used technique to employed facial expression features. A variant of LBP features, examined in spatio-temporal volumes of the video after convolving with 2D Gabor filter-banks (LGBP-TOP), has recently been used as baseline features for automatic continuous emotion recognition from video modality [127] [165]. As for video geometric features, the primary step is to localise and track a dense set of facial points landmarks [165], or shoulder landmark [107], or the whole body expressions landmark [103]. These landmarks are then tracked to acquire low-level descriptors of the dynamics of facial or body gestures.

Despite its success in recognising emotions, hand crafted features have a number of problem in use. These hand-crafted features often lack generalisation ability, where disturbance such as high variation in scene lighting, camera view, image resolution, background, subjects head pose also skin colour may effect the overall performance. Thus an alternative approach is deep learning approach, where it learns the most appropriate feature abstractions directly from the data (image frame) and handle the limitations of hand-crafted features. Deep learning have become a successful approach in object recognition [88], face verification [156], human pose estimation [162] and so on. This success is contributed by the availability of computing power such as GPU and existing big databases such as ImageNet and MNIST that allow deep learning to extract highly discriminative features from the data samples. There have been enormous attempts at using DNNs in automated facial expression recognition and affective computing such as [105] [104] [58] which is very successful in determining emotion. The winner of the AVEC 2015 challenge [58], implement deep learning approach with hand-crafted features and still could improve the prediction accuracy.

Due to successful application in deep learning approach in computer vision, this Chapter intends to explore how far deep learning can recognize emotion in continuous dimension. Convolutional Neural Network (CNN) has been selected as deep learning architecture in order to do feature extraction on top of facial expression images. The main difference between proposed approach and the method implemented by winning team in [58] is that while they

were making use of Long Short Term Memory as machine learning approach to estimate arousal and valence, the proposed approach employed CNN architecture as feature extractor estimate arousal and valence.

4.3 Methodology

The proposed methodology is illustrated in Figure 4.1. At the initial stage, for each video clip, video and audio modality is dealt independently. Feature extraction is adopted in Stage 1. In this stage, feature extraction is implemented towards each of the visual frame and acoustic file. As for the visual frame, hand crafted features and deep learning features is adopted, in order to make feature representation. In Stage 2, Support Vector Regression is implemented, since it already becomes a baseline approach in many continuous emotion recognition tasks [125] [165]. Each of the features types, be it hand-crafted or deep learning feature, is feed consecutively into Support Vector Regression. Then, Stage 3 undergone post-processing method, where a time delay was optimised by maximising the performance on the development partition with the model, learned on the training partition. The initial estimation produced in Stage 3 is feed to Stage 4, where fusion on respective features occur, to produce the final estimation of continuous emotion dimension of arousal and valence. Each of the Stage will be discussed thoroughly in the following section.

4.3.1 Stage 1: Feature Extraction

Each of the video clip provided is converted to image frames, in order to capture the facial expression recorded in video data. Then, deep learning feature and the hand crafted feature is being implemented as feature extractor from the video frame. Deep learning feature consists of VGG-Face CNN descriptors, image recognition model named AlexNet and Deep Residual Network. For hand-crafted feature, two types of facial landmark: appearance and geometric feature is employed, where the first is adapting LGBP-TOP and the latter applied facial landmark on each video frame. As for audio, a set of LLD is set to encode the characteristic of the acoustic feature. Each of the features is detailed in the following subsection:

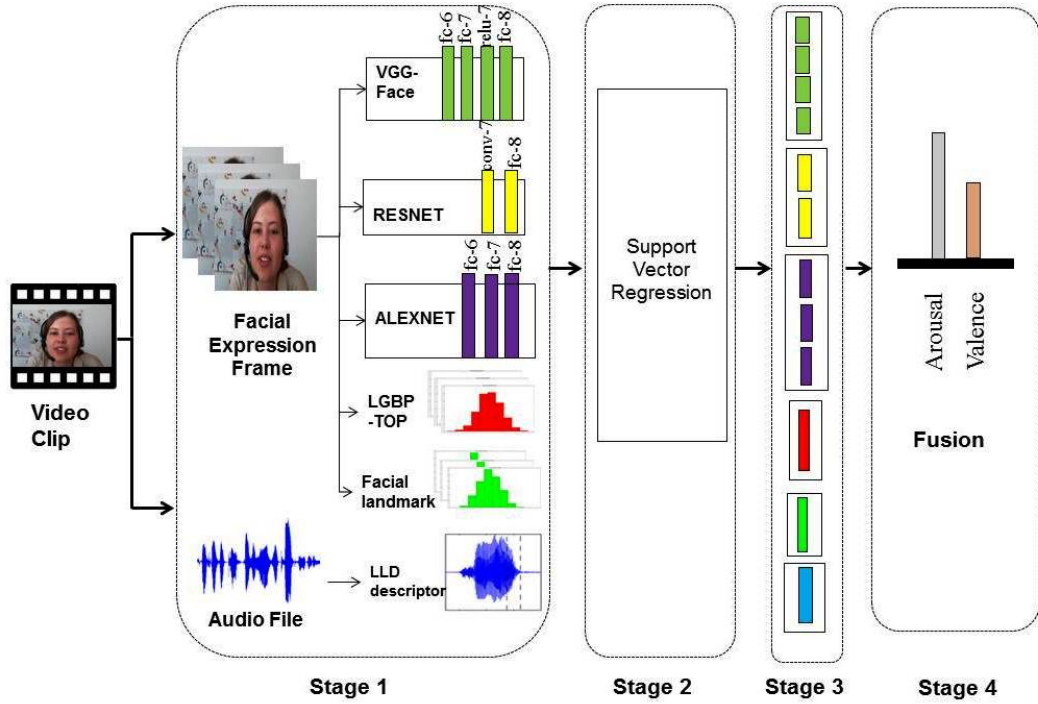


Figure 4.1 : A depiction of the pipeline of the proposed system. This depiction is specific to the combination of features from video modality and audio modality. In deep learning feature, $fc-6$ to $fc-8$ of VGGFace, $conv-7$ to $fc-8$ of ResNet and $fc-6$ to $fc-8$ of AlexNet is extracted from each CNN network. This is done for all frames of the video which produces a single feature vector in respective layer. Along with hand-crafted features and LLD descriptor, these feature is then feed into SVR for regression purpose.

4.3.1.1 Deep Learning Feature using Convolutional Neural Network

Inspired by deep learning recent success in computer vision, three different frameworks that train neural networks as feature extraction has been designed. There are VGGFace Network, image recognition model named AlexNet and Deep Residual Network. Given the usefulness of CNN features, this experiment aims to assess the features further and demonstrate how CNN features can be used for emotion recognition task.

In this experiment, pre-trained deep learning network has been used as feature extractor for the task of interest. It is conducted by removing the last few layers of a CNN, then treated the remainder of the CNN as a fixed feature extractor. Considering that these pre-trained CNNs are large multi-layer architectures, which encode a large amount of visual knowledge of varying complexity [182] [40], it is interesting to see what differences are there between convolutional layers and fully connected layers, so that all that is learnt by a CNN can be properly exploited in the future.

There are 3 hyperparameters needed change in order to study the behavior of each layer. There are depth, stride and zero-padding. First, the depth of the output volume corresponds to the number of filters we would like to use, each learning to look for something different in the input. For example, if the first convolutional layer takes as input the raw image, then different neurons along the depth dimension may activate in presence of various oriented edges, or blobs of color. Second, the stride need to be specified with which slide the filter. When the stride is 1 then the filters move at one pixel at a time. When the stride is 2 then the filters jump 2 pixels at a time as we slide them around. This will produce smaller output volumes spatially. Lastly, sometimes it will be convenient to pad the input volume with zeros around the border. This is called zero-padding. It is needed to control the spatial size of the output volumes.

CNNs are usually applied to image data. Every image is a matrix of pixel values. The range of values that can be encoded in each pixel commonly on 8 bit or 1 byte-sized pixels. Thus the possible range of values a single pixel can represent is $[0, 255]$. However, with coloured images, particularly RGB (Red, Green, Blue)-based images, the presence of separate colour channels (3 in the case of RGB images) introduces an additional ‘depth’ field to the data, making the input 3-dimensional. Hence, for a given RGB image of size, there will be 3 matrices associated with each image, one for each of the colour channels.

To understand the architecture going on in each stage of a deep convolutional neural network, the process based on mathematical principle is briefly explained as below:

- **Convolutional Layer:** Convolutional layers employ learnable filters which are each convolved with the layer’s input to produce feature maps $Z^l(x, y, i)$ for neuron i from each convolutional layer l is computed as:

$$Z^l(x, y, i) = X^{l-1}(x, y, c) * K_i^l(x, y, c) + B_i^l \quad (4.1)$$

- **Rectified Linear Unit:** A Rectified Linear Unit (ReLU) is a cell of a neural network which uses the following nonlinear activation function to calculate all convolved extracted features. ReLU is often assigned to the output of each hidden unit in a convolutional

layer and the fully connected layers. The output of the ReLU $P^l(x, y, i)$ is computed as:

$$P^l(x, y, i) = \max(0, Z^l(x, y, i)) \quad (4.2)$$

- **Normalization layer:** In this process, local response normalization is used for normalizing the output of the ReLU. This step is needed to yield better generalization and introduces non-linearity that is absent in the right hand side of the ReLU responses. It can be computed as:

$$Q^l(x, y, i) = P^l(x, y, i) \left(\gamma + \alpha \sum_{j \in M^l} (P^l(x, y, j))^2 \right)^{-\beta} \quad (4.3)$$

where $Q^l(x, y, i)$ computes the response of the normalized activity from the ReLU output $P^l(x, y, i)$. This is done by multiplying the output with an inverse sum of squares plus an offset γ for all ReLU outputs within a layer l

- **Max pooling layer:** The max-pooling operator computes the maximum response of each feature channel obtained from the normalized output. It can be computed as:

$$R^l(\bar{x}, \bar{y}, i) = \max_{x, y \in M(\bar{x}, \bar{y}, i)} Q^l(x, y, i) \quad (4.4)$$

where (\bar{x}, \bar{y}) is the mean image position of the positions (x, y) inside $M(\bar{x}, \bar{y}, l)$ that denotes the shape of the pooling layer, and $R^l(x, y, i)$ is the result of the spatial pooling of the convolutional layers. Noted that, max pooling reduces the dimensionality by applying the maximum function over the input R

- **Average pooling layer:** The average pooling operator computes the mean response of each feature channel obtained from the normalised output. It can be computed as:

$$R^l(\bar{x}, \bar{y}, i) = \frac{\sum_{x, y \in M(\bar{x}, \bar{y}, i)} Q^l(x, y, i)}{|M(\bar{x}, \bar{y}, i)|} \quad (4.5)$$

- **Classification layer:** The probability of the class labels from the output of the fully connected layer is computed using the softmax activation function. It computes the probabilities of the multi-class labels using the sum of weighted inputs from the previous

layer and is used in the learning process [114]:

$$y_d = \frac{\exp(x_d)}{\sum_{d=1}^D \exp(x_d)} \quad (4.6)$$

where y_d is the output of the softmax activation function for class d , x_d is the summed input of output unit d in the final output layer of the fully connected network and D is the total number of classes.

In the end, an end-to-end architecture as illustrated Figure 4.1 is developed, where architectures that trained all-together, accepting raw data colour images (facial expression images), learned to produce an estimation of valence and arousal. In particular, the following architectures have been evaluated:

- An architecture based on the structure of the VGGFace network [118]
- An architecture based on the structure of the ResNet-50 network [59]
- An architecture based on the structure of the AlexNet network [88]

Each of the networks is applied directly to facial expression frames of the database, trained to produce features from the respective layer, where the layer is detailed in the following subsequent:

VGG-Face [118]

In this subsection, pre-trained VGG-Face CNN descriptor as described in Figure 4.2 is leveraged due to the limited size of the dataset. The pre-trained VGG-Face CNN was learned from a large face dataset containing 982,803 web images of 2622 celebrities and public figure. While the primary purpose of VGG-Face is to identify subjects in its training dataset, it can be employed as a feature extractor for any facial expression image by running the image through the entire network, then extracting the output of the fully-connected layer (FC). The VGG-Face model is initially designed for face recognition. Therefore, the deep model is utilized by pre-trained the model and change the final layer for continuous emotion recognition problem.

The extracted feature vector serves as a highly discriminative, compact, and interoperable encoding of the facial expression image. Once the features are derived from the fully connected

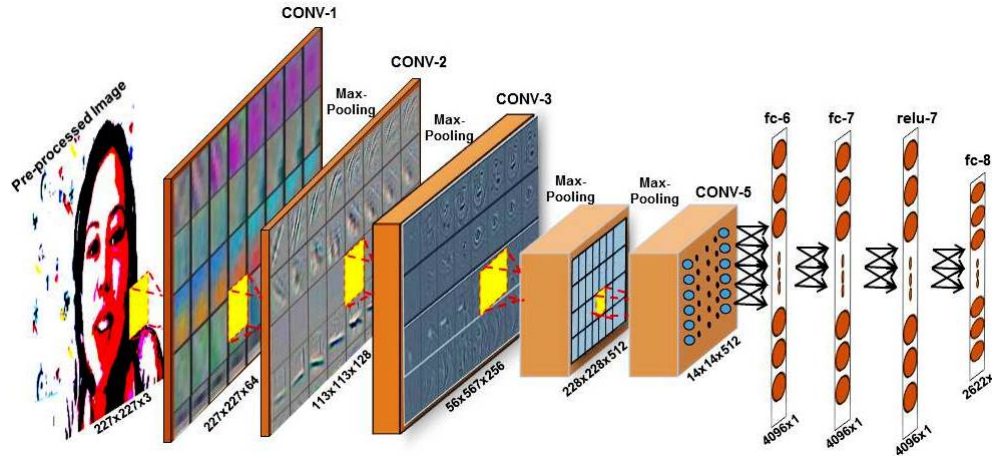


Figure 4.2: CNN architecture for VGG-Face

layer of the VGG-Face CNN, they can be used for training and testing arbitrary emotion regressors, which is the goal of this experiment.

MatConvNet toolbox by Vedaldi and Lenc [172] which consists of a library of MATLAB functions implementing CNN architectures have been employed as the software tool since it provides the researchers with the pre-trained implementation of VGG-Face CNN where it can be used as feature extraction. Then, the datasets contain RGB images which are fed to the convolutional neural network in their original colour channels.

The VGG-Face network described in Figure 4.2 has a deep architecture and is composed of 3×3 convolution layers, 2×2 pooling layers, and 4 fully-connected layers. While the network was initially trained to perform classification rather than feature extraction, the output layer of the network after softmax is not used in the experiments, rather the output layer before softmax, which contains 4096-dimensional descriptors are instead extracted from the first fully-connected layer.

VGG-Face is modelled as feature extractor by fixing all of its parameters and removing the regression layer. In order to extract the features, firstly, the facial expression images are preprocessed and fed to the CNN as a $224 \times 224 \times 3$ arrays of pixel intensities. At every convolutional layer, it performs a filtering operation on the former layer resulting in an activation volume. The activation then becomes the input of the following layer. Pooling is used throughout the network to reduce the number of nodes by down-sampling the activation maps using the max operation. Finally, features from fully connected layer (fc) are extracted

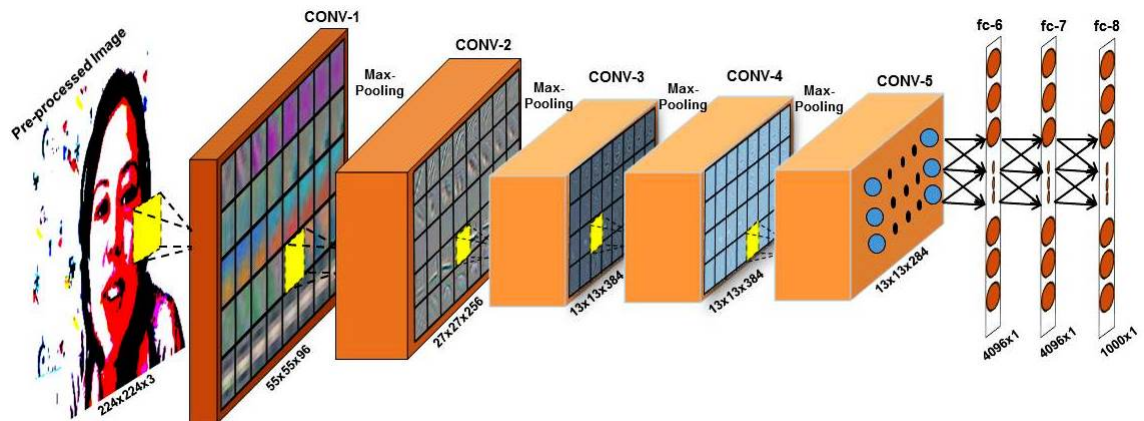


Figure 4.3: CNN architecture for AlexNet

when a frame is passed to VGG-Face network.

AlexNet [88]

AlexNet network is designed, competed and won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2012). It is an architecture that is based on the traditional CNN layered architecture - stacks of convolutions layers followed by max-pooling layers and rectified linear units (ReLU), with a number of fully connected layers at the top of the layer stack.

This network was trained on the ILSVRC-2012 training data, which contained 1.2 million training images belonging to 1000 classes. It brought down the error rate on image classification task by half, beating traditional hand-engineered approaches. This network also has revolutionised the way of thinking about the effectiveness of CNNs.

In this subsection, AlexNet trained CNN is being employed to extract the images descriptors, which is in this case, facial expression image. The input images are resized to 227×227 pixels using bicubic interpolation. Same as VGGFace, AlexNet also requires three channel images at the input (RGB format). Facial expression images are then processed by the Alexnet network, where each frame is processed by square convolutional layers of size 11, 5 and 3 each followed by max-pooling and local response normalisation (ReLU). The last three pooling layer output, $fc-1$, $fc-2$ and $fc-3$ is extracted for each facial expression images, where it has 4096 for each fully connected layer.

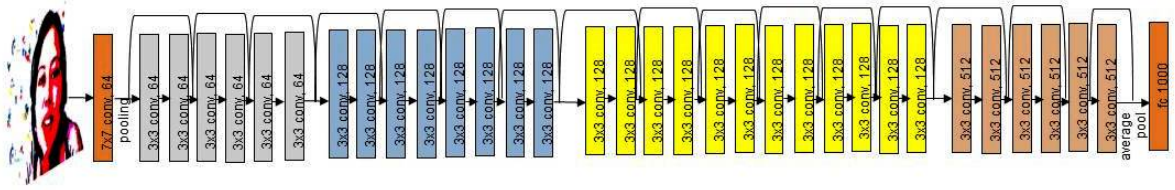


Figure 4.4 : CNN architecture for ResNet. (a) Residual network with 34 parameter layer. Noted that shortcut connection is added by skipping layer.

ResNet [57]

ResNet is another current state-of-the-art CNN architecture. Based on residual connections introduced by He et al., [57] it has shown remarkable improvement in recognition rates in ILSVRC-2015 competition. ResNet architecture is similar as VGG-16 proposed by [149] but with the additional identity mapping capability as shown in Figure 4.4. The first layer of ResNet is a 7x7 convolutional layer with 64 feature maps, followed by a max pooling layer of size 3x3. The rest of the network comprises of 4 bottleneck architectures, where after these architectures a shortcut connection is added. These architectures contain 3 convolutional layers of sizes 1x1, 3x3, and 1x1, for each residual function. After the last bottleneck architecture, an average pooling layer is inserted. The last *fc-8* layer, and another layer before average pooling layer *conv-7* is extracted from ResNet network.

Layer-wise Features of Deep Learning Architecture

From VGGFace, AlexNet and ResNet architecture, each of the deep networks is formed by a stacked structure, where each of the stacked structure gives feature representation from each layer. From left to right of a deep neural network, the features learned are from general to specific. For example, the earlier layer is known to learn the features that are similar to Gabor filters and colour blobs [181] while the latter layer (high-level layer) are usually well trained to specific task, e.g., image classification task [88]

Most previous work by default directly use the feature outputs from the high-level layer, since the high-level semantics expressed in these high-level layers are more related to the specific task. However, detecting emotional content from video modality is quite tricky because inside the video it is more diverse and more sparsely expressed video emotions. No previous work touched the effects of layer-wise features towards emotion recognition domain. Therefore,

these Chapter explore these question by evaluating fully connected layer (fc) from each deep network employed in these dataset. The output of each layer is considered as a visual descriptor of each frame. The difference in accuracy between layers is crucial to get the intuition on their suitability for video emotion analysis.

4.3.1.2 Hand Crafted Features

It is easy to interpret smiles, frowns, eyebrow-raising and more subtle actions that can be performed by the facial muscles. Those facial muscles movement can be extracted using hand-crafted features, using a manually predefined algorithm based on the expert knowledge. Local Binary Pattern [112], Local Scale-Invariant Features [94] and Histogram of Oriented Gradient [21] features are commonly known examples of hand-crafted features. In these Chapter, two type of facial descriptor: appearance and a geometric-based feature are employed in the experiment.

Face Geometric Feature

The geometric features are based on 49 facial landmarks detected and subsequently tracked with the Supervised Descent Method (SDM) facial point detector/tracker proposed by Xiong and De la Torre [180]. These 49 facial landmarks have been aligned with a mean shape from stable points (located on the eye corners and on the nose region). The features are computed as follows:

- The difference between the coordinates of the aligned landmarks and the mean shape is computed, at the same time the difference between the aligned landmark locations in the previous and the current frame. This provides 196 features.
- The facial landmark has been divided into three different regions: i) the left eye and left eyebrow, ii) the right eye and right eyebrow and iii) the mouth. For each region, Euclidean distances (L2-norm) and the angles (in radians) between the points are computed, resulting 71 features.
- The Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame is also computed, providing 49 features.

In total, geometric feature yields 316 features as a face representation from this method.

LGBP-TOP Features

The local dynamic appearance descriptor LGBP-TOP features take 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. In this Chapter, feature reduction is then performed on single feature histogram by applying a Principal Component Analysis (PCA) from a low-rank (up to rank 500) approximation. Finally, 84 features representing 98% of the variance is computed as a face representation from this method.

4.3.2 Stage 2: Support Vector Regression as Modelling Approach

To handle the regression task, *Support Vector Machine for Regression* (SVR) is employed owing to its mature theoretical foundation. It is also regarded as the baseline approach for many continuous affective recognition task [125] [143] [165]. By applying SVR onto regression task, the main target is to optimize the generalisation bound for regression in the feature space by using a ε -insensitive loss function which is used to measure the cost of error in the estimation. Hyperparameter C is also set accordingly to balance the errors and the generalisation performance.

In this Chapter, SVR was implemented in the LIBLINEAR toolkit with linear kernel and trained with the L2-regularised L2-loss dual solver. The tolerance value of ε set to be 0.1, and complexity (C) of the SVR was optimised by the best performance of the development set among [.00001, .00002, .00005, .0001, . . . , .2, .5, 1] for each modality and task.

4.3.3 Stage 3: Post-processing

After obtaining the initial estimation from SVR, the number of post-processing method to see the impact of SVR performance. First, the annotation delay was introduced to the training partition. It is undergone by shifting back in time the gold-standard training partition. The last value after shifting is duplicated, to realigned back the gold-standard after shifting.

Initial investigations focused on searching best delay for arousal and valence, which were achieved using a number of delay values from zero to eight seconds, as similar in [165]. Table

Table 4.1: Delay in seconds applied to the gold-standard, according to the emotional dimension (A=arousal, V=valence). The delay were obtained by maximising the results in development partition while applying the delay in training partition.

Modality	Type	Features	Layer	Delay _A	Delay _V
Audio	Acoustic	LLD	#	2.8	3.6
Video	Hand-Crafted	Geometric	#	2.4	2.8
		LGBP-TOP	#	2.8	2.4
		VGGFace	32	2.0	7.2
	34		5.6	4.8	
	35		0.0	2.8	
	36		0.0	2.8	
	Deep Learning	ResNet	514	1.6	8.0
			515	1.2	4.0
		AlexNet	16	1.2	1.6
	18		8.0	5.6	
	20		8.0	7.6	

4.1 shows the best delay achieved on training partition while maximising the concordance correlation in development partition.

4.3.4 Stage 4: Fusion Approach for Final Estimation

The initial estimation from respective hand crafted and deeply learned features are combined to achieve the highest possible final, single estimation performance. In this stage, exponent weighted decision fusion is incorporated where weighting each of them according to initial correlation from development dataset. It is a type of fusion decision which introduced in [84] for classification purpose. These method is being leveraged into regression purpose, where a decision weight in terms of (C^q) , and exponent q is found by hyper-parameter tuning. The value of q is found by scanning procedure over $[-50:0.1:150]$ then selected to provide the maximum correlation after the fusion. The scanning procedure proved to be useful to enhance the performance as illustrated in Figure 4.5.

4.4 Experimental Evaluation

The proposed framework was trained on the training set and tested on development sets for each of affect recognition, where the level of affect has to be estimated for each frame of the recording. The effectiveness of the proposed framework on emotion recognition using

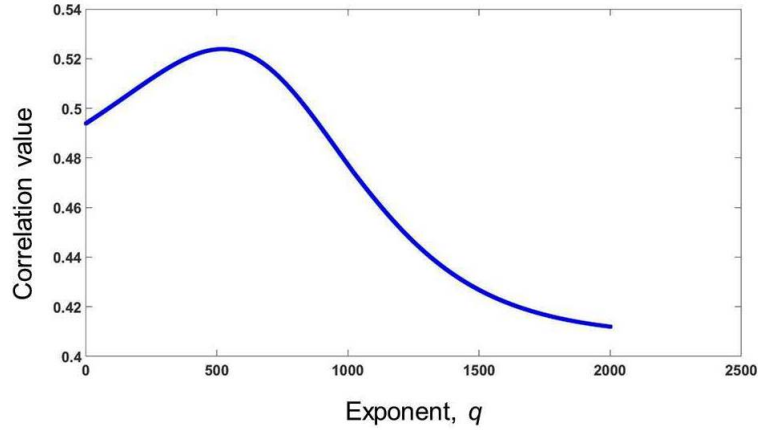


Figure 4.5 :Performance value in term of correlation as an exponent q is scanned in the exponentially weighted decision fusion. Noted that when proper q is selected, it gives maximum performance in development sets

deep layer-wise features and hand-crafted features is validated using Concordance Correlation Coefficient (CCC). It is the combination of Pearson Correlation Coefficient (CC) with the square difference between the mean of the two compared time series, as shown in Equation 4.7

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (4.7)$$

where ρ is the Pearson Correlation Coefficient between two-time series (eg., estimation and gold-standard), σ_x^2 and σ_y^2 is the variance of each time series, and μ_x and μ_y are the mean value of each time series.

4.4.1 Dataset of AVEC 2016

Remote Collaborative and Affective Interaction (RECOLA) dataset was recorded in the context collaborative work. Spontaneous interactions were collected in dyads and remotely through video conference. The corpus consists of multimodal modality, e.g., audio, video, ECG, and EDA which were recorded continuously and synchronously from 27 French-speaking participants. Each of the subjects has different mother tongue (French, Italian, and German), which provides further diversity in the recording of affect. In this experiments, only audio and modality is taken into consideration.

In order to ensure speaker-independence, the corpus is divided into three partitions (Training, Development, Testing) with each partition containing nine unique recording approxi-

mately balanced for gender, age, and mother tongue of the participants. In this experiment, only Training and Development partition has been fully utilised, since only the gold-standard label for this partition is publicly available for the researchers.

As for annotation of the corpus, six-French speaking raters (three male and three females) were asked to perform annotation in terms of arousal and valence for the first five minutes of all recording sequences. The obtained labels were then resampled at a constant frame rate of 40 ms, and averaged over all raters by considering the inter-evaluator agreement, to provide a *gold standard* label.

4.5 Experimental Results and Discussion

In this section, the results of the proposed deep learning and hand-crafted features from visual modality is discussed, focusing on estimation accuracy as evaluated by the concordance correlation (CCC).

For deep learning network, the activations of the three to four layers that are closest to the output (softmax) layer for the three deep networks is taken under consideration. These layers are denoted as *fc-6*, *fc-7*, *relu-7*, *conv-7*, *fc-8* respectively by the creators of each networks. VGGFace *fc-8* size has 2622 neuron, denoting of 2,622 identities in face verification task. Meanwhile, ResNet and AlexNet has 1000 neuron, denoting 1000 attribute in image classification task. *fc-7* features are the activations of the layer before *fc-8* while *fc-6* are the activations layer before *fc-7*. *fc-7* are the activation function from final hidden layer before propagating to softmax function. AlexNet and VGGFace are built with three fully connected layers, where two with 4096 neurons, and one with 1000 neurons for AlexNet and 2622 neurons for VGGFace, which outputs the class probabilities. However, ResNet only has one fully connected layer with 1000 neurons which again outputs the class probabilities. It is because, unlike VGGFace and AlexNet, ResNet focus on fasten training performance by deepening and lowering the parameter, resulting 8x smaller in depth than VGGNet. Noted that, the goal of this experiment is to investigate the use of the networks hidden layer activations as features and the objective is to purely evaluate the baseline emotion approach and not the flexibility of the network. Therefore, the weights of the initial layers are frozen [172], while retrain only the higher layer using SVR.

Table 4.2: CCC obtained in development partition after SVR and post-processing method. Features are taken from deep-learned features.

Deep Network	Layer		Neuron	Corr _A	Corr _V
AlexNet	16	<i>fc-6</i>	4096	0.09	0.05
	18	<i>fc-7</i>	4096	0.13	0.16
	20	<i>fc-8</i>	1000	0.06	-0.02
ResNet	514	<i>conv-7</i>	2048	0.18	0.21
	515	<i>fc-8</i>	1000	0.20	0.27
VGGFace	32	<i>fc-6</i>	4096	0.33	0.39
	34	<i>fc-7</i>	4096	0.35	0.37
	35	<i>relu-7</i>	4096	0.31	0.39
	36	<i>fc-8</i>	2622	0.29	0.31

The results of the experiments, across all nine combinations of the layer-wise deep learned network, are shown in Table 4.2. The empirical analysis is initiated with valence dimension. Overall, it shows that each fully connected layer provides higher correlation in valence rather than arousal dimension. It is consistent with previous research that the facial expression cues are very informative for predicting valence rather than arousal. While looking at the layer-wise feature, *fc-6* as well as a *relu-7* layer in VGGFace network gives higher results in terms of concordance correlation of 0.39 for respective network. It can be seen that valence has an increment in terms of correlation after ReLU activation, from 0.37 to 0.39. This increment partly because while the feature output is taken after Rectifier Linear Unit (ReLU) activation function, it is undergone by removing negative neuron by setting it to zero. As for ResNet, *fc-8* gives the higher results than *conv-7* layer. In arousal dimension, the results show that VGGFace *fc-7* gives higher results, with concordance correlation of 0.35, beating the other two deep network. Looking closely at layer-wise, it appears that decrement occurs after ReLU activation function is applied, from 0.35 to 0.31. Then at the last layer before softmax, that is *fc-8*, each of emotion dimension has the lowest correlation, 0.29 for arousal and 0.31 for valence. Noted that, each of earlier layer in *fc-7*, *fc-8* and *relu-7* has 4096 neuron per frame, while *fc-8* only has 1000 neuron per frame. The reduction of features at high-level layer may contribute to the decreasing in performance as can be seen in this case. Also, keeping a negative node as features are preferred, which might give more information towards emotion recognition. As for ResNet, it appears that deeper layer *fc-8* gives higher results with concordance correlation 0.20 compared to 0.18 in the previous layer. Meanwhile, in the AlexNet network, it does not provide the significant improvement in terms of concordance correlation for both emotion

dimension compared to the other two network. It could be because the face images from the dataset are very different from images in ImageNet, most likely it comes from different distribution. As a result, transfer learning does not capture informative neuron at the lower layer of the network.

Looking back at the results in Table 4.2, it is clear that the employment of VGGFace trained for face recognition task is more efficient than the deep network trained for image classification task. It is expected because ImageNet is comprised of 1000 classes ranging from cat, dogs to cars. By the last fully connected layers of the AlexNet and ResNet network, the network is close to predicting what class of the objects the image falls in. Thus, the neurons are understandably not good at differentiating between depressed face or neutral face.

Table 4.3: CCC obtained in development partition after SVR and post-processing method. Features are taken from hand-crafted features.

Hand Crafted	Corr _A	Corr _V
LLD Descriptor	0.79	0.45
Facial Geometric	0.38	0.61
LGBPTOP	0.48	0.47

Among the single hand-crafted feature in Table 4.3, LLD descriptor feature from audio modality perform best in arousal dimension, and LGBP-TOP features perform better in valence dimension. Overall, deep network demonstrated to be the weakest when compared to hand crafted features. It could be explained by the fact that the deep network in Table 4.2 are significantly different from hand crafted in Table 4.3 in several key features. A possibility for this cause can be due to the lack of direct training of the deep models on the emotion dimension task data. Hand crafted features, on the other hand, are crafted directly on the original face image, and mostly they are generic in nature.

Looking closely at Table 4.2 and Table 4.3, encouraging results also obtain in VGGFace features where it performs similar results to the geometric features in arousal dimension and LLD descriptor in valence dimension. At one sense the deep network can generate much more compact features and a allows natural cross-modal matching in continuous emotion task.

Following the trend of fusing hand crafted features and deep learning features to incorporate more information, late fusion has been performed from audio and visual modality across each dimension to prove that they contain different information. To this end, late fusion of the scores up from audio and visual modality, weighting each of them in an equal manner, to

Table 4.4: Comparison of CCC on fusion from baseline and proposed approach. Noted that proposed approach results are obtained on 2 fold cross validation.

All	Corr _A	Corr _V
Baseline [165]	0.82	0.68
Proposed Approach	0.81	0.69

further improve the concordance correlation results, as shown in Table 4.5. However, there is two main difference between the proposed methodology and system baseline in [165]. Firstly, the system in benchmark was trained using both training, development and test partition, whereas the proposed approach was trained using only the training partition, and tested on development partition using 2-fold cross validation. Secondly, the baseline system operates using video, audio and physiological signal modality, while proposed system only runs on audio and video modality.

In fusion stage, AlexNet features have been omitted in the fusion stage since it has the lowest performance in both emotion dimension. Looking at the results, it shows the concordance correlation increase to 0.81 in arousal and 0.67 in valence dimension after late-stage fusion is employed. This shows that though deep network features do not work well on their own, they provide valuable information which complements the handcrafted features. It can be added that deep learning features are directly learned from the image pixels with less loss of information whereas the handcrafted features are learned on high-level representations of the original face image and are prone to oversimplification. By fusing each of the features together, it contributed to the performance improved when all features are combined.

4.6 Comparison on the best performer of the Challenge

The proposed approach is also compared with the 7 state-of-the-art results [4] [9] [69] [174] [121] as in Figure 4.6. Among state-of-the-art approaches, Sun et al., [154] and Brady et al., [9] introduces deeply learned model into the system, however they also use physiological modality in fusion approach, which not taken into consideration the proposed fusion system. The best results are also from [9]. However, apart from physiological signal modality, [9] also incorporating high-level audio features, while the proposed approach only uses baseline audio features in the current system. The proposed approach achieve competitive performance in

correlation for both arousal and valence emotion dimension.

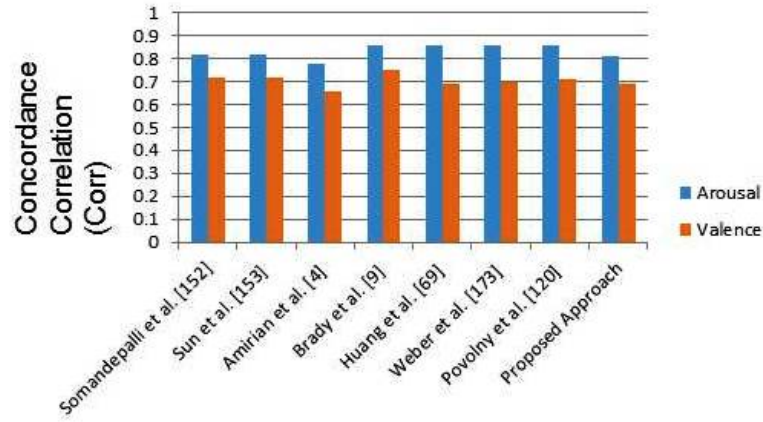


Figure 4.6 Bar chart comparison with state-of-the-art in AVEC 2016 in terms of concordance correlation in Arousal and Valence dimension.

Table 4.5: Comparison of CCC on fusion of AVEC 2016 state-of-the-art and proposed approach.

All	Corr _A	Corr _V	Method
Somandepalli et al. [153]	0.82	0.72	KF
Sun et al. [154]	0.82	0.72	add physiological modality
Amirian et al. [4]	0.78	0.66	RF+late Echo State
Brady et al. [9]	0.86	0.75	add physiological modality
Huang et al. [69]	0.86	0.69	OA framework
Weber et al. [174]	0.86	0.70	add geometric feature
Povolny et al. [121]	0.86	0.71	add audio feature
Proposed Approach	0.81	0.69	add deeply-learned features

4.7 Chapter Summary

This Chapter investigates the effectiveness of deep learning approach in terms of layer-wise in continuous emotion recognition task. By incorporating CNN features as deep learning architecture, VGGFace, AlexNet and ResNet network is being employed as feature extractor from deep learning. Numerical results show that VGGFace, which initially is trained for face verification task, performs better in detecting continuous emotion, outperformed the other two network. The combination of hand crafted features with deeply learned features yields significant improvement over the former and gives competitive performance over the state-of-the-art. Hence, suggesting that advance fusion strategy may yields better performance as it

will be discussed in the next Chapter.

Chapter 5

Continuous Affect Estimation in Two Stage Regression Framework

This Chapter proposes two-stage regression framework, where initial estimation of affect is feed into subsequent model such as Time Delay Neural Network, Long Short Term Memory and Kalman Filter. The two-stage approach separates the emotional state dynamics modelling from an individual emotional state prediction step based on input features. In doing so, the temporal information used by the subsequent model is not biased by the high variability between features of consecutive frames and allows the network to exploit the slow changing dynamics between emotional states more efficiently. The results further show that the proposed framework is competitive with the other state-of-the-art approaches, but being with a simple implementation.

5.1 Introduction

In recent years, a considerable amount of research in automatic continuous emotion has undergone to enable natural, intuitive and friendly human-machine interaction. Early works have focused on the recognition of primary discrete emotion, with the data being collected in a laboratory setting, where speakers act in specific emotional states [47] [141] [142] [184]. Recently, considerable amount of literature focus on emotional behavior in continuous dimensions such as arousal and valence [46] [48] [119] [175] [89] [151]. One possible explanation is that a single label may not reflect the complexity of the affective state conveyed by multiple sources of infor-

mation [129]. Hence a number of research areas have started to model, analyses and interpret the continuity of affective dimensions, rather than discrete emotion. Furthermore, the affective computing is moving towards combination of multiple modalities such as audio, video, text and physiological signal [97] [117] [152] [179] [186]. Traditionally, most of the studies have focused on a single modality - such as audio [96], video [155] or physiological signal [152]. With the advancement of devices such as the camera and microphone, multi-modality has been widely implemented for emotion recognition [10] [107] [102]. The combination of various modalities can be more useful for identifying and classifying emotions, which can boost emotion accuracy since each modality can provide complementary information [151] [165] [107].

In order to learn the relationship between the feature from various modalities and the multi-dimensional affective space, a variety of machine learning models have been investigated, such as k-Nearest Neighbor [122], continuous conditional random fields (CCRF) [70] and Relevance Vector Machine (RVM) [69]. Support Vector Regression (SVR) is a popular technique that has been frequently employed and is regarded as the baseline regression approach for many continuous affective computing tasks [127], [165] [145]. Recently, Long Short-Term Memory Recurrent Neural Networks (LSTM- RNN) [63] has become one of the state-of-the-art modeling technique in continuous emotion recognition, thanks to its ability to incorporate knowledge about how emotions typically evolve over time so that the inferred emotion estimates are produced under consideration of an optimal amount of context [179]. LSTM-RNN was first applied on acoustic features [178], and consecutively it has been successfully employed for other modalities such as video, and physiological signals [16] [107] [127].

There is a large volume of published studies covering wide range of machine learning techniques for continuous emotion recognition [127] [159]. However the methods discussed above make use of modeling techniques that are still tied at the features level. The aspect of decision level has been an uncommon approach when compared to feature level. The approach employed in [99] tackles this issue by proposing multistage classification by taking into account temporal information at the decision level. By taking into account temporal information on the decision level of a multistage system, the classification of a unit (e.g., a video-frame) of an emotion expression improved significantly.

Motivated by this initial promising results, this Chapter extended this approach by capturing the temporal relationship at decision level by proposing a two-stage regression framework.

This method will try decoupling the modeling of the temporal dynamics of an emotional state from the high variability of the feature level by using a two-stage regression framework. In this approach, SVR model, which captures the strength of the features, represents the initial estimation in the first stage regression. Then, it become an input in subsequent model for regression analysis to produce final estimation of affect. By using this approach, it will allow the network to easily exploit the slow changing dynamic between emotional state.

In summary, the main contribution for this Chapter are:

- Develop two-stage regression approach, where it separates the emotional state dynamics modeling from an individual emotional state prediction step based on input features.
- Investigate the effectiveness of subsequent model selected, say TDNN, LSTM and KF and how far it captures the temporal relationship between estimation on continuous instances of each modality selected.

The remainder of the Chapter is organized as follows: Section 5.2 discusses the related works; Section 5.3 presents the methodology of two-stage regression framework; Section 5.4 describes the databases and its corresponding features; Section 5.5 offers an experimental evaluation towards proposed approach; followed by Section 5.6 for results and discussion. Finally, Section 7.1 concludes this work and discusses potential avenues for future work.

5.2 Related Features and Modeling Technique

Automatic continuous emotion recognition typically consists of two systems: feature extraction, which provides low-level representation from given modality and modeling approaches that translate the low-level representation into estimation of affect-related features. Audio features, typically referred to as acoustic low-level descriptors (LLD), include a wide range of features that cover spectral, cepstral, prosodic and voice quality information. Video features, mainly by focusing on facial expression and facial landmark. It can capture the change and intensity of the detected face over the duration of a task. Video features can be divided into two. The first one is video appearance features, where the common example are histogram of oriented gradients (HOG), local binary patterns (LBP), edge orientation histogram (EOH). A variant of LBP features, examined in spatio-temporal volumes of the video after convolving with 2D Gabor filter-banks, (LGBP-TOP), has recently being used as baseline features

for automatic continuous emotion recognition [127] [165]. The second one is video geometric features, where the primary step is to localize and track a dense set of facial points landmarks [165], or shoulder landmark [107], or the whole body expressions landmark [103]. These landmarks are then tracked to acquire low-level descriptors of the dynamics of facial or body gestures.

A physiological signal such as ECG and EDA has been used extensively to measure continuous affect via wearable sensors that can be available at affordable costs. At rest, these signals are tonic in nature, and thus phasic changes are used to measure more immediate stimuli responses. Slowly evolving changes to the tonic frequency for both ECG and EDA signals have been correlated with higher levels of arousal [153]. Additionally, HRHRV can be extracted from the filtered ECG signal. Typically used to quantify physiologic changes in the autonomic nervous system. EDA reflects a rapid, transient response called SCR, as well as a slower, basal drift called SCL [25].

Then, all features will be composed to form a feature vector, which in turn feeds a regular classifier or a regressor. As a result, given a sufficient set of features of the time series, it permits any classical supervised learning algorithm such as linear classifier, k-NN, Support Vector Machine and so on to be applied. This approach has been experimented, for example, in emotion recognition from body movement [12]. Also, measuring level of distress recognition of post-traumatic stress disorder patients by analyzing their speech [169]. Last but not least, predicting a level of depression recognition from audio-visual features [165]. In continuous emotion recognition, it is usually performed with human-annotated arousal and valence as the gold-standard labels. Modeling approaches here are generally supervised, and various regression models have been proposed [44]. The most widely used regression method, which becomes the baseline for continuous affect is Support Vector Regression (SVR) [127] [165]. Other than SVR, Relevance Vector Regression (RVR), or linear regression models [170] are useful in predicting continuous affect.

However, a major disadvantage of feeding a set of descriptive features into regular classifier is that even though features summarizing the whole sequence can provide a meaningful description for the purpose of classification/regression, the temporal structure of the sequence (of emotions) is neglected and not being taken into consideration. For example, a randomly shuffled version of a sequence would result in the same statistical feature representation, but it

would correspond to an entirely different temporal pattern. These regression models are useful in predicting continuous affect, but are insufficient in capturing the temporal information of the affective dimensions. Exploiting the temporal patterns within the sequence of emotions can potentially enhance the final predictive power of the model, for instance, certain transitions of sequence of emotion can carry temporal information and dynamical evolution of these dimensions. Modeling a sequence of emotion is necessary, given the fact that emotion does not change rapidly with respect to time [53].

To overcome this limitation, LSTM-RNN has been successfully applied to continuous emotion estimation [127], since it has the ability to incorporate knowledge about how emotions typically evolve over time so that the inferred emotion estimates are produced under consideration of an optimal amount of context [179]. In their work, automatic continuous emotion estimation from several raters as well as window size in multimodal fusion is analysed. Apart from emotion prediction, LSTM-RNN is also fit for other regression tasks such as speech dereverberation [185] and non-linguistic vocalisation classification [119].

Apart from LSTM-RNN, continuous state-model approaches such as Kalman Filter (KF) [78] is employed to allow a deeper insight into the emotional prediction. In [9], the authors leverage a KF based approach by treating emotional state (x) as a function of time of the features information (z) from the respective modality using the standard state space framework. This model is being used to fuse each of the modality/sensor measurement per time step, and at the same time, it models the time-varying nature of the model to improve system performance further. In [136], good performance was achieved simply by modelling acoustic feature from music using KF. A possible rationale behind the success cases of KF in affect recognition is that its ability to propagate the emotion predictions mean and covariance of the current state in time. On the other hand, only a few parameters are needed to be estimated with a small number of observations for KF modelling. A somewhat less explored NN method is Time Delay Neural Network (TDNN) [173]. TDNN has the ability to capture dynamic relationship between consecutive observations, due to its delay property in the TDN nodes. Emotional expressions which occur at a particular moment are classified by taking into account not only the input features describing that moment but also the input features describe the moment before. This sentence is parallel with [178], whose amongst the first to apply LSTM-RNN in continuous affect recognition, where emotions typically evolve over time. The delay property

in TDNN nodes can be set as the number of past instance of emotional expression, making it a perfect fit for a modelling continuous emotion recognition. However, only [100] fully exploits TDNN structure when modelling the temporal relationship at the semantic level. For each regression model stated above, despite having the ability to learn useful features directly from every modality, they completely ignore temporal information present in continuous affect recognition.

Therefore, in this Chapter, a two-stage regression framework is proposed to try decoupling emotional state dynamics by modelling using the emotional state prediction step based on the input features. The first stage regression model, which is based on SVR, generates the original training prediction based on training a feature vector. Then, the training prediction will become an input to the subsequent model, to learn the temporal information on decision label. The subsequent model, which consists of TDNN, LSTM and KF are chosen because it can incorporate the temporal information of the decision labels, and propagate this information from one-time point to the next (see arrow in second-stage regression). In the second stage, it will learn the expected prediction using the development prediction from the first stage regression approach.

5.3 Methodology

The proposed two-stage regression framework for continuous affect prediction is depicted in Figure 5.1. SVR is being chosen as the first regression model, generates the initial estimation based on the feature vector from each modality. Then, the subsequent model is trained with initial estimation to learn the expected estimation. The rationality behind two-stage regression framework is generally to tackle the high variability that is present in the input features due to the changing in emotional expression, illumination or head pose. It is conducted by decoupling the input features from modelling the temporal information, instead of modelling temporal information only at the decision level. The second reason is to deal with slow changing emotional aspects of the expressions by exploiting the temporal relationship, again only at decision level. To implement the two-stage regression framework, SVR and subsequent model are trained in order. In other words, subsequent model takes the predictive ability of SVR in an account for training. These two-stage models are both built using the same training

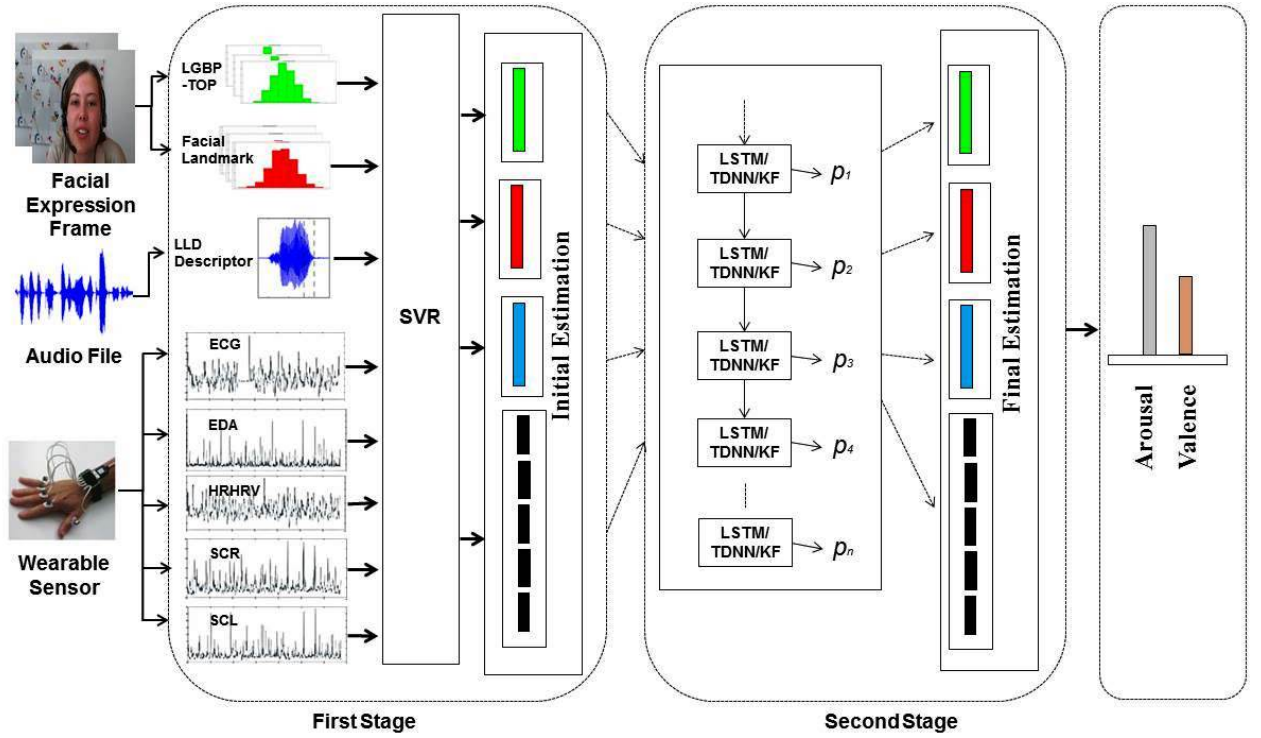


Figure 5.1: Architecture of Two-stage Regression Modeling for Continuous Emotion Recognition

dataset. Once the two-stage models have been trained, it can be treated as a two-layer system where the expected prediction x_t is produced continuously once a unit of sequences of features is received. While the framework should work with any arbitrary modelling technique, SVR has been chosen in the first stage of regression framework since it has been regarded as the baseline approach in the context of continuous emotion recognition. As for the second stage approach, the following subsequent model has been selected; continuous-state model such as KF, and neural network model such as TDNN, and recurrent neural network model such as LSTM-RNN. Each of the models stated above has the ability to propagate information from one time to next, by taking into account past information affective state.

5.3.1 First-Stage Regression

Feature extraction is adopted from each video, audio and physiological sensor modality. In this Chapter, the baseline features are fully utilized. In video modality, LGBP-TOP and facial landmark feature extractor is selected, while in audio modality, features such as low-level descriptor are implemented. As for physiological signal modality, two main features are

extracted that is ECG and EDA signal. HRHRV signal is designed from filtered ECG signal, while SCR and SCL reflected on EDA signal. Each of the feature information is explained clearly in [165].

SVR is implemented by fully utilising liblinear library [37]. However, in order to optimise the SVR generalisation bounds for regression in high-dimension feature space, it mainly depends not only a proper setting of ϵ -insensitive loss, but it also depends on a good predefined hyperparameter C and the chosen kernel parameters. In this experiment, the hyperparameter is chosen empirically to balance the emphasis on the error and generalisation performance. A more in-depth explanation of the SVR is in [27].

5.3.2 Second-Stage regression

In this Chapter, employing second stage regression approach in continuous emotion recognition is done for two reasons. First is to separate the emotional state dynamic from input features. By doing so, the temporal information is not biased by variation presents in the input features. Second is to exploits the slow changing property in emotion dimensions. Emotions are being known to change slowly through time. It is known that, many individual experience a gradual change of emotions from insecurity to security due to presence of trust. Change in an individual perception of the relationship also rely on changes in their emotion. Those gradual changes of emotion is studied in this Chapter.

There is a potential for further improving the efficiency of emotion dimension by classifying frames after short-term time intervals rather than including all frames. Therefore, in this Chapter methods accounting for short-term temporal correlations have been applied. Time Delay Neural Network (TDNN, Long Short Term Memory (LSTM) and Kalman Filter (KF) are employed on second stage regression approach, to cope with the slow varying trajectory of the affect.

Time Delay Neural Network

Time-Delay Neural Network (TDNN) is a type of feedforward network designed to capture dynamics of modelling process [173]. However, feedforward network has no internal memory to store information about the past, thus are insufficient for processing temporal sequences. Therefore, a memory of the past is introduced by extending the network input with the

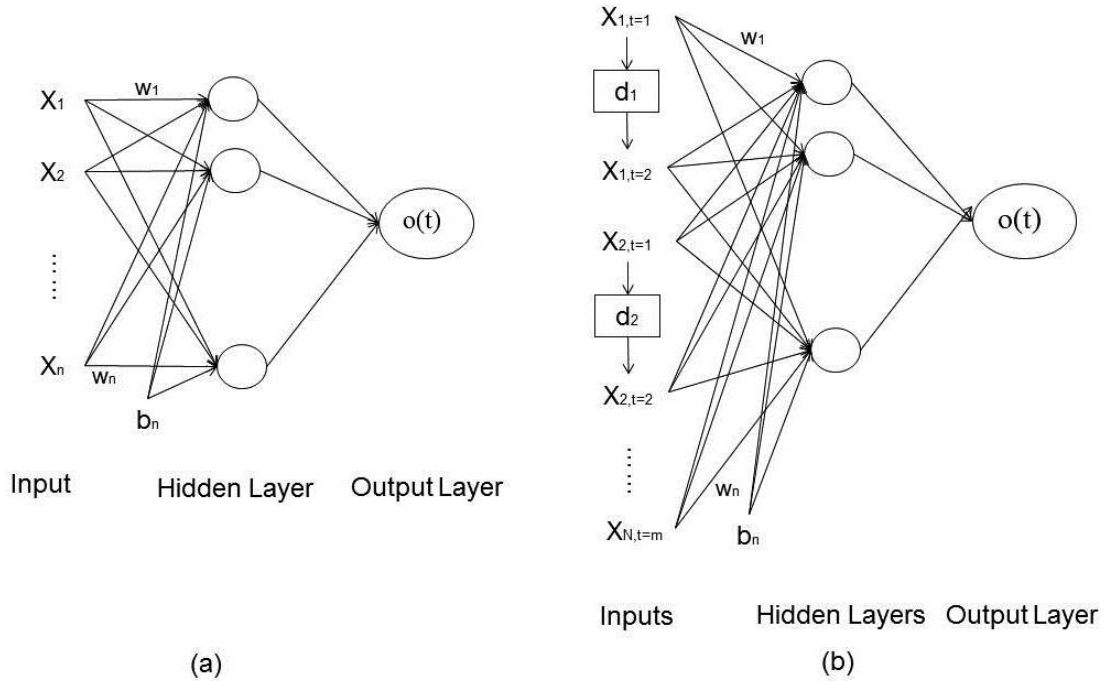


Figure 5.2: (a) Feedforward network (b) feedforward network with delay

previous input, which is known as a delay. The delays attempt to add a temporal dimension to the network. By doing so, TDNN capable of holding past samples of the input signal.

The delay is determined by how many past inputs are stored and how much memory can be correlated to future output. A too small delay may not capture dynamic of the emotion, because the network is blind to anything that happened before, and may create error results. A too large delay may consume more time.

A single TDNN has X_n input, such as $x_1(t), x_2(t), \dots, x_n(t)$ and one output $z(t)$. For each of X_n , there is a bias value, b and delay, d . These delay stores the memory of the past, $X(t-d)$ with $d = 1, 2, \dots, m$ and weights $w = w_1, w_2, \dots, w_n$. A single TDNN can be represent as

$$z(t) = f \sum_{n=1}^N \left[\sum_{d=0}^M X_n(t-d) * w_n d + b_n \right] \quad (5.1)$$

From Equation 5.1, it can be seen that the input X at the current time step t and previous time step $(t-d)$ contribute to the overall outcome of the neuron. This characteristic is needed to model dynamic emotion behaviour in continuous emotion recognition.

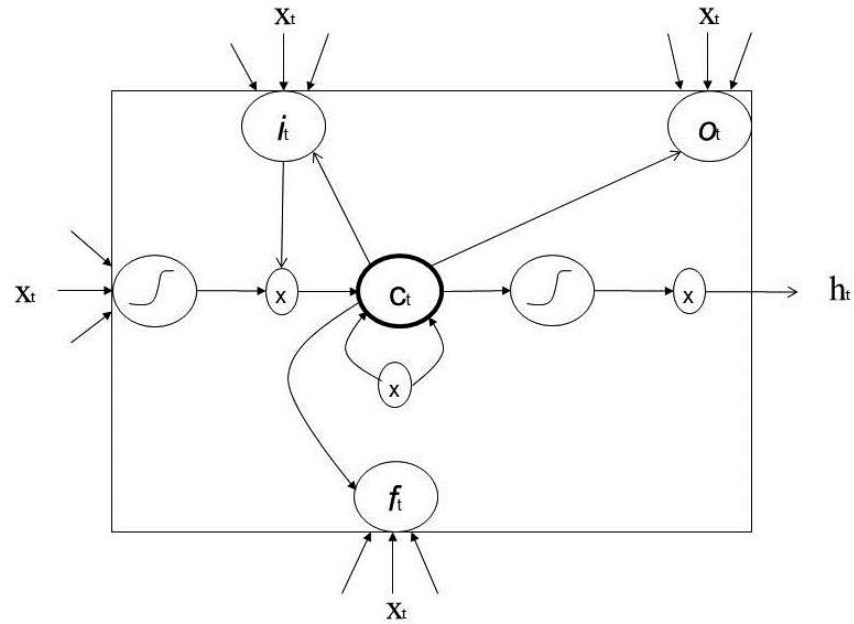


Figure 5.3 Each LSTM cell remembers a single floating point value c_t (Equation 5.6). This value may be diminished or erased through a multiplicative interaction with the forget gate f_t (Equation 5.5) or additively modified by the current input x_t multiplied by the activation of the input gate i_t (Equation 5.4). The output gate o_t controls the emission of h_t , the stored memory c_t transformed by the hyperbolic tangent nonlinearity (Equation 5.7 5.8). Images are reproduced from [43]

Long Short Term Memory

As noted before, emotions are inherently temporal. Even though TDNN are sequentially in nature, these architecture are implemented as feed-forward neural networks, where the flow of data in only one direction as in Figure 5.3 (b), forward from the input nodes through the hidden nodes and to the output nodes. There are no cycles or loops in TDNN network. Producing the cycles or loops in the network are necessary, in order to process the sequence of input features into a single, fixed-length vector.

In this Chapter, Long Short Term Memory is proposed to consider the sequence of initial estimation from first stage regression explicitly. Given an input sequences $\mathbf{x} = (x_1, x_2, \dots, x_T)$ a standard recurrent neural network computes the hidden vector sequence $\mathbf{h} = (h_1, h_2, \dots, h_T)$ and output vector sequence $\mathbf{y} = (y_1, y_2, \dots, y_T)$ by iterating the following Equations from $t = 1$ to T :

$$h_t = \mathcal{H}(W_{ih}x_t + W_{hh}h_{t-1} + b_h) \quad (5.2)$$

$$y_t = W_{ho}h_t + b_o \quad (5.3)$$

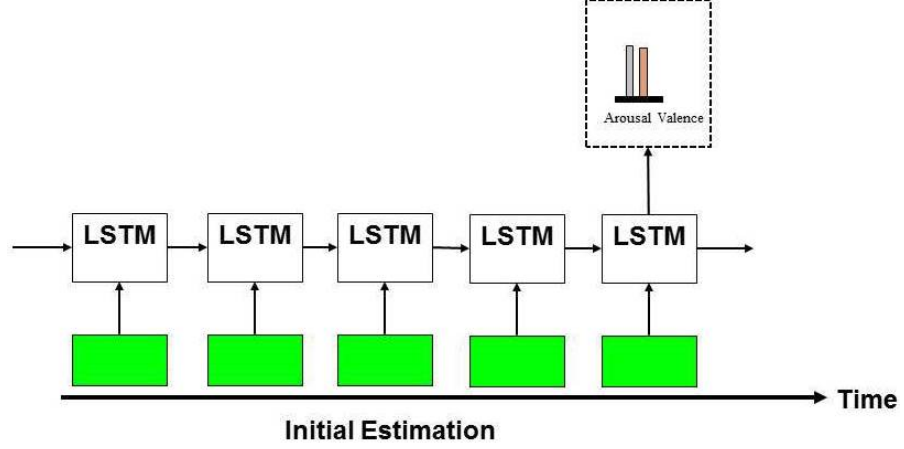


Figure 5.4 :At each time step, the input features are fed into the LSTM to compute the final estimation of arousal and valence.

where W denote weight matrices, b denote bias vectors and \mathcal{H} is the hidden layer activation function. In this experiment, the logistic sigmoid function is being employed in the architecture. Noted that W_{ih} is input hidden weight matrices, and b_h is biased in the hidden vector.

In LSTM architecture, memory cells are adopted to store an output information, allowing it to discover long-range temporal relationships in emotion dimension better. The hidden layer \mathcal{H} of LSTM is computed as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (5.4)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (5.5)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5.6)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (5.7)$$

$$h_t = o_t \tanh(c_t) \quad (5.8)$$

where σ is the function in terms of logistic sigmoid function and i , f , o , and c are respectively the *input gate*, *forget gate*, *output gate*, and *cell activation vectors*. The value stored in LSTM cell c is maintained unless it is added to by the input gate i or diminished by the forget gate f . The output gate o controls the emission of the memory value from the LSTM cell.

Figure 5.4 illustrates the proposed two-stage regression approach. LSTM with the tanh layer that computes arousal and valence score based on an initial estimation of the current time step and the hidden states and memory of the LSTM from the previous time step. To accurately detect the complete duration of continuous emotion activity, especially for relatively long and complex ones, it is essential for the model to capture the progression patterns of activities during training. An LSTM considers progression via the context that is passed along time in the form of the previous hidden state and memory as well. At every time step, it outputs a hidden representation, which encodes the input information from previous time step. Chou et al.,[15] proposed to averaged all the hidden representations from different time steps to have better results. However, in the proposed approach, when dealing with regression time series issues, final estimation is often calculated by the hidden representation of the last time step, as shown in Figure 5.4. Given the time-step is N , final estimation is performed only at the last time-step of the input sequences, whereas the previous $N - 1$ are automatically ignored by the system.

Kalman Filter

In an analysis of emotion, it has been found that human faces, in particular, are dynamic in nature [131]. It gives an intuition that emotions not only inherently temporal but are dynamic and evolve across time. However, in automatic continuous emotion recognition, system considering emotion dynamic has been relatively less explored. In this Chapter, initial estimation is modelled as observations in Kalman Filter to track each emotion dimensions continuously.

Emotion dynamic can be described by a state equation and observation equation as follows:

$$x_{t+1} = F_t x_t + w_t \quad (5.9)$$

where F_t is state transition matrix, x_t is the state of time t and w_t is process noise. To further up the assumption, the observation of the state can be made through a measurement system which can be represented by a linear equation in the form:

$$z_t = H_t x_t + v_t \quad (5.10)$$

where z_t is the observation or the measurement made at time t . x_t is the state of time t , H_t is the observation matrix and v_t is additive measurement noise.

There are few assumption to be made regarding Kalman Filter in second stage approach. To simplify the equation, F_t and H_t is being assumed as constant 1, $F_t = 1$; $H_t = 1$, the process and measurement noise random processes, w_t and v_t are uncorrelated, zero-mean, white noise processes with known covariance matrices, $w_t \sim N(0, Q_t)$ and $v_t \sim N(0, R_t)$. The idea of using KF is basically to smoothen the signal, in this context; is to find the optimal value of estimation of noise covariance Q_t and R_t . So in Equation 5.9, x_t can be treated as a surrogate for the unknown state, which can be assumed as gold-standard provided by AVEC 2016. In Equation 5.10, z_t is assumed to be the initial estimation from the first stage, and the difference between z_t and x_t , become the surrogate of the noise. Using all the information above, the estimation of noise covariance Q_t and R_t can be estimated easily. In this approach, the estimation of process covariance Q_t and noise covariance R_t of the prediction is estimated at the current index, or time t over a lookback window N using the gold-standard as a surrogate for the true process state. This method is called heuristic approach, where held out data over a lookout window is used to determine noise terms. At each lookback window N , a new estimation of the process covariance Q_t and noise covariance R_t is produced using Equation below

$$Q_t = cov(x_{t=2,N} - Fx_{t=1,N-1}) \quad (5.11)$$

$$R_t = cov(z_{t=1,N} - Hx_{t=1,N}) \quad (5.12)$$

Noted that, Q_t and R_t is produced from training prediction, therefore it can be incorporated as *predict* and *update* of the Kalman Filter iteration, where the input is the continuous initial estimation from the first stage.

5.4 Dataset and Features

In this experiment - RECOLA [128] has been adopted as standard dataset for the Audio-Visual Emotion Challenge in 2015/2016 [165] [127]. This database has been designed to study socio-affective behaviour, with more focused on multimodal data. In this database, it was recorded in in spontaneous interaction mode, during resolving of a collaborative task remotely through video conference. The corpus consists of multimodal data, such as; audio,

video, and physiological signal. It was recorded continuously and synchronously from 27 French-speaking participants, ranging from three nationalities/mother-tongue; French, Italian, German, to provide diversity affect in the dataset. To ensure balance data, the corpus were equally divided into three partitions (training, development, testing), with each partition containing nine recordings with approximately balanced for gender, age, and mother-tongue of the participants. The data is labelled in two affective dimensions, namely arousal and valence, and was manually annotated using a slider-based label tool. A combination of these individual ratings is used as the gold-standard label. The dataset is provided together with a set of pre-calculated features, in each modality, which can be incorporated into regression stage. In this Chapter, optimising the features is beyond the scope of this Chapter, because the main focus is to exploit the slow-changing temporal information in continuous affect recognition, by using baseline features.

5.4.1 Audio Features

Audio features are computed with openSMILE [34] and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [33]. LLD descriptors cover spectral, cepstral, prosodic and voice quality information are extracted as shown in Table 5.1. Overlapping fixed length segments, which shifted forward at a rate of 40ms is applied to deal with continuous recording. The arithmetic mean and the coefficient of variation is computed on all 42 LLD to extract functionals. To pitch and loudness the following functionals are additionally applied: percentiles 20, 50 and 80, the range of percentiles 20-80 and the mean and standard deviation of the slope of rising/falling signal parts. Functionals applied to the pitch, jitter, shimmer, and all formant related LLDs are applied to voiced regions only. The average RMS energy is computed, and six temporal features are included, which are; the rate of loudness peaks per second, mean length and standard deviation of continuous voiced and unvoiced segments and the rate of voiced segments per second, approximating the pseudo-syllable rate. Overall, the acoustic baseline features set contains 88 features.

5.4.2 Video Features

The RECOLA dataset provides a set of video features consisting of appearance features, namely Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [2], as well

Table 5.1: 32 Acoustic Low-Level Descriptor (LLDs)

1 energy-related LLD	Group
Sum of auditory spectrum (loudness)	Prosodic
19 spectral-related LLDs	Group
α ratio (50–1000Hz / 1–5kHz) ¹	Spectral
Energy slope (0–500Hz, 0.5–1.5kHz) ¹	Spectral
Hammarberg index ¹	Spectral
MFCC 1–4 ²	Cepstral
Spectral flux ³	Spectral
12 voicing-related LLDs	Group
F_0 (semi-tone)	Prosodic
Formants 1, 2, 3 (freq., bandwidth, ampl.)	Voice Quality
Harmonic difference H1–H2, H1–A3	Voice Quality
Log. HNR, jitter (local), shimmer (local)	Voice Quality

¹ computed on voiced and unvoiced frames.

² computed on voiced and all frames.

³ computed on voiced, unvoiced, and all frames.

as geometric features computed from 49 landmarks tracked in the video sequences using a Supervised Descent Method (SDM) [180].

LGBP-TOP is a dynamic appearance descriptor that is robust to illumination changes and misalignment [2]. To perform LGBP-TOP, the first step was to do a convolution with a number of Gabor filters on top of video frames. Then, LBP descriptor is applied through the set of Gabor magnitude response images. The resulting binary patterns are combined using histogram then concatenated into a single feature histogram. To be consistent with audio features, these features were calculated at a step size of 0.4s. Finally, Principal Component Analysis (PCA) is applied, obtained 84 features representing 98% of the total variance. As for video geometric features, 49 facial landmarks were tracked. as illustrated in Fig. 5.5. The detected face regions included left and right eyebrows (five points respectively), the nose (nine points), the left and right eyes (six points respectively), the outer mouth (12 points), and the inner mouth (six points). Then, the landmarks were aligned with a mouth (six points). Again, the landmarks were aligned with a mean shape from stable points (located on the eye corners and the nose region).

For each detected face in video frames, 316 features were extracted, which consists of three parts. The first one is 196 features computed by taking the difference between the coordinates of the aligned landmarks and those from the mean shape and between the aligned

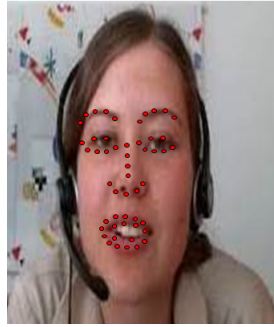


Figure 5.5: Illustration of the facial landmark features extraction from RECOLA dataset

landmark locations in the two consecutive frames. The second one is 71 features calculated on the Euclidean distances (L2-norm) and the angles (in radians) between the points in three different groups. The third one is 49 features computed by taking the Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame. Again, to keep the consistency with audio features, the mean and variance were computed over the sequential 316 features within a fixed length window (8s) that shifted forward at a rate of 0.4s.

5.4.3 Physiological Signal

The physiology features which consists of ECG signals, EDA signals, HRHRV signals, SCR and SCL signals were adopted precisely as in [165], with the size of a feature depends on each modality respectively.

The first physiological features are ECG, recorded the electrical activity of the heart. In this signal, it consists of the total of 19 features. They are zero-crossing rate, the four first statistical moments, the normalized length density, the non-stationary index, the spectral entropy, slope, mean frequency plus 6 spectral coefficients, the power in low frequency (LF, 0.04- 0.15Hz), high frequency (HF, 0.15-0.4Hz) and the LF/HF power ratio.

The second physiological features are EDA, which was derived from continuous variation in the electrical characteristics of the skin. From this signal, 8 features has been extracted including the four first statistical moments from the original time-series and its first order derivative.

The third physiological features are HRHRV, which is extracted from the heart rate and its measure of variability. It is being derived from filtered ECG signal by applying zero-delay

bandpass filter (3-27Hz) on the signal. 10 features have been extracted including the two first statistical moments, the arithmetic mean of rising and falling slope, and the percentage of rising values for each of those two descriptors.

The fourth physiological features are SCR. SCR is the event where the skin momentarily becomes a better conductor when either external or internal stimuli occur that are physiologically arousing. 8 features were extracted, including the four first statistical moments from the original time-series and its first order derivative.

The fifth physiological features are SCL, where its directly controlled by the sympathetic nervous system and indicates the activity of the sweat glands in the skin. 8 features were extracted, including the four first statistical moments from the original time-series and its first order derivative.

5.5 Experimental Evaluation

Separate continuous emotion recognition, arousal and valence estimations are obtained from individual modalities as described in the AVEC 2016 paper [165]. Firstly, the mean and variance were standardise on all features from training, development and testing partitions. It is to ensure that the data is normally distributed in all partitions. The regression task is performed using linear SVR provided with the liblinear library [37] in the unimodal setting. Then, second stage estimation is incorporated using subsequent model, such as TDNN, LSTM-RNN and KF, as shown in Figure 5.1. The initial estimation from first-stage is used as an input to incorporate with the second-stage approach, to capture the temporal information in the unimodal setting.

To demonstrate the effectiveness of proposed framework, firstly SVR modelling were individually trained in unimodal settings. SVR is implemented with linear kernel and trained with L2-regularised L2-loss dual solver, while all other parameters were kept the same. The complexity (C) was chosen in the range of [.00001, .00002, .00005, .0001 . . . , .2, .5, 1] and it is optimized by the best performance in development set. Note that all parameter were same for each modality in the unimodal setting. Then, each of the prediction outputs from SVR become an input into the subsequent model, that is TDNN, LSTM-RNN and KF models.

Estimation in TDNN is incorporated by taking into account the label assigned in previous

frames by the first-stage estimation, SVR.

During the training process, the output of the first-stage prediction (SVR) is used to train the model in second stage prediction (TDNN). Once the two models are trained, these two models can be treated in tandem as a two-layered system where the prediction value is produced continuously when the sequence of expressions is received. The implementation of TDNN is achieved by using MATLAB Neural Network Toolbox and by experimentally setting the parameters. To reduce the computational complexity, all the parameter such as the number of input nodes, the number of hidden nodes, the number of output nodes, the number of hidden delays are set to be the same setting for training, development and test partitions.

Estimation in LSTM-RNN is incorporated by taking into account the label assigned in previous frames by SVR. The parameter of W and b in Equation 5.2 can be learned by back-propagating through time from first- stage prediction, with the sum of the squared deviation between the output \hat{y}_t and output from the first-stage prediction y_t at $t = 1, \dots, T$ as the error function. The parameter of LSTM-RNN model are as follows: it has a visible layer with one input, a hidden layer with 64 LSTM blocks or neurons, and an output layer that makes a final prediction. The default sigmoid activation function is used for the LSTM blocks. The network is trained for 100 epochs, and a batch size of 24 is used. The training procedure was performed with KERAS Neural Network Library [19].

Unlike two modelling stated above, where the prediction is taken by taking into account past observation, KF modelling is act as a model to estimate emotional state as a function of time. In this approach, emotional state (x) is determined as a function of time from the information (z) from respective modality using the standard-space framework.

When generating estimation, post-processing such as smoothing, centering and scaling is being implemented towards estimation. Each post-processing step was kept and applied to the testing set if it improved the CCC score on the development set.

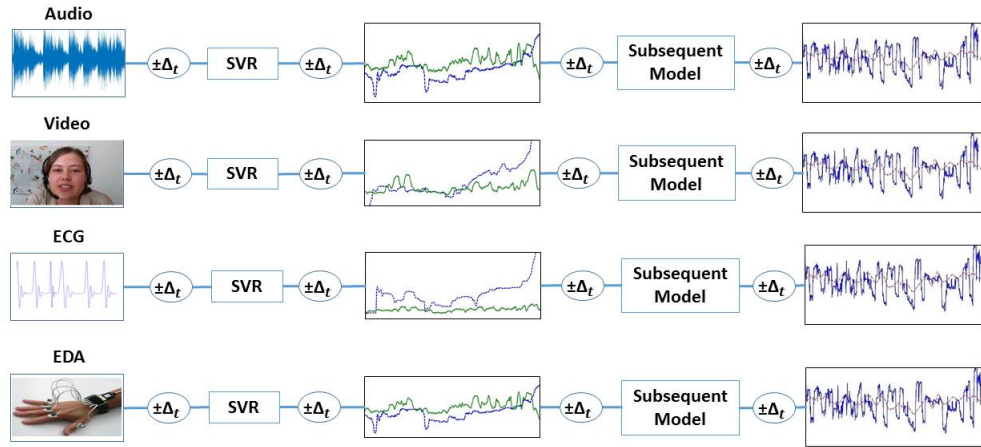


Figure 5.6 :Architecture of Two-stage Regression Modeling of Continuous Affect Recognition by using RECOLA dataset, Video provides two set of features, LBPTOP and video geometric features. Bio-signal such as ECG and EDA reflected by HRHRV and SCR with SCL respectively [165]

5.6 Results and Discussion

In this Chapter, The results of the proposed approach is reported in terms of *Concordance Correlation Coefficient* [127] metric:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (5.13)$$

where ρ is the *Pearsons Correlation Coefficient* between two time series (e.g: prediction and gold-standard); μ_x and μ_y are the means of each time series; and σ_x^2 and σ_y^2 are the corresponding variance. Unless mentioned otherwise, the proposed approach is being trained using training partitions and is being tested using development partitions. At the current stages, the test label for test sets is not being released yet from the AVEC organizer. Therefore all the results reported in this Chapter are solely based on development partitions.

The CCC values for unimodal performance in predicting arousal and valence separately are shown in Table 5.2. Results show comparable unimodal performance in LLD features, where no significant improvement in performance is observed here due to the audio recordings being mostly clean speech. Second-stage regression approach performs somewhat better on LGBP-TOP, geometric and physiological signal. As it can be seen, the second-stage regression

set-ups either match or outperform their corresponding individual baseline SVR models in the majority of the cases. The improvement could be because the initial baseline models such as SVR prediction actually helps the subsequent model to avoid its local minima.

Table 5.2: Comparison of baseline results with two-stage regression approach of unimodal performance on development sets

Features	Arousal			
	Baseline-SVR [165]	SVR-TDNN	SVR-LSTM	SVR-KF
LLD	0.796	0.794	0.769	0.780
LGBP-TOP	0.483	0.519	0.504	0.524
Geometric	0.379	0.430	0.402	0.415
ECG	0.288	0.356	0.325	0.337
HRHRV	0.382	0.427	0.417	0.423
EDA	0.077	0.125	0.130	0.163
SCL	0.101	0.105	0.161	0.247
SCR	0.071	0.157	0.175	0.214
Features	Valence			
	Baseline-SVR [165]	SVR-TDNN	SVR-LSTM	SVR-KF
LLD	0.455	0.381	0.428	0.432
LGBP-TOP	0.474	0.462	0.467	0.481
Geometric	0.612	0.624	0.630	0.638
ECG	0.153	0.130	0.153	0.153
HRHRV	0.293	0.274	0.293	0.298
EDA	0.104	0.188	0.194	0.197
SCL	0.124	0.156	0.166	0.170
SCR	0.110	0.066	0.085	0.086

In the arousal dimension, closer inspection of Table 5.2 indicates that the second-stage regression approach predominantly gives higher performance compared to the baseline in all modalities except audio. More specific, in each subsequent model, SVR-TDNN gives higher performance compared to SVR-LSTM and SVR-KF, in video-appearance, video-geometric, ECG and HRHRV modality. These results further support the idea of TDNN, where the model can capture dynamic behaviour between consecutive affective states through delay. It is noted that the delay is determined by experimentally setting the parameter. Detailed analysis of the number of delays chosen would be the part of future work. However, in contrast with earlier statements, SVR-KF gives superior results on EDA, SCL and SCR modality, compared with baseline, SVR-TDNN and SVR-LSTM. While the noisy observations or measurements in this context are the unimodal predictions acquired in the 1-stage approach, KF is able to model the process and measurement noise as a Gaussian efficiently, thus giving the optimal solution towards the performance, compared with the other two subsequent model.

Similarly, the second-stage approach brought additional performance improvements when

compared to baseline results on the valence dimension, although not as obvious as for the arousal dimension. From Table 5.2, it shows that video geometric modality appears more informative than audio modality. SVR-KF is the only models that contribute to the higher results across video-appearance, video-geometric, HRHRV, SCL and SCR modality, but not as significant as in arousal dimensions.

The relatively lower performance of predicting valence when compared to arousal dimensions in second-stage approach is likely due to the subsequent model which cannot easily capture the temporal aspect of the emotional expressions when mapping it to the valence dimensions. Taking a deeper look into its gold standard, it appears that, with an only small time interval, valence dimensional has a lot of starts, peak, and end points, making it harder for the model to capture temporal dynamics, and leads to the lower performance of improvement when compared to arousal dimensions. It is possible to hypothesise that having a high-level feature extraction method can possibly lead to further improvements of the results, but it is outside the scope of this Chapter. The future works would involve additional experiments about high-level features which can verify these assumptions.

Apparently, what stands out from Table 5.2 is that SVR-KF is the only subsequent model that gives higher performance when compared to the other two subsequent model. It is shown that for physiological signals segments containing emotion changes are more salient and useful for estimating emotions dimensions, especially for valence. These results confirm the assumption of KF that evolution of valence dimensions to be a linear dynamic system, it is explained by the fact that the ability of state dynamic means and covariance chosen in KF able to propagate through time, therefore, preserving the dynamic evolution of valence dimensions being tracked.

Looking back at the previous assumption that states that emotions typically have a slow dynamic. In the arousal dimension, SVR-TDNN and SVR-LSTM are able to model this temporal information of the sequence of emotion and storing them as a memory for future prediction. In valence dimension, however, it is matched with the ability of SVR-KF where KF is able to propagate the means and covariances as a function of time.

To further highlight advantages of 2-stage regression architecture, Table 5.3 and Table 5.4 shows the comparison of performance on online tracking by Kalman Filter [153] with SVR-KF and LSTM on physiological sensors with SVR-LSTM [9], respectively. From Table 5.3, in

Table 5.3: Comparison of unimodal performance by Somandepalli et al [153] and SVR-KF on the development set.

Features	Arousal		Valence	
	[153]	SVR-KF	[153]	SVR-KF
LLD	0.800 ¹	0.780	0.448 ¹	0.432
LGBP-TOP	0.481	0.524	0.474	0.481
Geometric	0.297	0.415	0.612	0.638
D-ECG	0.272	0.337	0.159	0.153
D-EDA	0.080 ²	0.163	0.178 ²	0.197
D-HRHRV	0.383	0.423	0.298	0.298

¹ computed on Teager energy-based MFCC and LLD

² computed on sparse EDA

arousal dimension, one can observe that SVR-KF achieves a higher CCC value than online tracking by Kalman Filter [153] for LGBP-TOP, geometric, ECG, EDA and HRHRV features. As for the valence dimension, it has achieved a balanced performance, SVR-KF performs better in LGBP-TOP, geometric and EDA features, while online affect tracking using KF is superior on audio and ECG features. Surprisingly, it achieved similar performances of CCC for both SVR-KF and online affect tracking in the HRHRV modality. From Table 5.4, similar observations were seen where SVR-LSTM gives superior results on arousal dimensions for both HRHRV and EDA modality. In valence dimensions, applying LSTM directly on physiological sensor features [9] gives superior results on HRHRV modality. However, SVR-LSTM gives comparable results on EDA modality.

Table 5.4: Comparison of unimodal performance by Brady et al. and SVR-LSTM on the development set.

Features	Arousal		Valence	
	[9]	SVR-LSTM	[9]	SVR-LSTM
HRHRV	0.357	0.417	0.364	0.293
EDA	0.082	0.130	0.177	0.194

5.7 Comparison on the best performer of the Challenge

The proposed framework is compared with the state-of-the-art, according to baseline feature in Figure 5.7 for arousal and Figure 5.8 for valence emotion dimension. In arousal dimension, features from physiological modality such as HRHRV, SCL and SCR gives best performance

using two-stage regression approach, while in valence dimension, it gives comparable performance across all baseline features. The best results from Figure 5.7 and Figure 5.8 are from Weber et al. [174]. In their case, they fully exploited fusion on each subject on development sets for each features, while in the proposed approach, no fusion is allowed, as the main focus is to investigate two-stage regression approach in each of the baseline features.

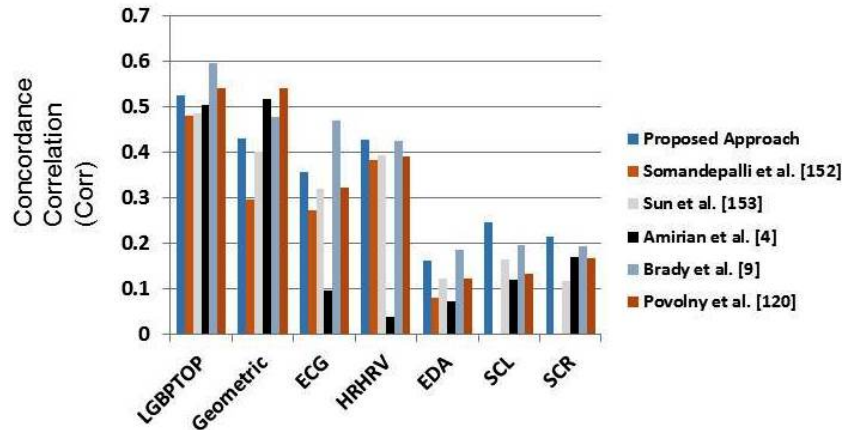


Figure 5.7 Bar chart comparison with state-of-the-art in AVEC 2016 in terms of concordance correlation in Arousal dimension. Results reported based on baseline features.

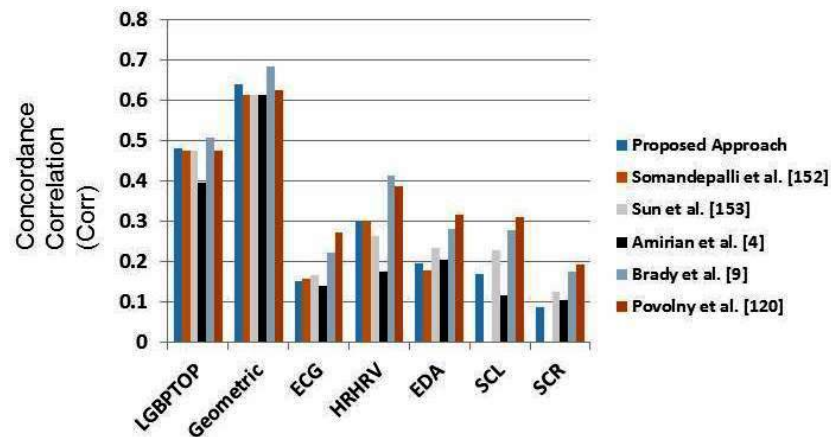


Figure 5.8 Bar chart comparison with state-of-the-art in AVEC 2016 in terms of concordance correlation in Valence dimension. Results reported based on baseline features.

5.8 Chapter Summary

This Chapter proposed and investigated 2-stage regression approach, for continuous emotion recognition from multiple modalities. This is achieved by firstly concatenating the strength

Table 5.5: Comparison of CCC on fusion of AVEC 2016 state-of-the-art and proposed approach.

All	Method
Somandepalli et al. [153]	Conditional Online Tracking
Sun et al. [154]	SVR with L1 L2
Amirian et al. [4]	RF+late Echo State
Brady et al. [9]	LSTM
Huang et al. [69]	OA framework
Povolny et al. [121]	LR with PCA
Proposed Approach	SVR + subsequent model

of an initial SVR model, as represented by its initial estimation, then using it as a basis for another regression analysis in a subsequent model. To demonstrate the suitability of the framework, the benefits from the state-of-the-art regression model in continuous emotion analysis, SVR is jointly explored with three other subsequent models, TDNN, LSTM and KF. Furthermore, these 2-stage regression approaches were evaluated in both unimodal settings, on eight different modalities, such as audio, video-appearance, video-geometric, ECG, HRHRV, EDA, SCL and SCR modality.

Results gained from RECOLA database indicate that the 2-stage regression approach can match or outperform the corresponding conventional SVR models when performing naturalistic affect recognition. An interesting observation was, apart from the audio modality, that all three subsequent models outperform baseline results of arousal dimension, and give competitive results towards baseline results of valence dimension in the unimodal setting. This demonstrates the effectiveness of proposed framework, in terms of capturing temporal information in prediction of emotion recognition and being able to work with a different combination of regression strategies.

There is a wide range of possible future research direction associated with 2-stage regression framework to build on this initial set of promising results. First, only baseline features in AVEC 2016 were investigated. Deep learning features can be taken into consideration. On top of that, accessing the suitability of more regression approach, which can capture the temporal dynamic of the prediction such as Particle Filter or Hidden Markov Model will be taken into consideration. In addition, it is interesting to extend the framework, not only on utilising the temporal information on decision level, but utilising temporal information on features level. For future works, the combination/fusion of all individual predictions will be investigated.

Chapter 6

Linear and Non-linear Multimodal Fusion for Continuous Affect Estimation in-the-Wild

This Chapter leverage continuous emotion recognition in-the-wild setting, by investigating mathematical modelling for each emotion dimension. Linear Regression, Exponent Weighted Decision Fusion, and Multi-Gene Genetic Programming are implemented to quantify the relationship between each affect. The proposed fusion methods were applied in the public Sentiment Analysis in the Wild (SEWA) multimodal dataset and the experimental results indicate that employing proper fusion can deliver a significant performance improvement for all affect estimation.

6.1 Introduction

Affective computing is a field of research that aims to enable intelligent systems to recognise, feel, infer and interpret human emotions. Early works in this field have focused on the recognition in terms of basic emotional states (6 basic emotion: *angry, disgust, fear, surprise, happy, sad, neutral*) and on the data collected in laboratory settings and acted data [47] [141] [142] [184]. Recent developments of sensors like camera and microphone have led to a renewed interest in emotion recognition, from recognizing discrete basic emotion to recognizing continuous emotion, or continuous affect estimation, in terms of arousal and valence [46] [48] [119]

[175] [151].

Numerous studies have been performed to compare the advantages offered by a wide range of modelling techniques for continuous affect recognition [107] [125] [124]. The introduction of Audio Visual Emotion Challenge (AVEC), in 2011 [144] marks the advancement of this continuous affect estimation. AVEC challenge aims to create a benchmark to evaluate modelling systems that are capable of recognising affect recognition beyond laboratory conditions. Using Pearson Correlation Coefficient (P_{corr}) as an objective function, the best results obtained by Nicolle et al., [111], which is 0.456. However, the best results for AVEC 2013 challenge were dropped to 0.1409 [101], but later improved to 0.5946 in AVEC 2014. From the 2015 edition of the AVEC challenge, Concordance Correlation Coefficient (C_{corr}) has been used as an official scoring metric of AVEC. It is mainly because Concordance Correlation reflects on a bias, variance and scaling issues of prediction and gold standard whereas Pearson Correlation is insensitive of those three.

This Chapter describes a multimodal approach on SEWA dataset, by leveraging the individual advantages of each modality, then quantifying the relationship between each modality. In this Chapter, decision fusion on an initial prediction by employing linear and non-linear fusion approach is applied. Some researchers advocate that combined multiple modalities will contribute to the recognition accuracy, and it can be achieved in a numerous way. Method for simple mapping such as averaging [75] to complex method such as linear regression [165] [126], SVR [123], random forests [11] or KF based [9] [153] has been used to combine prediction from multiple modalities. However, a systematic understanding of the relationship between modalities contribute to the higher recognition accuracy is not fully explored. Few kinds of literature build the fusion model using linear regression method. [165] [126]. However, these methods assumed that the continuous affect label is linear in time. Looking closely at the gold standard affect label in [75], potential nonlinearities behaviour may occur in continuous affect label. In other words, the contributions of this Chapter are three-fold:

- Investigated linear and non-linear behavioural modelling approach to model unimodal prediction from trained regressor to predict each affect dimension in multimodal fusion manner.
- Examined the possibility of constructing affect estimation prediction equation from initial prediction result. These modelling equations can provide a convenient way to express

the relationship between each modality and affect estimation in multimodal fusion manner.

- Critically view on the conclusion that arousal can be much better predicted using audio modality and valence can perform much better using video modality. By examine the weighted value for each modality, such conclusion can hold different context or different data.

The rest of this Chapter is organised as follows. Related work on video and audio based facial expression recognition in-the-wild settings is discussed in Section 6.2. Then, Section 6.3 briefly describes the database used. Meanwhile Section 6.4 gives complete detail on decision fusion approach. Section 6.5 evaluates the proposed algorithm, by producing mathematical equation from unimodal estimation. Finally, Section 6.7 conclude Chapter 6 along with its limitation and future works.

6.2 Related Works

The evolution of continuous affect estimation usually comprises of two system: standard features extraction methods which are grounded in statistical/mathematical notions, and modern machine learning which is based on algorithms from artificial intelligence field. In the literature of continuous affect recognition, typically there are two modalities present to estimate affect, audio, and visual modality [48]. Audio modality usually represented by audio features such as acoustic low-level descriptors (LLD), which include broader features such as spectral, cepstral, prosodic and voice quality information. As for video modality, it typically referred as video features which consist of appearance feature and geometric feature. Video modality also captures the change and intensity of facial expressions over time. For appearance feature, the most popular example would be local binary patterns (LBP) and a histogram of gradients (HOG) modelled using bag-of-words (BOW). A robust variant of Local Binary Pattern (LBP), called Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) is incorporated in spatio-temporal volumes of the video after convolving with 2D Gabor filter-bank. LGBP-TOP has been used as baseline feature in automatic affect recognition challenge [125] [165]. Video geometric features include identifying landmarks on the face [165] or shoulder [107] or the whole body [103].

Apart from audiovisual modality, physiological recording is introduced in recent literature of estimation arousal and valence [125] [165]. Electrocardiogram (ECG) and electro-dermal activity (EDA) recording are known to have been correlated with the higher level of Arousal. Heart rate (HR) and heart rate variability (HRHRV) extracted from the ECG signal are the most often reported emotion indicators, followed by skin conductance level (SCL) [87]. SCL provides the activity of the sweat glands in the skin and is controlled by the nervous system.

Experimenting with text modality is quite a new approach in continuous affect recognition. The semantics of the words used can be an important aspect of emotion detection. It is because the words chosen can say a lot on the current state of emotion of the person. In previous AVEC 2016 challenge, only Povolny et al. [121] addressing text feature by exploring automatic speech recognition, lexicon-based approach, and word embedding technique, to create a dictionary for each utterance. Recent AVEC 2017 [126] challenge introduce new modality, that is text modality as one of the baseline features. It is provided by using bag-of-words text feature representation based on the transcription of the speech as text features. The necessity of text modality in emotion recognition is quite straight-forward. For example, sob and laughter can reflect the current emotion state of the person indeed. In [22] the Suite of Linguistic Analysis Tools (SALAT) [91] were utilised extensively for emotion investigations for extracting a range of text-based features. However, the author only focuses on liking dimension, other than arousal and valence dimension, when employing SALAT tools in continuous emotion system. Word vector is also introduced in [18], where distributional word representations are learned from the massive textual dataset.

Signals such ECG and EDA recording carries information where slowly evolving changes to the tonic frequency for both ECG and EDA signals have been correlated with higher levels of arousal. Heart rate (HR; tonic) and heart rate variability (HRV; phasic) extracted from the ECG signal are typically used to quantify physiologic changes in the autonomic nervous system. Skin conductance level (SCL; tonic) and measures of skin conductance response (SCR; phasic) provide a complementary view [153].

Affect estimation is usually performed with human-annotated Arousal and Valence for gold standard ratings. Modeling approaches here are generally supervised, and regression-based method is the approach of estimating affect. SVR is perhaps the most widely used regression method for affect estimation and has been regarded as baseline approach for affect estimation

[125] [145] [165]. Recent literature takes into account short term temporal correlation such as Continuous Conditional Random Fields (CCRF) on top of SVRs [6] and various type of neural network including TDNN [100], RNN [17] and LSTM-RNN in [108] [17] [123]. Another study [107], employed a bidirectional LSTM model with an output-associative framework to achieve improved performance in affect prediction. Following this trend, a deep bidirectional LSTM was proposed [58] in which was gives the highest results in [125].

When dealing with several modality and modelling technique, the question of how to fuse them arises. Feature level fusion and decision level fusion is the most well-known approach for assessing continuous affect estimation. Feature level fusion is undertaken solely by concatenating each feature from multiple modalities then a single classifier is trained on the concatenated features [17] [68]. However, feature-level fusion is plagued by several challenges. Commonly, this approach tends to create very high dimensional feature vectors and lead to overfitting. Secondly, features from multiple modalities are obtained at different time scales. For example, HRV features from physiological modality typically extracted in minutes [115] while LLD features from audio modality can be in the order of milliseconds [165]. Recently, feature level fusion based on Canonical Correlation Analysis (CCA) has attracted the attention in the area of emotion recognition [55] [54]. CCA is incorporated by using the correlation between two sets of features to find two sets of transformations such that the transformed features have maximum correlation across the two feature sets while being uncorrelated within each feature set.

The second fusion approach, decision fusion is the process of first generating separate estimations fusing them into one final estimation. Each estimation from multiple modalities can be independently generated using separate models, and the results are joined using a multitude of possible methods. In this case, the fusion of prediction obtained from various modalities becomes easy compared to feature-level fusion, since the prediction resulting from multiple modalities usually have the same form of data. Another advantage is that each of every modality can utilize its best suitable model to learn its corresponding features. Among the notable decision-level fusion methods in continuous affect recognition is linear regression [125] [165] [126] has been implemented in several AVEC challenge to fuse the estimation from each modality. Other than linear regression, method such as averaging [75], SVR [123], random forests [11] or KF based [9] [153] has been used to combine prediction in decision fusion process.

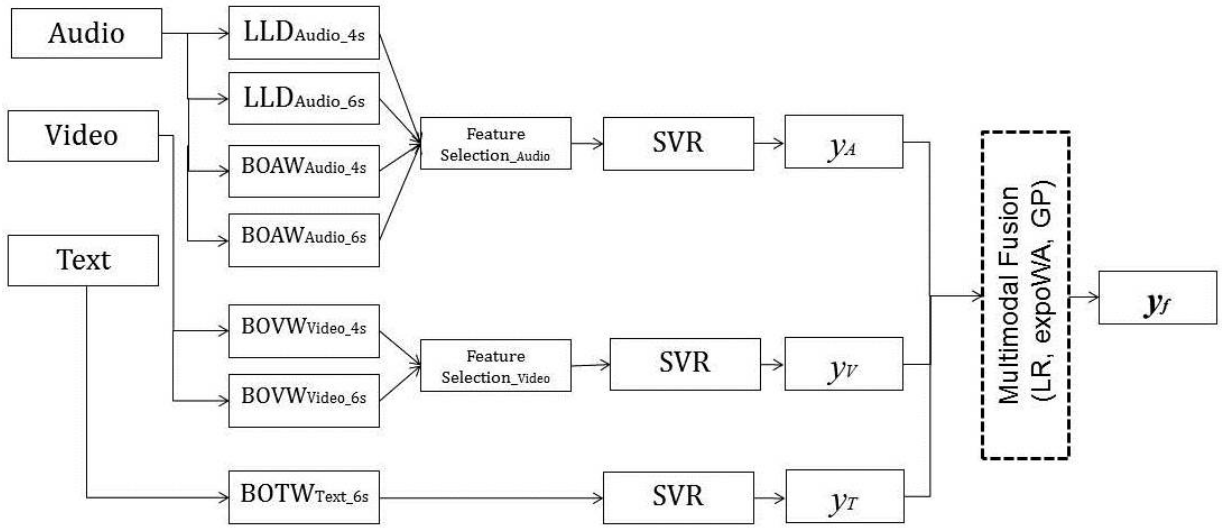


Figure 6.1 :Overview of the proposed system. Fusion of the predictions of the three modalities: audio, video and text.

Output-Associative (OA) fusion techniques, introduced by Nicolaou et al. [107], which take into account the contextual and temporal dependencies that exist within and between predicted affect values when performing fusion, are gaining popularity in continuous emotion prediction. The combination of Output-Associative Bidirectional Long Short Term Neural Networks (OA-BLSTMs) [107], then OA-Relevance Vector Machines (OA-RVMs) [109] [110] is used fairly in order to predict the affect. On AVEC challenge, Huang et al., [69] leverage the fusion methods by combining early, output-associative and late fusion in the final emotion prediction.

However, although such feature or modelling approach was successfully predicting affect in a continuous way, a systematic understanding of what is the relationship between each modality in multimodal fusion is still less frequently explored. Each of the modelling approaches reviewed usually does not give a definite function for the fusion rule. On top of that, it is not always possible to design a model that suits each modality because of the complexity. Therefore, the need to develop a model that can approximate the relationship between the predictions based on a measured set of data without a need of prior knowledge about the modality that produced the experimental data is desired.

6.3 Dataset and Features

The most recent versions of the challenge, AVEC 2017 [126] departed from the previous two years challenge and used SEWA (Sentiment Analysis in the Wild) dataset. It consists of audio-visual recordings of subjects showing spontaneous and natural behaviours. Unlike RECOLA dataset which is being used in previous two years challenge where the recording is strictly in laboratory setting, in SEWA, all recordings were done in-the-wild, e.g., using standard webcams and microphones from the computers in the subjects offices or homes, without any intervention of specific speakers, headphone, or sensors. Audiovisual were recorded during dyadic interactions, 32 pairs in total. The data is provided in three partitions (Training, Development, and Test), where both partners of one video chat appear in the same partition. The data is labelled in three affective dimensions, namely Arousal, Valence, and new emotion dimension, that is Likability, indicating the participants taste for the video/audio. It also was manually annotated by 6 annotators (3 female, 3 male), all were German native speakers, using a joystick. The data is labelled in three affective labels, namely arousal, valence and likability, manually annotated by 6 annotators (3 female, 3 male), all were German native speakers, using a joystick. The dataset is provided together with a set of pre-calculated features which will be incorporated into the model. To avoid repetition, the version of the feature extraction procedures are detailed in [126]. The dataset is provided together with a set of pre-calculated features which will be incorporated into the model, as can be seen in the next subsection.

6.3.1 Audio

For the audio modality, the database provides two sets of audio features, namely Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) LLDs: *functionals* extracted using openSMILE toolkit [34] and *bag-of-audio-words* (BOAW): extracted using openXBOW toolkit [137]. The latter features, BOAW is inspired by text mining research area and commonly used in document classification (bag-of-words). Using bag-of-words principle, LLD on a certain segment is quantised using a codebook of *audio words*, then a histogram of audio words is produced on a corresponding segment. The important parameter that needs to be taken into consideration is the codebook size, i. e., the number of audio words set into the framework. In the baseline features, the codebook size is set to 1000, then standardised to zero mean and unit

variance prior to vector quantisation. Both segment-level eGeMAPS LLDs and BOAW types were computed over segments of 4 and 6 seconds. In total, the audio baseline feature sets with functionals contain 88 features, while the BoAW features contain 1 000 features.

6.3.2 Video

As for video modality, the database provides two sets of video features: facial features and *bag-of-video-words* (BOVW) features. The facial features include face orientation (pitch, yaw, and roll - 3 features), pixel coordinates for 10 eye point (20 features) and pixel coordinates for 49 facial landmarks (98 features). In total, facial has feature value of 121. Then, each of the features is standardised to zero mean and unit variance on the frame level. The latter features, BOVW features were computed on top of standardised facial features with a codebook size of 1 000. The facial features have been extracted from each video frame using the Chehra face tracker [5] while BOVW features are extracted using openXBOW toolkit [137].

6.3.3 Text

Experimenting with text-based features is a quite new approach to continuous emotion recognition. AVEC 2017 firstly introduces text modality features, which are bag-of-words feature representation. In this Chapter, a *bag-of-text-words* (BOTW) feature representation is generated based on the transcription of the speech. By taking into account only the terms with at least two occurrences, the results in a dictionary contained 521 words. Therefore, openXBOW toolbox with a codebook size of 521 is used, resulting 521 features of BOTW.

6.3.4 Regression models

Separate arousal, valence and likability predictions are obtained from individual modalities as described in the last subsection. The regression task is performed using linear SVR provided with the liblinear library [37]. Data from the Development set is used to test the performance as well tune for different parameters after fitting the SVR models on the Training set. Unimodal predictions are first obtained from the five feature sets provided in SEWA dataset (LLD, BOAW, facial landmark video, BOVW and BOTW). Additional experiments have been conducted by scaling and shifting the unimodal estimation according to the training label to correct the bias and scaling issues. These unimodal estimations are used as an input

in multimodal estimation in the proposed late fusion approaches using Multi-Gene Genetic Programming (multi gene GP).

6.4 Decision Level Fusion

In the previous section, the individual strength of each feature is examined, but in an isolated way. Therefore, in this section, the individual advantage of each modality is leveraged by combining them in a multimodal fusion manner. The objective of the fusion is not just to enhance the performance, but to examine the possibility to construct prediction equation of each affect. The next subsection investigated in details each of the fusion approaches applied for continuous affect estimation. Each of the initial prediction from audio, video and text is denoted as y_A , y_V , y_T , and become an input in the following subsequent multimodal fusion.

6.4.1 Linear Regression (LR)

LR attempts to model the relationship between two variables by fitting a linear equation to observed data. In the case of continuous affect estimation, regression coefficients γ need to be weighted separately according to the contribution of each modality towards affects. Equation 6.1 is the linear regression formula where γ and ϵ_m are the regression coefficients and bias term computed in development sets, and \mathbf{y}_f is the final fused prediction.

$$\mathbf{y}_f = \gamma_A(y_A) + \gamma_V(y_V) + \gamma_T(y_T) + \epsilon_m \quad (6.1)$$

6.4.2 Exponent Weighted Decision Fusion

Exponent weighted (EW) decision fusion is a type of decision fusion which introduced in [84] for classification of emotion in EMOTIW challenge 2015. In this Chapter, the exponent weighted decision fusion approach is leveraged in regression manner, where its validation accuracy represented by the correlation between development dataset. Suppose an SVR model with the best correlation, C , where C_A is the best correlation for audio, C_V is for the best correlation for video and C_T is the best correlation for text, will provide an initial prediction for each modality. Then, the final ensemble of our initial prediction from each of the features

in the exponent weighted decision fusion become:

$$y_f = (C_A)^q(y_A) + (C_V)^q(y_V) + (C_T)^q(y_T) \quad (6.2)$$

where a decision weight in terms of $(C)^q$ reflects the significance of initial prediction according to each modality and an exponent q is a hyper-parameter tuning. Here, the value of q is found by a simple uniform search: scanned over $[-50:0.1:150]$ then selected to provide the maximum correlation after the fusion.

6.4.3 Genetic Programming (GP)

GP is inspired by biological evolution in nature. To improve their genomes, the evolution begins by iteratively processing randomly generated solutions (individuals). The objective function is the individual fitness. Iteratively, the reproduction generation is constructed by *survival-of-the-fittest* individuals, by employing *crossover* and *mutation*. In brief, *crossover* is the recombination of parent genome to produce child genome while *mutation* is a possible modification that happens to child genome. The iterative process stops when the maximum number of generations is reached, or the best fitness is visited.

GP is recognized with non-linear chromosomes (trees). The trees in GP can be different in depth and width. The breeding process in GP is performed by mutation and crossover process, as shown in Figure 6.2 and Figure 6.3. In mutation process, the randomly chosen sub-tree is

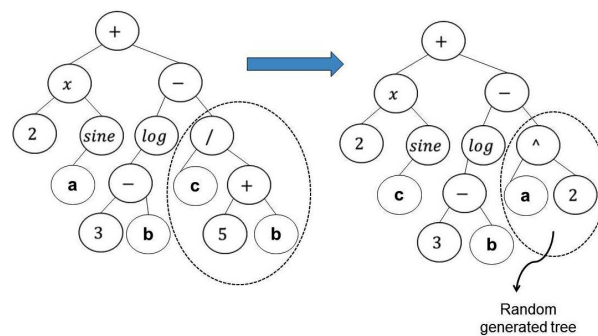


Figure 6.2 :Mutation process in GP. The dashed circle part of tree is replaced with a random generated tree

replaced by randomly a generated tree. In recombination process, the two randomly chosen sub-trees are exchanged among the parents to generate the children.

In the case of continuous affect estimation, the task is to estimate the affect in continuous

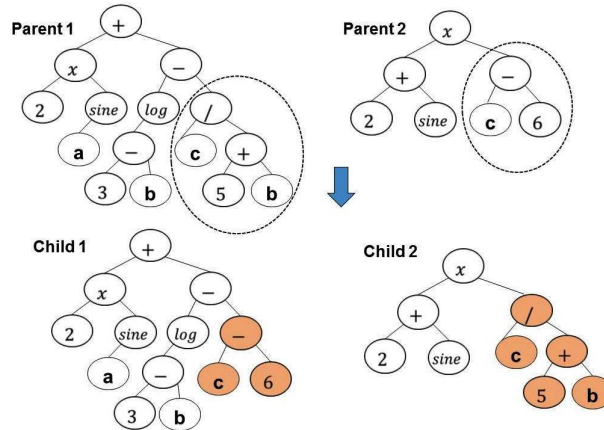


Figure 6.3: Recombination process in GP. Dashed circle parts in parents are exchange.

time scale. Therefore, a model that recombine GP ability and regression method is needed for this task. Multi-gene GP is suitable since the generated model has the advantage of combining regression method and the ability to represent non-linear behaviour [146] of the affect.

Multi gene GP is the results of a combination of GP, multiple gene, and linear regression. In other words, each solution is formed by a linear combination of one or more such functions, called genes. A graphical representation of formulation with three input variable, x_1, x_2, x_3 as shown in Figure 6.4. As can be seen, the structure of this model contains non-linear terms such as sin, exp, cos, and the overall model is a weighted linear combination with respect to each coefficient.

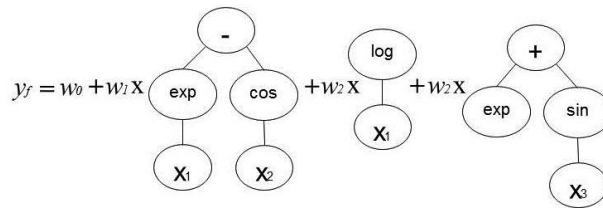


Figure 6.4: Graphical formula with three input variable.

From Figure 6.4, the solution is in the form of:

$$y_f = w_0 + w_1 g_1(x) + w_2 g_2(x) + \dots + w_n g_n(x) \tag{6.3}$$

where n is the number of gene. Each gene is applied to the feature matrix, producing $N \times 1$

vector where:

$$\mathbf{y}_f = [\mathbf{1}g_1g_2\dots g_n] \cdot \mathbf{w} \quad (6.4)$$

with $\mathbf{1}$ being $N \times 1$ vector of ones. The output y of the whole solution is then given by formula:

$$\mathbf{y}_f = \mathbf{G} \cdot \mathbf{w} \quad (6.5)$$

The optimal coefficient vector w^* can then be found using the least-squares estimation with respect to the true target vector y

$$\mathbf{w}^* = (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \mathbf{y} \quad (6.6)$$

6.4.4 Kalman Filter (KF)

In continuous emotion recognition, systems that consider emotion dynamic is less relatively explored. Since emotions are slow changing with time, it is not appropriate to assume that emotional states are static. Previous AVEC 2016 applying dynamic models as fusion approach in their system [9] [153]. One of the most attractive frameworks for exploiting emotion dynamics is KF, which has been used to capture the relationship between position, velocity and acceleration [78]. In Chapter 5, KF is employed to exploit the time series nature of the emotional label data. In this Section, KF is employed to fuse the emotional measure into a single fusion estimation. In this approach, the parameter of transition matrix A , process noise $w(k)$, measurement matrix C , and measurement noise $v(k)$ is estimated. Recall back Equation 2.30

$$x(k+1) = Ax(k) + w(k) \quad (6.7)$$

In the measurement equation z , the measurement matrix C relates the underlying emotional states to the measurements, $v(k)$ is the zero-mean measurement noise term, as in Equation 6.8:

$$z(k) = Cx(k) + \beta + v(k) = \begin{bmatrix} z_{audio} \\ z_{video} \\ z_{text} \end{bmatrix} \quad (6.8)$$

The $v(k)$ and $w(k)$ are important for capturing correlation between multiple states and it is used to exploit the emotion and emotion dynamic. Training a KF mainly concerns estimating the parameter from Equation 6.7 and 6.8, that is A , $w(k)$, C , and $v(k)$.

In this approach, KF implementation is adapted from [9], where the estimation of the state transition matrix and measurement transition matrix was addressed with linear least square. Let x and z be defined as gold standard and the corresponding measurement from the individual modality respectively:

$$\begin{aligned} X_N &= [x_1, \dots, x_{N-1}] \\ Z_N &= [z_1, \dots, z_{N-1}] \end{aligned} \quad (6.9)$$

State transition matrix A and the variance of the process noise Q can be found by:

$$\begin{aligned} A &= (X_{2,N}, X_{1,N-1}^T)(X_{1,N-1}, X_{1,N-1}^T)^{-1} \\ Q &= \text{cov}(w, w) = \text{cov}(X_{2,N} - AX_{1,N-1}) \end{aligned} \quad (6.10)$$

If the following substitution is followed:

$$\begin{aligned} \bar{X}_{1,N} &= \begin{bmatrix} X_{1,N} \\ 1_{1 \times N} \end{bmatrix} \\ \bar{C} &= \begin{bmatrix} C & \beta \end{bmatrix} \end{aligned} \quad (6.11)$$

Equation 6.8 can be rewrites as follows:

$$Z_{1,N} = \bar{C}\bar{X}_{1,N} + v(k) \quad (6.12)$$

Similarly measurement matrix C , the bias term β and variance of the measurement noise R are estimated in the same way by:

$$\begin{aligned} \bar{C} &= (Z_{1,N}, \bar{X}_{1,N}^T)(\bar{X}_{1,N}, \bar{X}_{1,N}^T)^{-1} \\ R &= \text{cov}(v, v) = \text{cov}(Z_{1,N} - C\bar{X}_{1,N} - \beta) \end{aligned} \quad (6.13)$$

After each of the parameters is obtained, the KF performs two operations at each time step: (i) the time update and (ii) the measurement update. Each of the operation is detailed in

Equation 2.37 till Equation 2.41 The fusion approach is done per time step, and at the same time, it models the emotion dynamic of the system.

6.5 Experimental Results

This section empirically evaluates the proposed algorithm in SEWA dataset. SVR modelling is performed for the continuous affect recognition in unimodal setting (audio, video and text modality) according to the features as in Figure 6.1. Once the unimodal estimation of each affect is optimised, multi-gene GP fusion and exponent weighted decision fusion strategies are incorporated to investigate its robustness in the multimodal setting settings. To evaluate the proposed approach, both fusion rule is evaluated by comparing it to the widely-used decision fusion rules in affect; linear regression method.

6.5.1 Experimental Set-ups and Evaluation Metrics

To illustrate the effectiveness of the proposed architecture, the baseline experiment is carried out, where SVR models were individually trained on the modalities of audio, video, text, or the combination, respectively. Specifically, the SVR was implemented in MATLAB with LIBLINEAR toolkit [38], with linear kernel, and trained with the L2-regularised L2-loss dual solver. The complexity (C) of the SVR was optimized by the best performance of the development set among $[2^{-15}, 2^{-14}, \dots, 2]$ for each modality and task. Finally, the early stopping strategy was used when no improvement in the development of the last three iterations. Herein the annotation delay was adapted, to compensate temporal delay associated with the participants and corresponding emotion reported by annotator [97]. In this Chapter, the corresponding delay was estimated in the preliminary experiments using SVR and by maximising the performance on the development partition, while shifting the gold standard annotations back in time, similar in [165]. The delay is identified by shifting the gold standard back in time with respect to all features in the corresponding modality.

The performance of proposed architecture is reported based on C_{corr} [126] metric:

$$C_{corr} = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \quad (6.14)$$

where ρ is the P_{corr} between two time series (e.g: prediction and gold-standard); $\mu_{\hat{y}}$ and μ_y

are the means of each time series; and σ_x^2 and σ_y^2 are the corresponding variance. In contrast to the largely used P_{corr} , C_{corr} also take the bias and variance, e.g., $(\mu_{\hat{y}} - \mu_y)^2$ between the two compared series into account. Hence, the value of C_{corr} is within the range of $[-1, 1]$, where ± 1 represents perfect concordance and discordance while 0 means no concordance between two-time series. Figure 6.5 shows the corresponding affect signal and the corresponding gold standard label signal. From the Figure, P_{corr} of 0.642, as well as C_{corr} of 0.573, has been obtained, with the obtained C_{corr} takes the bias of the mean and variance into account. Therefore, predictions of affect that are well correlated with the gold standard but shifted in value are penalised in proportion to the deviation. The metric of C_{corr} fits better in continuous affect estimation and has been used as an official score for the last three years of AVEC challenge.

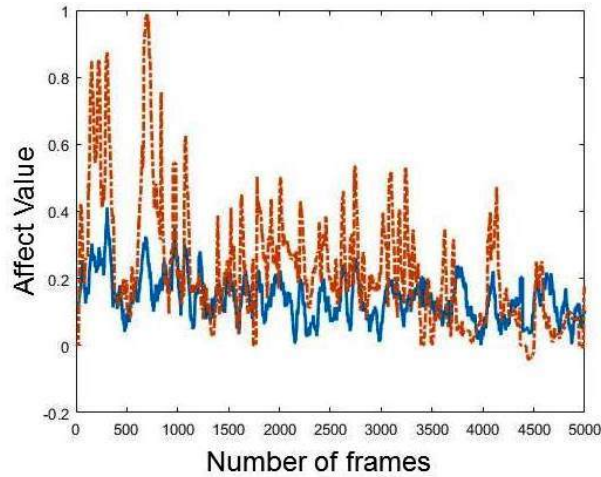


Figure 6.5 :Comparison of C_{corr} and P_{corr} between two time series. Dashed line denoted as prediction label and clear line is gold standard label

6.5.2 Affect Estimation in Unimodal Modality

Table 6.1 displays the results in terms of C_{corr} obtained from unimodal modality of SVR on the development sets of SEWA. On arousal, the best performance is achieved with video modality, more specifically on BOVW features. In valence, the highest results of C_{corr} is taken from audio modality, more specifically on BOAW features. Whereas in likability, the highest is from text modality, more specifically on BOTW features.

6.5.3 Affect Estimation in Mutimodal Modality

The results for multimodal performance of the proposed architecture is shown in Table 6.2. The baseline performance on the development sets as reported in [126] is also shown for comparison. In fusion stage, the best prediction results from unimodal modality are selected, then the model fusion is being tested by dividing the development prediction into two section. The first section is being used to extract the parameter in development sets and the second section is used to test the correctness of the fusion prediction made. A sanity check has been done by reversing the role of both sections.

Three different fusion models were developed for driven piles using LR, EW decision fusion and multi-gene GP algorithm. Statistical performance of the developed models was found in terms of Concordance Correlation Coefficient (C_{corr}) where the value can vary from -1 to 1. In the following subsection, Arousal will be denoted as AR , Valence as VA and Likability as LI respectively.

6.5.3.1 Linear Regression

In the first experiment, fusion model is built by a simple linear regression of the initial estimation obtained on the development partition, using Equation 6.1 in Weka 3.7 [56] on top of MATLAB with the same setting as mentioned above. Equation 6.15 to Equation 6.17 shows the final equation according to each affect, respectively.

$$\mathbf{y}_{f_{AR}} = 0.956y_A + 0.425y_V + 0.404y_T - 0.0915 \quad (6.15)$$

$$\mathbf{y}_{f_{VA}} = 0.299y_A + 0.302y_V + 0.249y_T - 0.0116 \quad (6.16)$$

$$\mathbf{y}_{f_{LI}} = 0.144y_A + 0.202y_V + 0.348y_T - 0.0069 \quad (6.17)$$

6.5.3.2 EW

For the second experiment using EW, the best exponent q is obtained from the first section, then the same q is applied in the second section. Each of the q is scanned in the range of [-50:0.1:150] and validated by using Equation 6.2 thus selected to provide the maximum performance after the fusion. Equation 6.18 to Equation 6.20 shows the final equation according

to each affect, respectively.

$$\begin{aligned} \mathbf{y}_{f_{AR}} = & (0.328)^{6.6}(y_A) + (0.455)^{6.6}(y_V) \\ & + (0.407)^{6.6}(y_T) \end{aligned} \quad (6.18)$$

$$\begin{aligned} \mathbf{y}_{f_{VA}} = & (0.401)^{2.1}(y_A) + (0.389)^{2.1}(y_V) \\ & + (0.386)^{2.1}(y_T) \end{aligned} \quad (6.19)$$

$$\begin{aligned} \mathbf{y}_{f_{LI}} = & (0.175)^{1.5}(y_T) + (0.249)^{1.5}(y_V) \\ & + (0.390)^{1.5}(y_T) \end{aligned} \quad (6.20)$$

6.5.3.3 GP Modelling

Three multi gene GP models are established in this Chapter for predicting the continuous affect for each of affect dimension, respectively. GPTIPS2 developed by Searson et al., [146] was used for model development. The parameters that were set in the multi gene GP algorithms include: a population size of 250, a tournament size of 20, maximum number of genes allowed in an individual 8, function set $\{+, -, \times, /, \sin, \cos, \exp\}$ and terminal sets $\{y_A, y_V, y_T\}$. Equation 6.21 to Equation 6.23 shows the final equation according to each affect, respectively.

$$\begin{aligned} \mathbf{y}_{f_{AR}} = & 5.5e^{-4} \sin(27y_V^3) + 0.31e^{y_T} - 200y_V^3y_T^9 \\ & + 3.4y_A(y_A^3 + y_Vy_A^2 + y_V) \\ & - 0.025e^{(-3y_V)} \sin(9.5y_T) \\ & + 0.1y_A^{1/4} - 0.1y_T^3 - 0.33 \end{aligned} \quad (6.21)$$

$$\begin{aligned} \mathbf{y}_{f_{VA}} = & 0.057 \sin(16y_Vy_T) - 0.32 \sin(y_Ay_Vy_T) \\ & + 0.13 \sin(y_V^2(y_A + 7.8)) \\ & + 0.12y_T^2e^{-y_T}(y_A + 7.8) \\ & + 0.16y_A(e^{-y_T})^{1/2}(y_V + 7.5)^{y_A} + 4.5e^{(-3)} \end{aligned} \quad (6.22)$$

$$\begin{aligned}
\mathbf{y}_{f_{LI}} = & 0.15y_V + 0.15y_T + 0.15 \sin(\sin(y_A)) \\
& -0.18|y_T| + 9.3y_V^4 y_T - 3.4e^3 y_V^7 y_T \\
& +0.36y_A^2 - 6.5y_V^3 + 79y_V^5 \\
& +399y_A y_V^3 y_T + 5.6e^3 y_A y_V^5 y_T - 6.9e^{(-3)}
\end{aligned} \tag{6.23}$$

6.5.3.4 Performance Comparison

Table 6.1: Unimodal performance on the development set

Modality	Features	C_{corr}		
		Arousal	Valence	Likability
Audio	LLD_4s	.380	.338	.062
	LLD_6s	.342	.274	.089
	BOAW_4s	.325	.390	.032
	BOAW_6s	.327	.392	.104
Video	BOVW_4s	.453	.384	.172
	BOVW_6s	.370	.340	.132
Text	BOTW_6s	.364	.382	.317

Closer inspection on Table 6.2 shows that in most cases, decision fusion gives better results than feature fusion method. It is believed that, given the fact that features are extracted in the same manner, there is a tendency of the features have similar or nearly similar distribution, which makes one of them is redundant when performing feature fusion. The finding confirms that in arousal and valence dimension, the multimodal system in Table 6.2 performs better than the best unimodal system in Table 6.1. The new dimension, likability, however, perform better on the unimodal system on text modality. In LR, better performance was achieved for estimating arousal than valence and likability consistent with existing linear modelling frameworks, as shown in Equation 6.15. From this Equation, it shows that audio modality gives the highest weighting factors which contribute significantly to the higher performance in arousal. However, when it comes to valence and likability dimensions, there seems to be a relatively lower performance in estimating those two affect, most likely due to non-linearities in the relationship between the features and those two affect ratings. A further investigation takes place on EW and multi-gene GP approach. The system performance is further improved upon using non-linearity behaviour in estimating valence and likability. By having proper q selection in EW approach gives a significant gain in C_{corr} results for valence and likability, from 0.507 to 0.549 and 0.215 to 0.231 respectively. However, the baseline approach has slightly

Table 6.2: Multimodal performance on 2-fold cross validation

Fusion Type	Fusion Method	C_{corr}		
		Arousal	Valence	Likability
Feature [126]	Concatenate	.525	.507	.235
Decision	LR	.592	.507	.215
	EW	.440	.549	.231
	multi gene GP	.572	.562	.258
	KF	.551	.538	.353

higher performance than the proposed multi gene GP approach, in likability dimension. This may be due to the fact that the SVR models in the first stage have already fit well for the likability with the original feature vector. Notably, the formula produced by multi-gene GP seems to be more compact than yielded by LR and EW, which generates the best results on C_{corr} in valence and likability dimensions, by 0.559 and 0.257 respectively. Looking at the performance increase, it has been concluded that a model with a simple structure is incapable of describing such complex functional mapping in a satisfactory manner. A lower C_{corr} on multi-gene GP and EW instead of LR confirms the assumption that evolution of arousal dimension is linear in time, consistent with the assumption in [153].

An interesting observation was when using KF for fusion approach, the emotion dimension of liking improve significantly from .235 using multi-gene GP to 0.353 using KF approach. This outcome is contrary to that of Dang et al. [22] who suggested that the emotion dimension liking is strongly associated with text modality, but barely correlated with the audio and video modalities. To examine which of the modality proved to be the most trusted when estimating emotion dimension liking, Kalman gain matrix in Equation 2.39 where K_k computed at every time step is accessed. After final computation of Kalman gain weights, it shows that each of the weighted considerably equal, 1.33 for video modality, 1.02 for audio modality and 0.82 for text modality. It shows that, when fusion approach is done per time step, the correlation between estimation and gold-standard can be higher and emotion dynamic can be modelled correctly, at least in liking dimension.

6.6 Comparison on the best performer of the Challenge

The proposed approach is also compared with the the state-of-the-art results [67] [18] and baseline results. It shows competitive performance between state-of-the-art higher results over

the baseline. In arousal and valence dimension, the highest performance is from Huang et al. [67]. In their approach, an additional features, such as IS10, MFCC features and Bottleneck features is adopted in audio modality while LGBP-TOP, HOG and deep visual features for video modality. In liking dimension, Chen et al. [18] achieves significant performance. Since text based features is associated with liking dimension, they fully utilize the word embedding features in German language and translated English languages as feature representation. An additional features most likely contributes to the significant performance in the state-of-the-art results. However, in the proposed approach, instead of feature, modeling approach has been investigated, in order to construct mathematical equation of each affect dimension.

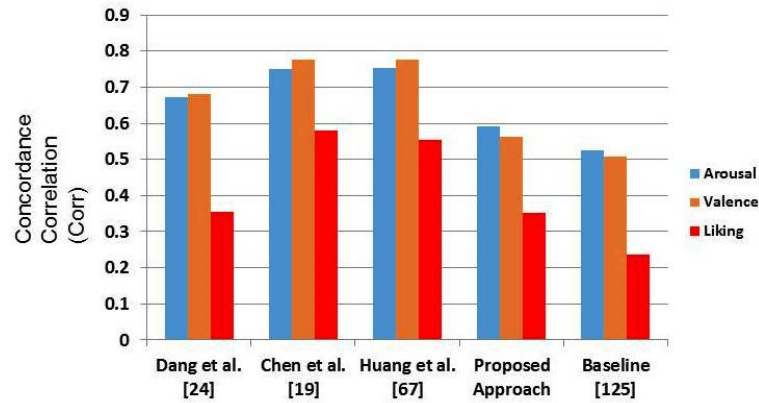


Figure 6.6: Bar chart comparison with state-of-the-art in AVEC 2017 in terms of concordance correlation in multimodal fusion approach.

Table 6.3: Comparison of CCC on fusion of AVEC 2017 state-of-the-art and proposed approach.

All	Method
Dang et al. [22]	OA-RVM framework.
Chen et al. [18]	Multi-Task Learning
Huang et al. [67]	LSTM-RNN
Proposed Approach	LR+EW+MGGP

6.7 Chapter Summary

In this Chapter, analytical investigations have been carried out to assess the performance of Linear Regression, Exponent Weighted Decision Fusion and multi-gene Genetic Programming in multimodal fusion continuous affect estimation in the wild. The results presented have suggested that, in most cases, decision fusion is superior to features fusion in terms of robustness

and accuracy. Multi-gene GP has consistently outperformed LR and EW in the estimation of valence and likability dimension. This work investigates the possibility of employing different modelling approach, including LR, EW, and multi-gene GP, for constructing prediction fusion rules at the decision level in continuous affect estimation in-the-wild. To train and verify these multimodal fusion approaches, a dataset containing text and audiovisual recording is used. LLD, BOAW, BOVW and BOTW features are extracted respectively from audio, video and text modality. Then SVR has been employed to estimate the initial prediction of each affect. In fusion stage, the best initial prediction from unimodal modality is selected, and LR, EW, and multi-gene GP are being employed to construct the prediction rules. Experimental results show that the prediction equation of multi-gene GP shows better modelling outcome than the benchmark results, outperform the baseline approach in all affect dimension. Result comparison with benchmark method such as LR shows that multi-gene GP significantly improve the performance in valence and likability dimension. It confirms an initial assumption that there exists non-linearity behaviour in those two affect dimension. As for arousal dimension, LR performs better than baseline, EW, and multi-gene GP fusion approach. It shows that arousal dimension is generally linear in time.

The fusion using KF approach enabled the multi-modality fusion of emotional state while leveraging the time-varying nature of the emotional states. This approach shows the best performance in liking dimension, suggesting that when fusion approach is done per time step, emotion dynamic can be captured and modelled correctly in the system.

It should be mentioned here that the conclusion might not be completely correct due to the use of the dataset. Although it is a very good dataset, however, the total number of samples is still limited, and the features and first baseline regression method are very basic. In future work, more multimodal datasets and features will be tested to improve the system and verify these assumptions.

Chapter 7

Conclusion and Future Works

7.1 Conclusion

In this thesis, feature representation that is beneficial in various emotion recognition settings is addressed. The proposed system, which consists of feature smoothing, deeply learned features, two-stage regression framework and fusing between multiple modalities is demonstrated. The results achieve competitive performance over the state-of-the-art approach. indicates that the proposed approach agrees with existing domain knowledge. The proposed techniques also show that it can generalise to a different definition of emotion space (Arousal, Valence etc..) and to different input modalities (video/audio/text etc..)

The first task on continuous emotion recognition is on feature smoothing. By adopting Haar Wavelet Transform towards selective features, then employing Partial Least Square Regression as regression approach, it demonstrates that feature smoothing correlated with emotion space and can achieve state-of-the-art performance.

The effectiveness of convolutional neural network features proceeds to video modality and at the same time perform additional analysis on hand crafted features. In-depth analysis of deeply learned features revealed which parts of convolutional and fully connected layer that had the influential effect on the output estimation

One step further has been taken to improve performance on continuous emotion recognition by introducing new modality that is physiological signal modality. At the same time, two-stage regression approach has been introduced, where an initial estimation of affect is feed into the subsequent model. By doing so, the initial estimation is not biased by the high

Table 7.1: Final performance in terms of CCC taken from AVEC 2016 development set.

Modality	Features	Corr	
		Arousal	Valence
Video	LGBPTOP	0.48	0.40
	VGG Face	0.58	0.37
Audio	LLD descriptor	0.40	0.55

variability caused by various sensors or various feature extraction method. Through additional analysis, which features were correlated the most towards gold standard is discovered. Adding physiological features and how it contributed towards overall performance were also examined.

Finally, an additional enhancement is taken by fusing each of the initial estimation from multiple modalities using a linear and non-linear model to produce a final estimation of affect. In this approach, multimodal fusion continuous affect estimation *in-the-wild* is investigated. This finding, while preliminary, suggests that potential non-linearities occur between affect dimensions.

This subsection is dedicated by experimenting on one common dataset for all the methods proposed in Chapter 3, 4, 5 and 6 and comparing the results. The chosen dataset is AVEC 2016. In this dataset, the modality chosen would be video and audio modality. We will investigate the effectiveness of wavelet transform as proposed in Chapter 3 for handcrafted features and also deep learning features as proposed in Chapter 4. Then, by extending the architecture by capturing the temporal analysis at decision level by proposing two stage regression framework, as indicate in Chapter 5. Finally, we leverage continuous emotion recognition by investigating mathematical modeling for each emotion dimension as proposed in Chapter 6.

At the initial stage, video and audio is dealt independently and preprocessed by using Haar Wavelet Transform. At Stage 2, each feature type is feed consecutively into SVR. The initial estimation in Stage 2 is feed to Stage 3, where two-stage regression approach takes place. Then, at Stage 4, fusion on respective features occur, to produce the final estimation of arousal and valence.

The results of experiments across all proposed approach are shown in Table 7.1. In arousal dimension, the results shows VGGFace gives high results. It is clear that, from Chapter 4, the employment of VGGFace trained for face recognition is more efficient than exhaustive approach of LGBPTOP features. However, in valence dimension, audio descriptor likely to capture temporal dependencies brought by two stage regression approach, as explained in

Table 7.2: Comparison of CCC based on mathematical modeling in Chapter 6. Noted that the proposed approach are obtained on 2 fold cross validation.

	Corr_A	Corr_V
VGGFace + LLD Descriptor	0.60	0.45

Chapter 5.

Table 7.2 shows that, mathematical modeling proposed in Chapter 6 gives better results than the best unimodal system in Table 7.1. VGGFace features give higher weighting factor which contribute significantly to the higher performance in arousal. However, when it comes to valence, relatively lower performance given when compared to the best performance in unimodal system. It is most likely due to potential non-linearities occur between features and valence dimension itself. It is suggested that more constructive mathematical equation is needed in order to capture the non-linearities effect occur between features and affect ratings.

7.2 Future Works

While the results are encouraging, noted that this thesis is only a small step towards building an automatic emotionally continuous intelligent systems. The performance on affect recognition can still be improved from two aspects: the variety of the dataset as the ground truth values and the feature extraction method as well as the modelling approaches. In the thesis, only two type of datasets is selected: naturalistic setting and *in-the-wild* setting. However, both datasets are captured in controlled conditions and under very specific scenarios. Obtaining and annotating more data in a wide variety of scenarios can validate the performance and provide a better way to design continuous emotion recognition system. On the other hand, a number of feature extraction method and regression approach is also investigated, however, but the space for finding the optimal approach is huge. For example, in addition to SVR, LSTM-RNN is also being tested in the thesis. By exploring CNN feature descriptor with LSTM-RNN as modelling approach to build the model maybe can enhance the overall performance. Also, instead of having facial expression images as a representation of video modality, other visual cues could be leverage such as hand gesture, body expression and so on to improve the robustness on the recognition performance. Improving the performance on continuous affect recognition can provide a fundamental step towards building an emotionally intelligent systems.

7.2.1 Wide Variety of Datasets

While the results on SEMAINE, RECOLA and SEWA datasets are encouraging, they are by no means sufficient. It is because each of the respective datasets is far too small to suggest that the affect recognition systems can generalise to unconstrained emotions in the wild. To properly test the boundaries of the proposed system, they need to be trained on datasets that are large scales and accurately mimic real-life situations.

7.2.2 Detecting Mental Health Disorders

Other than depression, another possible direction would be to study the detection of anxiety or autism, to help clinicians in medical diagnosis. Currently, the manner in which a patient is diagnosed is based on individual assessment. As the number of patients increases so does the need for accurate diagnosis. Standardisation of the diagnosis task can greatly help the doctors or counsellor to give personalised care to the risk patients.

7.2.3 Reinforcement Learning

The works in Chapter 5 focus solely on two-stage regression approach. It can be extended by modelling two-stage regression by applying reinforcement learning. It means that the computer is an agent that uses emotion recognition to interact with its environment (the user). The environment then outputs its state (detected emotion), and the agent (the computer) must perform an action that maximizes its expected reward. The reinforcement learning can act as a tandem where the estimation of affect is produced continuously when the sequence of expressions is received. Research in this direction can make the system to recognize person emotion then the system can respond it with an appropriate action in emotionally intelligent systems.

Bibliography

- [1] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, pages 579–586, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [2] T. R. Almaev and M. F. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 356–361, 2013.
- [3] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.
- [4] M. Amirian, M. Kächele, P. Thiam, V. Kessler, and F. Schwenker. Continuous multimodal human affect estimation using echo state networks. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16*, pages 67–74, New York, NY, USA, 2016. ACM.
- [5] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, 2014.
- [6] T. Baltrusaitis, N. Banda, and P. Robinson. Dimensional Affect Recognition using Continuous Conditional Random Fields. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

- [7] R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636, 1996.
- [8] I. Bloch. Information combination operators for data fusion: a comparative review with classification. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 26(1):52–67, Jan 1996.
- [9] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16*, pages 97–104, New York, NY, USA, 2016. ACM.
- [10] R. A. Calvo and S. D’Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.
- [11] P. Cardinal, N. Dehak, A. L. Koerich, J. Alam, and P. Boucher. Ets system for av+ec 2015 challenge. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15*, pages 17–23, New York, NY, USA, 2015. ACM.
- [12] G. Castellano, S. D. Villalba, and A. Camurri. Recognising Human Emotions from Body Movement and Gesture Dynamics. In *Affective Computing and Intelligent Interaction*, pages 71–82. 2007.
- [13] M. Caudill. Neural networks primer, part i. *AI Expert*, 2(12):46–52, Dec. 1987.
- [14] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Multi-scale temporal modeling for dimensional emotion recognition in video. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14*, pages 11–18, New York, NY, USA, 2014. ACM.
- [15] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition. *AVEC workshop*, pages 65–72, 2015.
- [16] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Long short term memory recurrent neural network based encoding method for emotion recognition in video. In *2016 IEEE*

- International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2752–2756, March 2016.
- [17] S. Chen and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15*, pages 49–56, New York, NY, USA, 2015. ACM.
- [18] S. Chen, Q. Jin, J. Zhao, and S. Wang. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17*, pages 19–26, New York, NY, USA, 2017. ACM.
- [19] F. Chollet. keras. <https://github.com/fchollet/keras>, 2015.
- [20] A. Clerico, R. Gupta, and T. H. Falk. Mutual information between inter-hemispheric eeg spectro-temporal patterns: A new feature for automated affect recognition. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 914–917, April 2015.
- [21] 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.
- [22] 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.
- [23] C. Darwin and P. Ekman. The expression of the emotions in man and animals (3rd ed.). *The expression of the emotions in man and animals 3rd ed*, 1872.
- [24] R. Davidson, K. Scherer, and H. Goldsmith. *Handbook of Affective Sciences*. Series in affective science. Oxford University Press, 2002.
- [25] M. E. Dawson, A. M. Schell, and D. L. Filion. *The Electrodermal System*, 2007.

- [26] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using phog and lpq features. In *Face and Gesture 2011*, pages 878–883, March 2011.
- [27] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support Vector Regression Machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, 1997.
- [28] P. Ekman. Strong evidence for universals in facial expressions: A reply to Russell’s mistaken critique. *Psychological Bulletin*, 115(2):268–287, 1994.
- [29] P. Ekman and W. Friesen. *Pictures of Facial Affect*. Consulting psychologists Press, 1976.
- [30] P. Ekman and W. Friesen. *Facial Action Coding System: Investigator’s Guide*. Number v. 1 in Facial Action Coding System: Investigator’s Guide. Consulting Psychologists Press, 1978.
- [31] P. Ekman, W. Friesen, P. Ellsworth, A. Goldstein, and L. Krasner. *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. Pergamon general psychology series. Elsevier Science, 2013.
- [32] P. Ekman, W. V. Friesen, M. O’Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, K. Scherer, M. Tomita, and A. Tzavaras. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717, 1987.
- [33] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.
- [34] F. Eyben, F. Weninger, F. Groß, B. Schuller, F. Gross, and B. Schuller. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. *Proceedings of the 21st ACM International Conference on Multimedia (MM 2013)*, (May):835–838, 2013.

- [35] F. Eyben, F. Weninger, S. Squartini, and B. Schuller. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 483–487, May 2013.
- [36] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1459–1462, New York, NY, USA, 2010. ACM.
- [37] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
- [38] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.
- [39] Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 445–450, New York, NY, USA, 2016. ACM.
- [40] D. Garcia-Gasulla, J. Béjar, U. Cortés, E. Ayguadé, and J. Labarta. Extracting visual patterns from deep learning representations. *CoRR*, abs/1507.08818, 2015.
- [41] Y. F. A. Gaus, H. Meng, A. Jan, F. Zhang, and S. Turabzadeh. Automatic affective dimension recognition from naturalistic facial expressions based on wavelet filtering and pls regression. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 05, pages 1–6, May 2015.
- [42] F. A. Gers, J. A. Schmidhuber, and F. A. Cummins. Learning to forget: Continual prediction with lstm. *Neural Comput.*, 12(10):2451–2471, Oct. 2000.
- [43] A. Graves. *Supervised Sequence Labeling with Recurrent Neural Networks*, volume 12. 2013.
- [44] H. Gunes, M. Nicolaou, and M. Pantic. Continuous Analysis of Affect from Voice and Face. In *Computer Analysis of Human Behavior SE - 10*, pages 255–291. 2011.

- [45] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *Int. J. Synth. Emot.*, 1(1):68–99, Jan. 2010.
- [46] H. Gunes and M. Pantic. Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, 2010.
- [47] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(1):64–84, 2009.
- [48] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image Vision Computing*, 31(2):120–136, Feb. 2013.
- [49] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Face and Gesture 2011*, pages 827–834, March 2011.
- [50] H. Gunes, C. Shan, S. Chen, and Y. Tian. Bodily Expression for Automatic Affect Recognition. In *Emotion Recognition: A Pattern Analysis Approach*, pages 343–377. 2015.
- [51] 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. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14*, pages 33–40, New York, NY, USA, 2014. ACM.
- [52] R. Gupta, K. ur Rehman Laghari, and T. H. Falk. Relevance vector classifier decision fusion and eeg graph-theoretic features for automatic affective state characterization. *Neurocomputing*, 174(Part B):875 – 884, 2016.
- [53] L. A. Gutnik, A. F. Hakimzada, N. A. Yoskowitz, and V. L. Patel. The role of emotion in decision-making: A cognitive neuroeconomic approach towards understanding sexual risk behavior. *Journal of Biomedical Informatics*, 39(6):720–736, 2006.

- [54] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi. Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition. *IEEE Transactions on Information Forensics and Security*, 11(9):1984–1996, Sept 2016.
- [55] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi. Fully automatic face normalization and single sample face recognition in unconstrained environments. *Expert Syst. Appl.*, 47(C):23–34, Apr. 2016.
- [56] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.
- [57] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [58] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15*, pages 73–80, New York, NY, USA, 2015. ACM.
- [59] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80, 2015.
- [60] J. A. Healey and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, June 2005.
- [61] C. Hjortsjö. *Man's Face and Mimic Language*. Studen litteratur, 1969.
- [62] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, 92(1):84–89, jun 2004.

- [63] S. Hochreiter and J. Uergen Schmidhuber. LONG SHORT-TERM MEMORY. *Neural Computation*, 9(8):1735–1780, 1997.
- [64] R. Horlings, D. Datcu, and L. J. M. Rothkrantz. Emotion recognition using brain activity. In *Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, CompSysTech '08, pages 6:II.1–6:1, New York, NY, USA, 2008. ACM.
- [65] C. Hu, Y. Chang, R. Feris, and M. Turk. Manifold based analysis of facial expression. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 81–81, June 2004.
- [66] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *University of Massachusetts Amherst Technical Report*, 1:07–49, 2007.
- [67] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang, and J. Yi. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC '17*, pages 11–18, New York, NY, USA, 2017. ACM.
- [68] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15*, pages 41–48, New York, NY, USA, 2015. ACM.
- [69] 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.
- [70] V. Imbrasaitė, T. Baltrušaitis, and P. Robinson. CCFNF for continuous emotion tracking in music: Comparison with CCRF and relative feature representation. *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2014.

- [71] R. Jenke, A. Peer, and M. Buss. A comparison of evaluation measures for emotion recognition in dimensional space. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 822–826, Sept 2013.
- [72] C. Jones and J. Sutherland. *Acoustic Emotion Recognition for Affective Computer Gaming*, pages 209–219. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [73] E. Jovanov, A. O. Lords, D. Raskovic, P. G. Cox, R. Adhami, and F. Andrasik. Stress monitoring using a distributed wireless intelligent sensor system. *IEEE Engineering in Medicine and Biology Magazine*, 22(3):49–55, May 2003.
- [74] M. Kächele, M. Schels, and F. Schwenker. Inferring Depression and Affect from Application Dependent Meta Knowledge. *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*, pages 41–48, 2014.
- [75] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels. Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15*, pages 9–16, New York, NY, USA, 2015. ACM.
- [76] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R. Chandias Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio. EmoNets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111, 2016.
- [77] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülcehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI '13*, pages 543–550, New York, NY, USA, 2013. ACM.
- [78] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35, 1960.

- [79] I. Kanluan, M. Grimm, and K. Kroschel. Audio-visual emotion recognition using an emotion space concept. In *European Signal Processing Conference*, 2008.
- [80] H. Kaya, F. Çilli, and A. A. Salah. Ensemble cca for continuous emotion prediction. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14*, pages 19–26, New York, NY, USA, 2014. ACM.
- [81] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. *CoRR*, abs/1512.00596, 2015.
- [82] Z. Khalili and M. H. Moradi. Emotion detection using brain and peripheral signals. In *2008 Cairo International Biomedical Engineering Conference*, pages 1–4, Dec 2008.
- [83] P. Khorrami, T. L. Paine, and T. S. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, ICCVW '15, pages 19–27, Washington, DC, USA, 2015. IEEE Computer Society.
- [84] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15*, pages 427–434, New York, NY, USA, 2015. ACM.
- [85] S. Koelstra, C. M??hl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis; Using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.
- [86] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.
- [87] S. D. Kreibig. Autonomic nervous system activity in emotion: A review, 2010.
- [88] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

- [89] N. Kumar, R. Gupta, T. Guha, C. Vaz, M. V. Segbroeck, J. Kim, and S. Narayanan. Affective feature design and predicting continuous affective dimensions from music. In *Mediaeval Workshop, Barcelona, Spain*, 2014.
- [90] O.-w. Kwon, K. Chan, J. Hao, and T.-w. Lee. Emotion Recognition by Speech Signals. *Eighth European Conference on Speech Communication and Technology*, pages 125–128, 2003.
- [91] K. Kyle. Suite of automatic linguistic analysis tools (salat), Dec. 2017.
- [92] R. Lane and L. Nadel. *Cognitive Neuroscience of Emotion*. Series in affective science. Oxford University Press, 2002.
- [93] P. A. Lewis, H. D. Critchley, P. Rotshtein, and R. J. Dolan. Neural correlates of processing valence and arousal in affective words. *Cerebral Cortex*, 17(3):742–748, 2007.
- [94] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2, ICCV '99*, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.
- [95] Q. Lu, J. Ren, and Z. Wang. Using genetic programming with prior formula knowledge to solve symbolic regression problem. *Intell. Neuroscience*, 2016:1:1–1:1, Jan. 2016.
- [96] A. Mahdhaoui and M. Chetouani. Emotional speech classification based on multi view characterization. In *Proceedings - International Conference on Pattern Recognition*, pages 4488–4491, 2010.
- [97] S. Mariooryad, S. Member, C. Busso, and S. Member. Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108, 2015.
- [98] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.*, 3(1):5–17, Jan. 2012.
- [99] H. Meng and N. Bianchi-Berthouze. Affective state level recognition in naturalistic facial and vocal expressions. *IEEE Transactions on Cybernetics*, 44(3):315–318, 2014.

- [100] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas. Time-Delay Neural Network for Continuous Emotional Dimension Prediction From Facial Expression Sequences. *IEEE Transactions on Cybernetics*, 46(4):916–929, 2016.
- [101] H. Meng, D. Huang, H. Wang, H. Yang, M. Al-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 International Workshop on Audio/Visual Emotion Challenge, AVEC '13*, pages 21–30, New York, NY, USA, 2013. ACM.
- [102] A. Metallinou, A. Katsamanis, M. Wollmer, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification (Extended abstract). In *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pages 463–469, 2015.
- [103] A. Metallinou, Z. Yang, C. chun Lee, C. Busso, S. Carnicke, and S. Narayanan. The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. *Language Resources and Evaluation*, 50(3):497–521, 2016.
- [104] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, March 2016.
- [105] A. Mollahosseini, B. Hassani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor. Facial expression recognition from world wild web. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 00, pages 1509–1516, June 2016.
- [106] A. Nakasone, H. Prendinger, and M. Ishizuka. Emotion recognition from electromyography and skin conductance. In *The Fifth International Workshop on Biosignal Interpretation (BSI-05)*, pages 219–222, 2005.
- [107] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

- [108] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, April 2011.
- [109] M. A. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. In *Face and Gesture 2011*, pages 16–23, March 2011.
- [110] M. A. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. *Image Vision Computing.*, 30(3):186–196, Mar. 2012.
- [111] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 501–508, New York, NY, USA, 2012. ACM.
- [112] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.
- [113] V. Ojansivu and J. Heikkilä. *Blur Insensitive Texture Classification Using Local Phase Quantization*, pages 236–243. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [114] E. Okafor, P. Pawara, F. Karaaba, O. Surinta, V. Codreanu, L. Schomaker, and M. Wiering. Comparative study between deep learning and bag of visual words for wild-animal recognition. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8, Dec 2016.
- [115] H. A. Osman, H. Dong, and A. E. Saddik. Ubiquitous biofeedback serious game for stress management. *IEEE Access*, 4:1274–1286, 2016.
- [116] M. Pantic and L. J. M. Rothkrantz. Expert system for automatic analysis of facial expressions, 2000.
- [117] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. In *Proceedings of the IEEE*, volume 91, pages 1370–1390, 2003.

- [118] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.
- [119] S. Petridis and M. Pantic. Prediction-Based Audiovisual Fusion for Classification of Non-Linguistic Vocalisations. *IEEE Transactions on Affective Computing*, 7(1):45–58, 2016.
- [120] R. W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.
- [121] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel. Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16*, pages 75–82, New York, NY, USA, 2016. ACM.
- [122] R. Prakash Gadhe, R. R. Deshmukh, and V. B. Waghmare. KNN based emotion recognition system for isolated Marathi speech. *International Journal of Computer Science Engineering (IJCSE)*, 2015.
- [123] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J. P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30, 2015.
- [124] F. Ringeval, E. Marchi, C. Grossard, J. Xavier, M. Chetouani, D. Cohen, and B. Schuller. Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 08-12-September-2016, pages 1210–1214, 2016.
- [125] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic. AVEC 2015 – The 5th International Audio/Visual Emotion Challenge and Workshop. *Proceedings of the 23rd ACM International Conference on Multimedia, MM 2015*, pages 1335–1336, 2015.
- [126] 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.

- [127] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. Av+ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15*, pages 3–8, New York, NY, USA, 2015. ACM.
- [128] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, April 2013.
- [129] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.
- [130] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, June 2015.
- [131] P. Sarkheil, R. Goebe, F. Schneider, and K. Mathiak. Emotion unfolded by motion: A role for parietal lobe in decoding dynamic facial expressions. *Social Cognitive and Affective Neuroscience*, 8(8):950–957, 2013.
- [132] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12*, pages 485–492, New York, NY, USA, 2012. ACM.
- [133] K. Scherer, A. Schorr, and T. Johnstone. *Appraisal Processes in Emotion: Theory, Methods, Research*. Series in Affective Science. Oxford University Press, 2001.
- [134] K. R. Scherer. *Psychological models of emotion*, 2000.
- [135] K. R. Scherer, T. Banziger, and E. B. Roesch. *Blueprint for affective computing : a sourcebook*. 2010.

- [136] E. M. Schmidt and Y. E. Kim. Prediction of time-varying musical mood distributions using Kalman filtering. In *Proceedings - 9th International Conference on Machine Learning and Applications, ICMLA 2010*, pages 655–660, 2010.
- [137] M. Schmitt and y. v. p. Björn W. Schuller, journal=Journal of Machine Learning Research. openxbow - introducing the passau open-source crossmodal bag-of-words toolkit.
- [138] B. Schuller. Recognizing affect from linguistic information in 3d continuous space. *IEEE Transactions on Affective Computing*, 2(4):192–205, Oct 2011.
- [139] B. Schuller, D. Arsic, F. Wallhoff, G. Rigoll, et al. Emotion recognition in the noise applying large acoustic feature sets. *Speech Prosody, Dresden*, pages 276–289, 2006.
- [140] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2, pages 881–884, 2007.
- [141] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll. Speaker independent speech emotion recognition by ensemble classification. In *IEEE International Conference on Multimedia and Expo, ICME 2005*, volume 2005, pages 864–867, 2005.
- [142] B. Schuller, G. Rigoll, and M. Lang. Hidden Markov model-based speech emotion recognition. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 2:401–404, 2003.
- [143] B. Schuller, M. Valstar, R. Cowie, and M. Pantic. AVEC 2012: The continuous audio/visual emotion challenge - an introduction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12*, pages 361–362, New York, NY, USA, 2012. ACM.
- [144] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011 - the first international audio/visual emotion challenge. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction - Volume Part II, ACII'11*, pages 415–424, Berlin, Heidelberg, 2011. Springer-Verlag.

- [145] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. AVEC 2012: the continuous audio/visual emotion challenge. *Proc. 14th Int'l Conf. Multimodal Interaction Workshops*, pages 449–456, 2012.
- [146] D. Searson, D. Leahy, and M. Willis. GPTIPS: An open source genetic programming toolbox for multigene symbolic regression. *Proceedings of the International of the MultiConference of Engineers and Computer Scientists*, I:17–20, 2010.
- [147] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vision Comput.*, 27(6):803–816, May 2009.
- [148] L. C. D. E. Silva and I. T. Miyasato. Facial Emotion Recognition Using Multi-modal Information. In *International Conference on Information, Communications and Signal Processing ICICS*, number September, pages 9–12, 1997.
- [149] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [150] A. Sinha, H. Chen, D. G. Danu, T. Kirubarajan, and M. Farooq. Estimation and decision fusion: A survey. In *2006 IEEE International Conference on Engineering of Intelligent Systems*, pages 1–6, 2006.
- [151] M. Soleymani, S. Asghari Esfeden, Y. Fu, and M. Pantic. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 3045(c):1–1, 2015.
- [152] M. Soleymani, S. Asghari-esfeden, M. Pantic, and Y. Fu. Continuous emotion detection using EEG signals and facial expressions. *IEEE Conference on Multimedia and Expo (ICME)*, 231287(231287):3–8, 2013.
- [153] K. Somandepalli, R. Gupta, M. Nasir, B. M. Booth, S. Lee, and S. S. Narayanan. Online affect tracking with multimodal kalman filters. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16*, pages 59–66, New York, NY, USA, 2016. ACM.

- [154] B. Sun, S. Cao, L. Li, J. He, and L. Yu. Exploring multimodal visual features for continuous affect recognition. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16*, pages 83–88, New York, NY, USA, 2016. ACM.
- [155] Y. Sun, N. Sebe, M. S. Lew, and T. Gevers. Authentic emotion detection in real-time video. *Human Computer Interaction, European Conference on Computer Vision*, pages 94–104, 2004.
- [156] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, June 2014.
- [157] Y. Tang. Deep learning using support vector machines. *CoRR*, abs/1306.0239, 2013.
- [158] J. Tao and T. Tan. Affective computing: A review. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3784 LNCS, pages 981–995, 2005.
- [159] L. Tian, J. D. Moore, and C. Lai. Emotion recognition in spontaneous and acted dialogues. In *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pages 698–704, 2015.
- [160] S. Tomkins. *Affect Imagery Consciousness: Volume I: The Positive Affects*. Springer Series. Springer Publishing Company, 1962.
- [161] S. Tomkins. *Affect Imagery Consciousness: Volume II: The Negative Affects*. Springer Series. Springer Publishing Company, 1963.
- [162] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, June 2014.
- [163] K. P. Truong, D. A. Van Leeuwen, M. A. Neerinx, and F. M. De Jong. Arousal and valence prediction in spontaneous emotional speech: Felt versus perceived emotion. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2027–2030, 2009.

- [164] R. Valitutti. Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, 2004.
- [165] 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.
- [166] 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 *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, AVEC '14*, pages 3–10, New York, NY, USA, 2014. ACM.
- [167] 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 *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC '13*, pages 3–10, New York, NY, USA, 2013. ACM.
- [168] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th International Conference on Multimodal Interfaces, ICMI '07*, pages 38–45, New York, NY, USA, 2007. ACM.
- [169] E. L. van den Broek, F. van der Sluis, and T. Dijkstra. Telling the story and re-living the past: How speech analysis can reveal emotions in post-traumatic stress disorder (ptsd) patients. In J. Westerink, M. Krans, and M. Ouwkerk, editors, *Sensing Emotions: The impact of context on experience measurements*, volume 12 of *Philips Research Book Series*, pages 153–180. Springer Science+Business Media B.V., Dordrecht, The Netherlands, August 2011.
- [170] L. van der Maaten. Audio-visual emotion challenge 2012: a simple approach. *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 473–476, 2012.

- [171] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
- [172] A. Vedaldi and K. Lenc. MatConvNet - Convolutional Neural Networks for MATLAB. *Arxiv*, pages 1–15, 2014.
- [173] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, Mar 1989.
- [174] R. Weber, V. Barrielle, C. Soladié, and R. Séguier. High-level geometry-based features of video modality for emotion prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16*, pages 51–58, New York, NY, USA, 2016. ACM.
- [175] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller. Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016*, New York City, NY, July 2016. IJCAI/AAAI. 7 pages.
- [176] F. Weninger, M. Wöllmer, and B. Schuller. Emotion Recognition in Naturalistic Speech and Language-A Survey. In *Emotion Recognition: A Pattern Analysis Approach*, pages 237–267. 2015.
- [177] Wikipedia. Human computer interaction, 2017. [Online; accessed 11-December-2017].
- [178] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 597–600, 2008.
- [179] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, feb 2013.
- [180] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face

- alignment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, June 2013.
- [181] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, pages 3320–3328, Cambridge, MA, USA, 2014. MIT Press.
- [182] J. Yosinski, J. Clune, A. M. Nguyen, T. J. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015.
- [183] M. You, C. Chen, J. Bu, J. Liu, and J. Tao. Emotion recognition from noisy speech. In *2006 IEEE International Conference on Multimedia and Expo*, pages 1653–1656, July 2006.
- [184] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, Jan 2009.
- [185] Z. Zhang, J. Pinto, C. Plahl, B. W. Schuller, and D. Willett. Channel mapping using bidirectional long short-term memory for dereverberation in hands-free voice controlled devices. *IEEE Trans. Consumer Electronics*, 60(3):525–533, 2014.
- [186] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller. Enhanced semi-supervised learning for multimodal emotion recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2016-May, pages 5185–5189, 2016.