UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Automated Deep Audiovisual Emotional Behaviour Analysis in the Wild

Jing Han

Angaben zur Veröffentlichung / Publication details:

Han, Jing. 2020. "Automated Deep Audiovisual Emotional Behaviour Analysis in the Wild." Augsburg: Universität Augsburg.

Nutzungsbedingungen / Terms of use:

CC BY-SA 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY-SA 4.0: Creative Commons: Namensnennung - Weitergabe unter gleichen Bedingungen
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by-sa/4.0/deed.de

Automated Deep Audiovisual Emotional Behaviour Analysis in the Wild

Inaugural-Dissertation
for the degree of
Doctor of Natural Sciences (Dr. rer. nat.)
in the
Faculty of Applied Computer Science
University of Augsburg

by

Jing Han

2020

Evaluation committee:	Prof. Dr.-Ing. habil. Björn Schuller Prof. Dr. Elisabeth André Prof. Dr. Ngoc Thang Vu
Date of Defence:	June 25th, 2020

Abstract

Automatically estimating a user's emotional behaviour via speech contents and facial expressions plays a crucial role in the development of intelligent human-computer interaction systems. Thus, considerable efforts have been made to develop emotion-aware systems to be useful in real-life applications. However, several main challenges still remain and need to be tackled. For example, hand-engineered features are not effective or discriminative to represent the emotional contents from the raw audio or video inputs. Likewise, conventional deep learning structures and models have not taken the characteristics of emotions into account and thus need to be adjusted. Furthermore, continual emotion perception and empathic behaviour analysis have not been investigated so far, which however is highly related when implemented into real-life products.

To deal with these challenges, this thesis proposes and presents a set of representation learning approaches and emotion modelling frameworks. In particular, for the representation learning task, with the advent of deep learning techniques, data-driven representation learning approaches are introduced, aiming to learn discriminative, context-aware, largely modal-invariant features to represent the emotional states. Further, various novel deep network structures are conceived and investigated to enhance present emotion recognition systems. More precisely, this is achieved either by incorporating the strengths of different sub-networks, or by exploiting the disagreement level of the annotations as difficulty indicators. Alternatively, models can be trained jointly with heterogeneous data, or grasp additional knowledge through adversarial learning. Extensive experiments conducted with various spontaneous emotional datasets demonstrate that these introduced methods are superior to the current state-of-the-art methods in both dimensional emotion regression and categorical emotion classification tasks. Moreover, this thesis sheds light on how to deploy deep learning techniques to effectively address lifelong emotional recognition and automatic empathy detection issues.

Acknowledgement

First of all, I would like to thank my supervisor Prof. Björn Schuller for providing me the opportunity to study in University of Augsburg, for inspiring me to carry out research into the field of affective computing and deep learning, for his kind help during the past five years of my studies, for his guidance and the numerous discussions that we had, and for his supervision of this thesis.

Moreover, I want to take this opportunity to express my gratitude to all the people for their help they gave me to accomplish this thesis. Particularly, for discussing and exchanging ideas and thoughts, I sincerely thank my colleagues and collaborators Fabien Ringeval, Gil Keren, Hesam Sagha, Jean Kossaifi, Jie Shen, Jun Deng, Maximilian Schmitt, Nicholas Cummins, Vedhas Pandit, Yanan Guo, Zhao Ren, Zixing Zhang, and many others. Their supports during all this time are greatly appreciated. I also extend my thanks to all the people in the chair for making the time of studying and working here interesting and pleasant.

Most of all, I want to thank especially my parents for their constant support and understanding. Last but not least, I am very grateful for my husband Zixing for his guidance and advice, for his love and support, and for his encouragement and companionship throughout my PhD study.

This research was supported by the EU's Horizon 2020 Programme through the Innovation Action No. 645094 (SEWA), and by the EU's Horizon 2020/EFPIA Innovative Medicines Initiative under grant agreement No. 115902 (RADAR-CNS).

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Aims of the Thesis . . . 5
- 1.3 Structure of the Thesis . . . 6

2 Theoretical Background **9**
- 2.1 Emotion Recognition in General . . . 9
- 2.2 Monomodal Emotion Recognition . . . 11
- 2.3 Audiovisual Emotion Recognition . . . 13
- 2.4 Deep Learning for Feature Extraction . . . 14
- 2.5 Deep Learning for Emotion Prediction . . . 17

3 Contributed Methodology **19**
- 3.1 From Hand-crafted to Data-driven Representations . . . 19
  - 3.1.1 Deep Crossmodal Latent Representations . . . 20
    - 3.1.1.1 EmoBed Framework: System Overview . . . 21
    - 3.1.1.2 Joint Training with Audiovisual Data . . . 22
    - 3.1.1.3 Crossmodal Emotion Embedding . . . 23
  - 3.1.2 Deep Bag-of-X-Words . . . 27
    - 3.1.2.1 Bag of Audio Words . . . 29
    - 3.1.2.2 Bag of Context-Aware Words . . . 30
- 3.2 From Shallow to Deep Modelling . . . 32
  - 3.2.1 Strength Modelling . . . 32
    - 3.2.1.1 Strength Modelling in Monomodal System . . . 33
    - 3.2.1.2 Strength Modelling for Multimodal System . . . 35
  - 3.2.2 Dynamic Difficulty Awareness Training . . . 37
    - 3.2.2.1 Difficulty Indicators: RE and PU . . . 38
    - 3.2.2.2 DDAT Framework: System Overview . . . 39

3.2.2.3 Difficulty Information Retrieval . . . 40
3.2.2.4 Difficulty Information Exploitation . . . 43
3.2.3 Adversarial Training . . . 44
3.2.3.1 Vanilla Generative Adversarial Networks . . . 45
3.2.3.2 Conditional GANs . . . 46
3.2.3.3 Conditional GANs for Emotion Prediction . . . 47
3.2.3.4 Optimising with Wasserstein Distance . . . 49
3.3 From Isolated to Continual Learning . . . 49
3.3.1 Lifelong Learning and Catastrophic Forgetting . . . 50
3.3.2 Elastic Weight Consolidation . . . 51
3.3.3 Continual Emotion Recognition . . . 53
3.4 Synchronisation Behaviour Analysis Based on Autoencoders . . . 54
3.4.1 Introduction of Synchronisation Behaviour . . . 55
3.4.2 Autoencoder-based Synchronisation Behaviour Analysis . . . 56

4 Experimental Evaluations **57**
4.1 Spontaneous and Multimodal Emotional Databases . . . 57
4.1.1 RECOLA Database . . . 57
4.1.2 SEWA Database . . . 58
4.1.3 OMG-Emotion Database . . . 60
4.2 Experimental Setup . . . 61
4.2.1 Audiovisual Features . . . 61
4.2.2 Performance Measures . . . 63
4.3 Emotion Prediction with Deep Crossmodal Latent Representations . . . 65
4.3.1 Experimental Evaluation . . . 65
4.3.2 Performance . . . 66
4.3.2.1 Results on RECOLA . . . 66
4.3.2.2 Results on OMG-Emotion . . . 69
4.3.2.3 Visualisation of Emotion Embeddings . . . 70
4.3.2.4 Impact of Auxiliary Modalities and Triplet Loss . . . 70
4.3.3 Summary . . . 74
4.4 Emotion Regression Based on Deep Bag-of-X-Words . . . 75
4.4.1 Experimental Evaluation . . . 75
4.4.2 Performance . . . 75
4.4.3 Summary . . . 77
4.5 Strength Modelling-based Emotion Recognition . . . 78
4.5.1 Experimental Evaluation . . . 79
4.5.2 Performance . . . 80
4.5.3 Summary . . . 85
4.6 Emotion Regression via Dynamic Difficulty Awareness Training . . . 85
4.6.1 Experimental Evaluation . . . 85
4.6.2 Performance . . . 86

4.6.3 Summary . . . 91
4.7 Emotion Prediction with Adversarial Training . . . 91
4.7.1 Experimental Evaluation . . . 91
4.7.2 Performance . . . 92
4.7.3 Summary . . . 93
4.8 Continual Emotion Prediction via Lifelong Learning . . . 94
4.8.1 Experimental Evaluation . . . 95
4.8.2 Performance . . . 96
4.8.2.1 Cross-cultural Emotion Recognition . . . 96
4.8.2.2 Hyperparameter Selection . . . 102
4.8.3 Effectiveness Verification . . . 103
4.8.3.1 Discussion . . . 103
4.8.4 Summary . . . 105
4.9 Behaviour Synchronisation Analysis . . . 105
4.9.1 Experimental Evaluation . . . 105
4.9.2 Performance . . . 106
4.9.3 Summary . . . 108

5 Discussion and Outlook **111**
5.1 Contributions . . . 111
5.2 Limitations and Future Prospects . . . 113

Acronyms **115**

List of Symbols **119**

References **123**

List of Figures

3.1 The proposed crossmodal Emotion emBedding (EmoBed) framework for monomodal emotion recognition [87]. 21
3.2 Structure comparison among the proposed joint audiovisual training (e), and other related multimodal learning frameworks (i. e., early fusion (a), late fusion (b), model-level fusion (c)), and multi-task learning (d) [87]. 22
3.3 Framework for learning latent discriminative representations. For each training instance $\mathbf{x}_i$, $\mathbf{x}_i^+$ indicates a randomly selected instance within the same category as for $\mathbf{x}_i$; $\mathbf{x}_i^-$ indicates another randomly selected instance from a different category; $\mathbf{d}_i^+$ and $\mathbf{d}_i^-$ respectively denote the distances between the two latent representations learnt from the instances with the same or different categories. 25
3.4 Diagram of the Bag-of-Audio-Words approach pipeline. 29
3.5 Diagram of the Bag-of-Context-Aware-Words approach pipeline. . . . 31
3.6 Overview of the Strength Modelling framework. 34
3.7 Strength Modelling with early fusion strategy. 35
3.8 Strength Modelling (SM) with late fusion. Fused predictions are from multiple independent modalities with the same model (denoted by the red, green, or blue lines), multiple distinct models within the same modality (denoted by the solid or dashed lines), or the combination. . 36
3.9 Dynamic Difficulty Awareness Training Frameworks with two stages. Difficulty information can be indicated by either the input reconstruction error (i. e., an error vector or the sum of all errors), or the emotion perception uncertainties. Figures are adapted from [234]. . . 39
3.10 Framework of a vanilla Generative Adversarial Network (GAN) [84]. . 46
3.11 Framework of a Conditional Generative Adversarial Network (CGAN) [84]. 47

3.12 Framework of a conditional GAN framework for prediction: the first model (NN_1) predicts time-continuous labels $\hat{y}_t$ from a set of acoustic features $\mathbf{x}_t$, whereas the second model (NN_2) infers a binary decision whether the input source comes from the real data y_t or from the first model NN_1, given the context $\mathbf{x}_t$. 48

3.13 Illustration of Elastic Weight Consolidation (EWC) 52

4.1 Screen-shot taken from one example recording with two subjects in the SEWA database [91]. 59

4.2 Visualisation of the learnt representations of the development set of the RECOLA database when using the proposed EmoBed systems or the classic monomodal systems. Red, green, and yellow markers: representations from audio (eGeMAPS), video (appearance), and video (geometric) modalities; solid and hollow markers: high and low arousal/valence. 71

4.3 Visualisation of the learnt representations of the development set of the OMG-Emotion database when using the proposed EmoBed systems or the classic monomodal systems. Red and green markers: representations from audio and video modalities; solid and hollow markers: happy and sad categories. 72

4.4 Impact of the joint auxiliary modality loss on the *joint audiovisual training* systems for either arousal (a) or valence (b) regression with the RECOLA database. 72

4.5 Impact of the crossmodal triplet loss on the *crossmodal triplet training* systems for either arousal (a) or valence (b) regression with the RECOLA database. 73

4.6 Impact of the joint auxiliary modality loss on the *joint audiovisual training* systems (a), and impact of the crossmodal triplet loss on the *crossmodal triplet training* systems (b), with the OMG-Emotion database. 74

4.7 The effect of the sub-bag's window size on the performance (CCC) when predicting arousal and valence separately. Performances are averaged over all examined time step sizes on the development partition. 77

4.8 Automatic prediction of arousal via audio signals (a) and valence via video signals (b) obtained with the best settings of the *strength*-involved models and *individual* models for a subject from the test partition. 82

4.9 Performance comparison between the single-task learning, the proposed dynamic difficulty awareness training approach based on reconstruction error (RE) or perception uncertainty (PU), and their dynamically-tuned (DT) versions. Results pertain to the test partition for both arousal (a) and valence (b) targets using three feature sets (audio-eGeMAPS, video-appearance, and video-geometric). . . . 88
4.10 Percentage of the contribution of each information stream (a) or model (b) for achieving the best arousal or valence predictions. 89
4.11 Automatic predictions of arousal (a) and valence (b) obtained by conducting conditional adversarial training, for a randomly selected subject from the test partition on RECOLA database. 94
4.12 Visualisation of the performances in terms of CCC of the proposed methods comparing with other baseline approaches on the test sets of FR and GE. Results are separately shown for arousal and valence regressions via audio, video, and their combination (AV), respectively. Note that, the white bars indicate the performance of a matched culture-specific model, while the red dotted lines denote the performance of a joint training model. a: mismatched culture-specific model, b: mismatched model w/ L2-norm, c: sequential training w/o EWC, d: sequential training w/ EWC (proposed). 101
4.13 The effect of the hyper-parameter to control the regularisation λ in the proposed *GE after FR, w/ EWC* model, when predicting arousal and valence on the development sets via three various feature types, i. e., audio, video, and audiovisual (AV). The average performance of both FR and GE is calculated and denoted as avg. 102
4.14 Venn diagrams to visualise the relations among three parameter sets by analysing the important parameters obtained in three models for audiovisual *aro*usal and *val*ence predictions, where each of them can be viewed as a circle. In particular, the red circle denotes a culture-specific model, i. e., 1-FR or 1-GE; the green circle represents another culture-specific model, i. e., 2-GE or 2-FR; and the purple circle indicates a sequential training model that learns task 1 and 2 sequentially. The values in the circles show the number of important parameters, which belong to one set only or lie at the intersection of two or even three sets. 104
4.15 Slope of RMSE sequences of 70 Chinese subjects from 35 recordings. In each recording, there are two subjects as denoted with blue and red bars, respectively. 106
4.16 Slope of RMSE sequences of paired subjects from all recordings of six cultures (C1: Chinese, C2: Hungarian, C3: German, C4: British, C5: Serbian, C6: Greek). In each recording, there are two subjects as denoted with blue and red bars, respectively. 107

List of Tables

4.1 Three partitions of the RECOLA database 58

4.2 SEWA corpus: Number of conversations and subjects and total duration in minutes for each culture. 59

4.3 Performance comparison in terms of CCC for the arousal prediction among the proposed EmoBed systems, related baselines, and other state-of-the-art systems. These results pertain to the experiments conducted on the *dev*elopment and *test* partitions of the RECOLA database. Three feature sets (audio-eGeMAPS A, video-appearance V_{app}, and video-geometric V_{geo}) were employed to evaluate all approaches. The cases where EmoBed systems have a statistical significance of performance improvement over the classic monomodal systems are marked by the "$*$" symbol. 67

4.4 Performance comparison in terms of CCC for the *valence* prediction among the proposed EmoBed systems, related baselines, and other state-of-the-art systems. These results pertain to the experiments conducted on the *dev*elopment and *test* partitions of the RECOLA database. Three feature sets (audio-eGeMAPS A, video-appearance V_{app}, and video-geometric V_{geo}) were employed to evaluate all approaches. The cases where EmoBed systems have a statistical significance of performance improvement over the classic monomodal systems are marked by the "$*$" symbol. 68

4.5 Performance of the proposed EmoBed systems, related baselines, and other reported systems in terms of F1 on the development set of the OMG-Emotion dataset. The cases where EmoBed systems have a statistical significance of performance improvement over the classic monomodal systems are marked by the "$*$" symbol. 69

4.6 Performances in terms of Concordance Correlation Coefficient (CCC) of the proposed BoCAW features with various window sizes in the first stage (W_1), for both *arousal* and *valence* regressions, evaluated on the *dev*lopment and *test* partitions. Note that, for each W_1, only the best performance among four examined time step sizes (Ts_1) is reported, by calculating the averaged predictions of arousal and valence on the development set. The best results achieved are highlighted. The symbol of $*$ indicates the significance of the performance improvement over the bag-of-audio-words (BoAW) baseline method. 76

4.7 Performances in terms of CCC of the proposed method comparing with other state-of-the-art approaches on the RECOLA dataset. The best results achieved are highlighted. The symbol of $*$ indicates the significance of the performance improvement over the bag-of-audio-words (BoAW) method. 78

4.8 The optimised complexity (C) of SVR and number (N) of hidden nodes per layer of BLSTM-RNN for different types of modality and task. 80

4.9 Performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) on the *development* and *test* partitions from the *audio* signals. The best achieved CCCs are highlighted. The symbol of $*$ indicates the significance of the performance improvement. . . . 80

4.10 Performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) on the *development* and *test* partitions from the *video* signals. The best achieved CCCs are highlighted. The symbol of $*$ indicates the significance of the performance improvement. . . . 81

4.11 Performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) with *early fusion strategy* on the development and test partitions. The best achieved CCCs are highlighted. The symbol of $*$ indicates the significance of the performance improvement. . . . 83

4.12 Performance comparison in terms of RMSE and CCC between the *str*ength-involved models and the *ind*ividual models of SVR (S) and BLSTM-RNN (B) with *late fusion strategies* (i.e., modality-based, model-based, or the combination) on the *dev*elopment and *test* partitions. The best achieved CCC is highlighted. The symbol of $*$ indicates the significance of the performance improvement. 84

4.13 System performance comparison in CCC for the conventional single-task learning (baseline) framework, the multi-task learning (MTL) framework, and the proposed Dynamic Difficulty Awareness Training (DDAT) framework using reconstruction error (RE, a vector as v or a scalar of sum as s) and perception uncertainty (PU) variants. These results pertain to the experiments conducted on the *dev*elopment and *test* partitions for both arousal and valence targets. Three feature sets (audio-eGeMAPS, video-appearance, and video-geometric) were employed to evaluate all approaches. The best results achieved on the test set are in bold. The cases where DDAT has a statistical significance of performance improvement over MTL are marked by the "$*$" symbol. 87

4.14 Late fusion performance in terms of CCC in different fusion strategies (i. e., modality-based, modality- and model-based, and dynamically-tuned modality- and model-based) for the *dev*elopment and *test* partitions of both *aro*usal and *val*ence regressions. The predictions are generated from the reconstruction-error-based DDAT framework (P_{re}) or the perception-uncertainty-based DDAT framework (P_{pu}); their dynamically-tuned versions ($P_{re,dt}$ or $P_{pu,dt}$); or the baseline model (P_{bs}). The best results achieved on the test set are in bold. Note that P_{re}, $P_{re,dt}$, P_{pu}, $P_{pu,dt}$, and P_{bs} are the fused predictions from three diverse feature sets. 90

4.15 Performance in terms of Concordance Correlation Coefficient (CCC) of the proposed conditional adversarial training approaches, as well as its variation (+ Wasserstein distance), for both *arousal* and *valence* regressions, evaluated on the *dev*elopment and *test* partitions. 93

4.16 CCC performances via various training strategies for emotion regression based on *audio*, *video*, or the *combination*. Performance on the development sets and test sets of the two databases ($FR_{dev}, FR_{test}, GE_{dev}, GE_{test}$) as well as the average performance on the two test sets (μ_{test}) are reported for *arousal*, respectively. 97

4.17 CCC performances via various training strategies for emotion regression based on *audio*, *video*, or the *combination*. Performance on the development sets and test sets of the two databases ($FR_{dev}, FR_{test}, GE_{dev}, GE_{test}$) as well as the average performance on the two test sets (μ_{test}) are reported for *valence*, respectively. 98

4.18 Average slope of RMSE sequences of all subjects within six different cultures is listed in the upper row, respectively; the correlation coefficient denoted as *pcc of pairs* indicates the correlation of behaviours of two subjects and is listed in the last row for each culture (C1: Chinese, C2: Hungarian, C3: German, C4: British, C5: Serbian, and C6: Greek). 108

1 Introduction

1.1 Motivation

Emotional behaviour analysis of human beings is a multidisciplinary research field, spanning anthropology, cognitive science, linguistics, psychology, and computer science [14, 125, 152, 166, 176, 207]. In particular, in the field of Human-Computer Interaction (HCI), detecting and understanding emotional states of humans automatically is essential, to improve the effectiveness of HCI systems and devices, via providing an affective-based personalised user experience [38, 146]. In addition, thanks to the great developments of machine learning, especially the tremendous advancements in deep learning in recent years, we have witnessed fruitful theoretical and empirical works, which enable machines to recognise meaningful patterns of emotional behaviours by people [47, 84, 196, 201, 211]. In this light, these innovative technologies have facilitated or are urging various real-world applications to handle affective information. For example, in the context of healthcare, an emerging trend in medical diagnostic methods is to detect cognitive and developmental diseases (e. g., autism and depression) based on automatic emotional behaviour analysis [3, 44]. Likewise, for education applications, adaptive e-learning platforms can provide personalised support, by taking into account the learners' emotion states in the learning process [9, 168]. Furthermore, making good use of users' emotions can lead to better user experience in applications such as conversational systems [240], recommender systems [81], and socially assistive robotics [208].

In principle, in order to acquire affective information of an individual, a large variety of modalities can be exploited. These modalities include, but are not limited to, facial expressions [55], hand gestures [70], speech [229], text [4], and physiological signals such as Electrocardiogram (ECG) [202], Electroencephalogram (EEG) [2], and Electromyogram (EMG) [106]. Furthermore, it is believed that integrating the complementary information of diverse modalities can contribute to further performance and robustness improvement of emotion recognition systems, when compared with a unimodal system. For this reason, various data fusion strategies across modalities

have been extensively investigated [154, 186, 199]. Specifically, due to its practical importance in real-world applications, a considerable amount of achievements have been reported by combining acoustic and visual observations, to improve emotion recognition accuracy [83, 109, 142]. In other words, automatic analysis of audiovisual behaviour is an active and promising research area, and has been a primary focus in diverse HCI applications. This can probably be attributed to two main reasons. On the one hand, facial expressions and speech are deemed as two of the most direct channels to transmit human emotions [213]. On the other hand, audiovisual sensors such as cameras with integrated microphones are easily accessible and becoming indispensable of daily life more and more.

As aforementioned, the recent advent of deep learning techniques has highlighted the possibility of automatic emotion analysis from audiovisual signals. For instance, the first attempt to apply Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) for long-range context modelling in Speech Emotion Recognition (SER) was by Wöllmer et al. [211]. After that, the same architectures were investigated for audiovisual emotion recognition in [212]. In addition, Neagoe et al. [140] presented a Facial Emotion Recognition (FER) approach using Convolutional Neural Networks (CNNs), to learn affect-salient, discriminative, and high-level features. On top of that, recently, the combination of CNNs and LSTM-RNNs, firstly constructed in [196], has been shown to be a promising approach for estimating dimensional emotions in an end-to-end manner.

Despite the great progress that has been made so far in the development of audiovisual-based emotion recognition approaches, a number of challenges remain in order to reach full applicability of current systems, and five main of them are covered in this thesis and listed below:

1. Many systems are trained on the data collected in controlled settings, and therefore their performances are usually severely degraded in real-world applications and tools trained on such data. In early days, most, if not all, emotional data were recorded under laboratory or controlled conditions, and acted [23, 76]. On the one hand, these data can be of great interest in certain circumstances, e. g., when the objective is to find the portrayal of specific emotional details [25, 57]. On the other hand, however, acted data cannot be equally helpful for training if one wants to predict the naturalistic display of emotions. Note that, it is inevitable that there is a 'mismatch' between laboratory conditions and realistic scenarios. For example, spontaneous emotions are much more complex and subtle in comparison with acted ones. As a consequence, to minimise the gap between the training and inference phases and further to facilitate affect-aware systems into our daily life interface, the crux in devising such systems is the collection of realistic behavioural recordings made *'in-the-wild'*, that is, in truly unconstrained and real-world contexts.

2. Armed with these 'in-the-wild' data from multiple sensors, a succeeding challenge is how to represent these raw data to picture emotional cues of our interests. Efforts have constantly been made in the context of machine learning, to represent data in a format that a computational model can work with [10]. In particular, the ability to extract *salient representations* specific to emotions is crucial, and forms the backbone of any emotion prediction model [84]. From video recordings, for instance, appearance-based features can be computed based on the detected facial landmarks, to obtain compact but informative facial representations [112]. Similarly, in speech emotion recognition, comprehensive and standardised audio feature sets are provided in the widely used openSMILE toolkit, covering spectral, cepstral, prosodic, and voice quality information [60, 61], and been successfully deployed in various paralinguistic recognition tasks [177, 178, 203]. Furthermore, inspired by the success of deep learning, an emerging trend is to learn data-driven representations rather than hand-crafted features for specific tasks [10]. So far, several promising data-driven representation learning approaches have been presented and helped to prevail over the state of the art, in particular for emotion recognition [69, 85, 127]. Yet, relevant research work is still urgently needed to deal with difficulties involved and to improve the performance of automated audiovisual emotion recognition in the wild.

3. Another important challenge is how to tailor existing *deep learning algorithms*, to address a number of special requirements arising from the task at hand, i. e., emotional behaviour analysis. Up to now, most of these techniques have been originally proposed for some other tasks, and then certainly shown beneficial throughout many applications. However, it is arguably inadequate to expect the same benefit arising in emotion recognition, by applying these frameworks directly without considering the property of the task. To this end, these algorithms and structures have to be adjusted accordingly, to fulfil the needs in several aspects such as effectiveness and robustness, in the context of emotion recognition. For a more in-depth discussion on these requirements, the interested reader is referred to [173]. Moreover, given as another example, meta information such as age and gender can provide rich speaker trait information. When the auxiliary information is exploited, remarkable performance improvements have been observed [108, 228]. Overall, this is an active field of research and new approaches are still introduced.

4. A fourth challenge is how to empower the systems to learn in a *lifelong context*. That is, instead of learning several models in isolation, one per task, one general model can be built via learning from multiple given tasks sequentially. This is rather essential for applying emotion recognition systems in real life, where reliable and effective emotion predictions are expected in various

scenarios, e. g., across different cultures and languages. To deal with this issue, in previous work, training strategies such as transfer learning [47] and multi-task learning [226] have been explored in affective computing, aiming at transferring or sharing the knowledge among multiple tasks and relaxing the task-dependent constraint to some degree. Nevertheless, a huge amount of training data from all tasks of interest are demanded, leading to large storage memory requirement and heavy computational load. Hence, advanced and innovated training strategies are needed which can endow machines the ability to retain and accumulate knowledge learnt over past tasks and then to apply this knowledge to future tasks.

5. The fifth challenge is *empathic behaviour analysis*, one of the most overlooked mechanisms in intelligent systems today. Empathy can be defined as a complex process, whereby "an observer reacting emotionally because he perceives that another is experiencing or about to experience an emotion" [210]. Indeed, it has been exemplified that, empathic agents are perceived as more likeable, smart, and trustworthy than non-empathic ones [21]. In another previous study, it has been shown that empathic devices can foster empathic feelings in users [145]. From this perspective, future machines should be endowed the ability to behave in an empathic manner, aiming at establishing and maintaining positive and long-term relationships with users [116, 150]. Nevertheless, in contrast to well-established emotion categories such as happy and sad, empathetic behaviour is more complex and difficult to be analysed quantitatively and qualitatively [116]. Whilst limited research efforts have been made up to now, deep empathic behaviour analysis is still at its very beginning and has a long way to go.

In this respect, the main focus of this thesis is to discuss and address the above mentioned five challenges and limitations, by leveraging and investigating diverse deep learning techniques. The next section will present the objectives of this thesis in an explicit way.

Apart from these five limitations, there are some other open issues and emerging challenges, and finding robust solutions to these challenges is an open and ongoing research interest. As a consequence, to provide competing platforms and benchmark test sets for emotion recognition researchers, a series of challenges have been organised, such as Audio Video Emotion Challenge (AVEC) [180], Emotion Recognition in the Wild Challenge (EmotiW) [49], Multimodal Emotion Recognition Challenge (MEC) [117], and One-Minute Gradual Emotion Recognition (OMG) [13]. In these challenges, different tasks and issues have been proposed and discussed, including depression analysis, engagement recognition, group emotion recognition, and so on and so forth [50, 162]. It is noteworthy that, since these challenges are based on spontaneous emotional data, it is more practical for researchers to compare the pros and cons of various algorithms in real-world environments. Additionally, most, if

not all, algorithms and approaches presented in this thesis have been verified on at least one of these benchmarks.

1.2 Aims of the Thesis

In the light of the above considerations, this thesis attempts to deal with the five key challenges (cf. Section 1.1), aiming to help allow for improved emotional behaviour analysis in the new generation of emotional AI systems. In particular and by that, the contributions of this thesis are to find answers to three primary Research Questions (RQs) as well as another two side RQs, and can be summarised as follows.

RQ-1: How do deep learning approaches perform when models are trained and evaluated on spontaneous affective data? To tackle the first challenge, the focus of this thesis will be set on analysing spontaneous affective data that are automatically acquired from smart audiovisual sensors. For this reason, instead of other small, prototypical, and often non-natural datasets, three standard spontaneous databases are targeted, i. e., the "Automatic Sentiment Analysis in the Wild" database (SEWA) created by the author and her colleagues, the OMG-Emotion behaviour dataset (OMG), and the "Remote COLlaborative and Affective interactions" dataset (RECOLA). These three databases are featuring in recordings captured under realistic scenarios, and will be employed in this thesis to construct models that are more beneficial in realistic applications.

RQ-2: How can we strengthen the representation learning process to learn useful representations for emotion recognition? To deal with the second challenge, i. e., the representation learning challenge, one aim of the thesis is to distil temporal context-aware emotional and less modal-variant discriminative representations from audio and video signals. In particular, the objective is two-fold, i. e., (a) to encapsulate successive low-level representations to create temporal context-aware high-level representations for emotion recognition, and (b) to attain a shared embedding space to explore the latent correlation between audio and video modalites. Moreover, this may also shed new light on the data sparsity issue met in affective computing domain, as knowledge learnt from a rich-resource modality can merit the training of a system for a low-resource modality.

RQ-3: How can we customise conventional deep learning-based models when concerning emotional behaviour analysis? To face the third challenge, i. e., the deep modelling challenge, substantial efforts are made in this thesis by leveraging suited and innovative deep learning algorithms to help approach increased performance and robustness of affective-aware intelligent systems. The efforts can mainly be grouped into two categories: First, several algorithms are proposed to concurrently reap the benefits and restraining the drawbacks of distinct deep learning-based models, to advance the benchmarks for emotion recognition in

the wild. Second, in the context of emotional behaviour analysis, approaches to exploit the subjective characteristic of emotion perception are also investigated, with an aim to boost the learning process of deep learning models.

RQ-a (side): Can emotion recogniser adapt continually and keep on learning over time? To tackle this challenge, a lifelong learning paradigm is, for the first time, introduced in this thesis for audiovisual emotion analysis. In particular, instead of learning culture-dependent models in a conventional isolated learning strategy, a widely used lifelong learning approach of elastic weight consolidation is investigated. With this manner, a general model can be obtained, which is capable of performing emotion recognition in a cross-culture scenario.

RQ-b (side): Can deep learning structures be exploited for automatic empathy behaviour analysis, even for unlabelled emotional data? To deal with the fifth challenge, i. e., the empathy emotion detection challenge, a preliminary antoencoder-based empathetic behaviour analysis model is also performed in this thesis. It extends the present emotional behaviour analysis which studies explicit emotion patterns, by showing the potential of automatically detecting mimicry behaviours via deep neural networks. This topic is still in its early development, which, however, can be of a broad interest in both academic and commercial communities.

1.3 Structure of the Thesis

To provide a good overview for the reader, the present study is structured into five main chapters as follows.

Chapter 1 primarily presents an introductory motivation and the raised objectives of the thesis, followed by a structure of the present study.

Chapter 2 covers the theoretical background of intelligent emotional recognition systems. First, fundamental knowledge of emotion recognition is concerned, by providing a brief survey of existing feature sets, models, and tools that are commonly used in affective computing, for audio and video, respectively. Next, this chapter discusses the recent developments in audiovisual emotion recognition, emphasising various fusion approaches in particular. Finally, it describes the state-of-the-art deep learning techniques and frameworks in the field of affective computing, associated with representation learning and emotion modelling, respectively.

Chapter 3 mainly concentrates on a set of concrete methods, which are proposed in this thesis, for dealing with challenged outlined in Section 1.1. This will include deep latent representation learning and bag-of-context-aware-words representation learning, with the goal of obtaining salient and meaningful representations. Furthermore, this also consists of a series of emotion recognition frameworks and training strategies, involving strength modelling, dynamic difficulty awareness training, adversarial training, and crossmodal joint training. Next, cross-cultural emotion

recognition based on lifelong learning is investigated. Lastly, for the first time, the advantage of deep learning is exploited for automatic empathetic behaviour detection in this chapter.

Chapter 4 describes practical evaluations of the methods presented in Chapter 3, yielding better performance in automated audiovisual emotional behaviour analysis in the wild. All experiments are performed with databases that are publicly available to the research community. After introducing the collected databases and related experimental setups, such as the evaluation measures, the performances of the corresponding approaches are then reported and analysed.

Chapter 5 summarises the contributions and their weaknesses. In addition, future research potential is outlined.

2 Theoretical Background

The focus of this chapter is on providing with readers the theoretical background of audiovisual emotion recognition for a better comprehension of the following parts thereafter. In particular, basic knowledge of emotion recognition will be covered in Section 2.1. Further, monomodal and multimodal emotion recognition will be discussed in Section 2.2 and Section 2.3, respectively. Finally, a brief overview of the state of the art in deep learning-based feature extraction and emotion prediction will be given in Section 2.4 and Section 2.5, respectively.

2.1 Emotion Recognition in General

In affective computing, emotion recognition can be defined as a process of automatically acquiring the affect of human beings, and it usually leverages methodologies and techniques from multiple research areas covering signal processing, machine learning, and so on.

Along with the studies developed in the theories of emotion, two kinds of emotion models are frequently explored in affective computing, namely, *categorical models* and *dimensional models*. That is, the perceived emotions can be presented as discrete labels from multiple discrete categories or continuous values in dimensional spaces. More specifically, in a typical categorical emotion model, six basic categories are suggested, covering anger, disgust, fear, happiness, sadness, and surprise, also known as Ekman's Six Basic Emotions [54]. Moreover, other than these 'big six' emotions, there are still a large number of additional emotions of interest that are understudied, such as contempt, distress, guilt, interest, panic, shame, etc.

Alternatively, considering dimensional models, it is presupposed that any emotion can be described in terms of certain dimensions, although the dimensions can vary from one research to another. The most popular dimensional model is Russell's Circumplex model of affect [165], where each emotion is defined by its position on the dimensions of arousal (the degree of intensity of an emotional state) and valence (how positive or negative an emotional state is) in a circular representation.

Likewise, other variations of Russell's dimensional models refined this model and suggested additional dimensions such as dominance, engagement, and pleasantness.

Apart from these two models, further models of interest include cognitive models [144] and interactional models [18] which are much more complicated but might be useful for us to understand the complexity of emotions and develop the next generation of affective-sensitive HCI systems.

Having discussed various models of emotions, in the following, various modalities where emotions can be displayed will be discussed. In human communications, emotions can be transmitted via multiple modalities, such as audio, text, video. In other words, information of a person's emotional state can be found simultaneously or sequentially by signals of multiple modalities. For example, from audio signals, emotions can be presented by linguistic features such as the choice of word as well as paralinguistic features such as the tone of voice and the rate of speech. Similarly, human emotion and affective states can also be detected from visual-based features such as facial expression, gesture, and body postures.

With that said, different emotion recognition systems can be raised by the selection of the modality to form *monomodal emotion recognition systems*. For instance, one can build a speech emotion estimator based on audio signals. Alternatively, one can constitute *multimodal emotion recognition systems* where information from different modalities are taken into consideration when estimating human emotion and affective states. One rationale behind is that various modalities might deliver complementary information. While a particular modality might dominate during conveying a certain type of emotion, the combination of it with other modalities might be critical for other types of emotions. Hence, although early studies mainly focused on recognising basic emotions from facial expressions, an ongoing trend in the field is to apply multimodal information to enhance the effectiveness and efficiency of emotion recognition systems [10, 24].

Also, when establishing an emotion recognition system, one crux is the availability of emotional data to be utilised for training. In this regard, the quantity of labelled data should be suitable. Nonetheless, collecting and labelling these emotional data are prohibitively expensive, time-consuming and at some point subjective. Further, people are reluctant to share these data as it raises security and privacy concerns. Given these reasons, a major effort has been made is to manage data collection and annotation in an efficient manner [139, 93]. Other than efficient data collection, promising techniques have also been proposed to cope with the scarcity issues via exploiting additional unlabelled data, including active learning, cooperative learning, and semi-supervised learning. Over the past years, these learning mechanisms have been applied successfully to efficiently exploit additional unlabelled data for emotion recognition [229, 231, 237, 235].

In addition to the data challenge, two further active research subfields in the domain are the design of suitable emotional representations from the raw data (generating representations that can well be processed by a machine), and the selection

of approaches to model the emotional pattern (developing algorithms that can automatically perform the task with the given representations). Specifically, general discussions of the two challenges in the context of deep learning will be provided in Section 2.4 and Section 2.5, respectively.

Finally, for a comprehensive overview and discussion of the state-of-the-art, remaining challenges and open issues in this field, interested readers are kindly referred to [152, 223, 79, 175, 195].

2.2 Monomodal Emotion Recognition

As mentioned in Section 2.1, monomodal emotion recognition systems normally independently explore the prominent features for the emotions of interest, from a certain modality [2, 56, 94, 34]. Particularly, with the rise of deep learning in the field, monomodal emotion recognition systems have been successfully employed and achieved appealing prediction performance [84, 225, 1]. In general, a typical monomodal emotion recognition system consists of two stages: *feature extraction* and *feature exploration*. More specifically, in the first stage, effective features need to be generated from the raw data to capture relevant information on emotional characteristics, whereas in the second stage appropriate machine-learning algorithms are selected and performed to automatically recognise emotions from features acquired from the prior stage. In this section, mainly two types of such monomodal systems for HCI are discussed, i. e., speech emotion recognition systems and facial emotion recognition systems, while these two modalities are two of the most important communicative modalities in human-human interaction [223]. Particular, as one may notice, the feature extraction is critically vital to the whole process, since good recognition performance cannot be gained without appropriate features. Hence, hereinafter the primary focus lies on the feature extraction stage of the two aforementioned systems.

First, concerning Speech Emotion Recognition (SER), discriminative and concise features need to be defined to best reflect the emotional content. Moreover, the extracted features should be robust against various conditions, e. g., recording devices and environments, language and cultural differences. In this regard, most commonly used acoustic features are rather low-level such as energy and loudness.Additionally, further prosodic features such as pitch and zero-crossing rate can be explored, as these features can describe the intensity, intonation, and rhythm of the speech which are highly associated to the emotional information in speech [174]. For instance, high intensity might indicate anger or surprise, and in contrast, low intensity may imply disgust or sadness. Alternatively, various spectral features and cepstral features are commonly utilised to recognise emotions from speech data. Commonly used spectral features involve the formants, spectral centroid points, spectral energy, spectral sharpness, and spectral slope. Likewise, frequently used

cepstral features cover Mel-frequency bands, Mel-Frequency Cepstral Coefficients (MFCCs), and Linear Prediction Cepstral Coefficient (LPCCs). All these aforementioned acoustic features are extracted on frame levels, and are usually referred to as Low-Level Descriptors (LLDs). Once these frame-level features are extracted, statistics can be gained by applying computing statistical functionals on these LLD contours over a series of frames to derive supra-segmental level features. Functionals that are often used contain extremes, means, moments, peaks, and percentiles.

To facilitate the acoustic feature extraction process, the openSMILE toolkit [61] that provides a number of predefined and well-established feature sets can be employed. Further detailed discussions of the acoustic features applied in this thesis will be given in Section 4.2.1. For an overview of the conventional features applied in speech emotion recognition, the reader is referred to [58]. Moreover, other than these hand-crafted acoustic features, learning data-driven features directly from raw data via a deep neural network is another promising research subfield, and more related discussions will be given in Section 2.4. Further, for a summary of the state-of-the-art techniques, open competitive challenges and future tendencies in and for speech emotion recognition, the interested readers are kindly referred to [56, 174].

Next, facial emotion recognition, or FER for short, is another vitally important component in recent HCI systems. In order to take non-verbal cues from facial emotions as humans, conventional FER systems usually consist of four main stages. First, in the pre-processing stage, noises are removed and meanwhile, the image sequence is enhanced. Second, in the face registration stage, the region of interest is registered, which could be either the whole face or parts of the face such as eyes and the mouth. Thereafter, the third stage is for face representation extraction where features can be extracted from the face with different approaches. Generally speaking, the approaches can be categorised into geometric-based versus appearance-based approaches, local versus holistic approaches, or static versus dynamic ones [37]. In particular, in geometric-based methods, features are extracted to describes the facial shape and activity by considering the location and deformation information of the facial components, whereas appearance-based paradigms simply encode textural information by using the intensity information of the images. Also, a variety of popular representations have been studied in the community, including but not limited to Gabor filters, Histogram of Oriented Gradients (HOG), Landmark locations and distances, Local Binary Pattern from Three Orthogonal Planes (LBP-TOP), Local Phase Quantisation (LPQ), and Scale-Invariant Feature Transform (SIFT). Further discussion of the features applied in the thesis will be given in Section 4.2.1. Finally, in the fourth stage, similar to SER, various machine learning algorithms can be performed to model the emotional characteristics accordingly.

It should be noted that, in most cases of the facial representation extraction process, it is essential to apply an additional dimension reduction process, as the original dimension of the obtained visual features is prohibitively large. In this respect, a subset of features could be selected with a range of techniques such as

AdaBoost and GentleBoost. Alternatively, transformation methods, e. g., Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), could be performed to map the higher dimensional representations into a lower-dimensional space. With dimension reduction, issues such as illumination variation, registration errors, and identity bias can be relieved to some extent [169]. Other than that, as an active research area, further discussion and comprehensive analysis associated of FER can be found in [37, 169, 63], covering the state-of-the-art techniques, key challenges as well as future research perspectives.

2.3 Audiovisual Emotion Recognition

Over the past two decades, considerable research efforts have been reported in developing SER and FER systems but largely independent of each other. However, recently more and more researchers advocate integrating audio and visual cues for emotion recognition [141]. One reason behind is that, as humans, we can process audio and visual information simultaneously for recognising emotions. Also, previous work indicated that some emotions are visually dominant while some others are auditory dominant [42]. As a consequence, different from these aforementioned monomodal systems in Section 2.2, advanced multimodal emotion recognition systems are developed to jointly exploit information from multiple modalities, to make use of complementary or supplementary information from various media cues. In this respect, such systems have consistently demonstrated superior performance when compared with the monomodal systems in numerous previous works [83, 234].

In particular, with the aim of addressing the problem of the automatic analysis of spontaneous affective behaviours, authentic affective data recorded in the wild are essential. That is, the data is supposed to be captured in completely unconstrained real-life environments. This is because data acquired in the laboratory are often with controlled conditions, and are not identical in real life. Hence, models trained with such controlled data do not generalise well and thus cannot be applied directly in real-world applications. For an overview of existing emotional databases, the reader is referred to [223, 112]. Nonetheless, it should be noted that many emotional databases are not publicly available, mainly due to the data privacy issue and the data administration issue [223].

Also, aiming to provide well-defined benchmarks, a range of challenges have been organised in this growing field to boost multimodal emotion recognition system performance. Among these efforts, the outstanding AVEC challenge can be dated back to 2011, and the latest AVEC challenge held in 2019 concentrated on estimating the spontaneous affective states across three different cultures with data recorded in the wild [180, 160]. Likewise, some other challenges have been run with the data captured from media-materials such as monologues of YouTube videos and clips of

films, namely, the annual EmotiW challenges since 2013 [49], the MEC challenge in 2016 and 2017 [117], and the OMG challenge in 2018 and 2019 [13].

Another critical issue in multimodal systems is how to combine the information from the multisensory data. To this end, in the context of audiovisual emotion recognition, a variety of fusion strategies have been extensively investigated, and they normally can be divided into three categories: *feature-level* fusion (also known as early fusion), *decision-level* fusion (also known as late fusion), and *model-level* fusion [223]. Typically, feature-level fusion straightforwardly concatenates audio and visual features into one combined feature vector, which is then used as the input for modelling. With feature-level fusion, better results have been gained and reported [209, 83, 239]. However, it often suffers from the high dimensionality of the feature space and the synchronisation problem of different data streams [239]. On the contrary, in the decision-level fusion, the predictions, rather than the features, reaped from different modalities are combined to come up with a final decision by the use of certain suited criteria [134]. Consequently, the decision-level data fusion has been reported to be highly beneficial for audiovisual emotion recognition [141, 198, 91]. In this method, however, the mutual correlation among different streams is overlooked. To tackle this issue, model-level fusion approaches have been studied to fuse the intermediate representations instead, with the aim of exploiting the correlation between audio and visual features while at the same time releasing the synchronisation requirement of the data streams [223]. Typical solutions associated with model-level fusion are mainly based on probabilistic graphical models, such as Bayesian Networks (BN), Dynamic Bayesian Networks, Hidden Markov Models (HMM), as well as their hierarchical variants [187, 183, 224]. However, it is still a challenging problem to integrate the information from different modalities, and consequently an increasing number of research efforts are made towards handling associated issues, such as the different time scales and temporal structures of multimodal signals, the varied reliabilities of different streams, ways to handle the missing data, the segmentation and alignment, etc.

Beyond these two challenges and other challenges described in Section 1.1, further recent advances and emerging issues under multimodal settings are extensively discussed in [10], such as learning joint or coordinated representations from multimodal data, and facilitating the modelling of a low-resource modality through exploiting another rich-resource modality. This thesis will address some of these typical issues elaborately to enhance audiovisual emotional behaviour analysis in real-life settings.

2.4 Deep Learning for Feature Extraction

Rather than the engineered acoustic and visual features discussed in Section 2.2, with the advent and development of deep learning algorithms and techniques, data-

driven features have emerged and achieved competitive or even superior prediction performance than conventional hand-designed features for specific applications [174]. In particular, recently a huge amount of efforts have been made towards learning affect-salient, discriminative, and high-level representations for emotional behaviour analysis [69, 115, 222, 138]. In the following, basic information of some popular deep representation learning structures is provided, which will be used in later parts of the thesis. For a broader and more in-depth overview of a variety of deep representation learning techniques, the reader is referred to [15, 118].

First, one may learn high-level representations with RNNs by capturing the context information along a temporal sequence of low-level descriptors. In the following, two popular hidden units employed in an RNN architecture are described, namely, the Long Short-Term Memory (LSTM) unit and the Gated Recurrent Unit (GRU), both of which have been frequently exploited in this thesis.

Concerning a typical LSTM unit, it consists of a self-connected memory cell c and three gate units to control the information flow, namely the input gate i, the output gate o, and the forget gate f. These three gates allow the network to learn when to read, write, or reset the value in the memory cell. Mathematically, the activation of the associated cell and gates can be represented as

$$\mathbf{i}_t = \sigma_g(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + b_i), \tag{2.1}$$
$$\mathbf{f}_t = \sigma_g(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + b_f), \tag{2.2}$$
$$\mathbf{o}_t = \sigma_g(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + b_o), \tag{2.3}$$
$$\mathbf{c}_t = \mathbf{f}_t \cdot \mathbf{c}_{t-1} + \mathbf{i}_t \cdot \sigma_c(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + b_c), \tag{2.4}$$
$$\mathbf{h}_t = \mathbf{o}_t \cdot \sigma_h(\mathbf{c}_t), \tag{2.5}$$

where $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$, $\mathbf{c}_t$, and $\mathbf{h}_t$ denote the corresponding activation vectors at the time t of the input gate, the forget gate, the output gate, the cell state, and the hidden state, respectively. Further, σ_g indicates a sigmoid function, whereas σ_c and σ_h are tanh activation functions. Moreover, $\mathbf{W}_{x(\cdot)}$ and $\mathbf{W}_{h\cdot}$ are weight matrices of the connections from each component to the input vector $\mathbf{x}$ and to the previous hidden state vector $\mathbf{h}_{t-1}$, respectively, and $b_{(\cdot)}$ indicates bias vectors. As a consequence, the weight matrices and bias vectors are the neural network parameters that need to be learnt during training. Such a structure grants LSTM-RNN structures to learn the context in both short and long range, and meanwhile, tackles the vanishing gradients problem met in a vanilla RNN framework. Beyond the aforementioned LSTM cell configuration, other LSTM variants have also been investigated in the literature, such as LSTM units with peephole connection or with combined forget and input gates. For a more in-depth explanation of LSTM-RNNs, the reader is referred to [98].

Then, a GRU cell consists of two gates only, namely, a reset gate r and an update gate z, to adaptively regulate how much each hidden unit remembers or forgets while reading a series of inputs. Further, different from an LSTM unit, the memory cell

state c is eliminated in GRUs. Therefore, it has fewer parameters and thus is faster to be trained than the LSTM unit. More important, it has been shown to exhibit comparable or even better performance than LSTM on certain tasks [33]. In formal, the activations associated within a GRU cell can be formulated as

$$\mathbf{z}_t = \sigma_g(\mathbf{W}_{xz}\mathbf{x}_t + \mathbf{W}_{hz}\mathbf{h}_{t-1} + b_z), \tag{2.6}$$
$$\mathbf{r}_t = \sigma_g(\mathbf{W}_{xr}\mathbf{x}_t + \mathbf{W}_{hr}\mathbf{h}_{t-1} + b_r), \tag{2.7}$$
$$\mathbf{h}_t = (1 - \mathbf{z}_t) \cdot \mathbf{h}_{t-1} + \mathbf{z}_t \cdot \sigma_h(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}(\mathbf{r}_t \cdot \mathbf{h}_{t-1}) + b_h), \tag{2.8}$$

where $\mathbf{x}_t$, $\mathbf{h}_t$, $\mathbf{z}_t$, and $\mathbf{r}_t$ indicate the input vector, output vector, update gate vector, and reset gate vector, respectively. Again, σ_g and σ_h are the sigmoid and tanh activation function. Again, the weight matrices $\mathbf{W}_{(\cdot)}$ and bias vectors $b_{(\cdot)}$ are trainable parameters. Moreover, some other GRU variations have been proposed based on this formation, and for more information the reader is referred to [33].

With novel architectures base on these RNN units, data-driven representations have been learnt with either fully labelled, or partially labelled, or even unlabelled datasets for emotion recognition tasks. For example, in [214] and [215], LSTM-RNN-based autoencoder frameworks were employed to provide discriminative representations by exploiting contextual information for speech emotion recognition. Besides, sparse autoencoders and denoising autoencoder RNNs were also investigated for feature transfer learning under cross-database conditions [47, 46].

Similarly, with the popularity of CNN frameworks, nowadays they are frequently applied to learn representations directly from the pixels of the face, as an alternative to classic hand-designed features. As core components in a typical CNN architecture, convolution operations are utilised to replace the affine transformation in conventional fully connected feedforward neural networks [113]. More specifically, a convolution layer consists of a set of learnable filters (or kernels) to extract specific local patterns at each local region of input images or feature maps. This relies on the assumption that if a filter is useful to detect a specific pattern at some spatial position, it should be suitable at other locations. Hence, filters normally have a small local receptive field but are replicated across the entire input visual field, and generate a set of feature maps, the number of which is equal to the number of the filters. Then, these obtained feature maps are stacked together, forming the full output of the convolution layer. Mathematically, this process can be represented as

$$(h_k)_{ij} = (\mathbf{W}_k * \mathbf{x})_{ij} + b_k, \tag{2.9}$$

where $(h_k)_{ij}$ indicates the (i, j) element of the k-th output feature map, and $\mathbf{x}$ represents the input feature maps. Further, $\mathbf{W}_k$ and b_k denote the trainable weight matrix and bias of the k-th filter, respectively. Last, the symbol $*$ implies a 2D spatial convolution operation between the filter and the input feature map. With this manner, parameters are shared and thereby constitutes the translation invariance

property of a CNN architecture. Furthermore, a pooling operation is often carried out after the convolution operation, to reduce the spatial size of the feature maps, resulting in further fewer parameters to be learnt in the subsequent layers. Two of the most common pooling operations are max-pooling (also known as maximum pooling) and mean-pooling (also known as average pooling). With suitable pooling operations, the amount of computation in the network is reduced and meanwhile, the overfitting issue is to some extent controlled. Normally, two or more fully connected layers are constituted after multiple blocks of convolution and pooling layers, the output of which can be deemed as the final deep representations of the original input image data.

Since the astonishing results obtained by CNNs on the well-known ImageNet dataset for object recognition task in 2012 [113], there has been a surge of interest in the potential of applying deep CNNs to produce meaning representations in the area of affective computing, particularly for facial representation extraction [172, 222] and for audio-based representation learning [101, 127]. For example, the authors of [52, 153] used CNNs to automatically extract salient representations from from facial images [153]. Similar work has also been done on speech spectrograms [104, 29].

In this thesis, two data-driven representation learning algorithms for emotion recognition will be discussed in Section 3.1.1 and Section 3.1.2, respectively.

2.5 Deep Learning for Emotion Prediction

Similarly to other applications where deep neural networks are employed, advances in emotion recognition have also benefited from deep learning techniques. In particular, beyond learning powerful feature representations with structures such as RNNs and CNNs, these deep structures can be exploited to model various emotion patterns from either engineered or data-driven input features. For instance, the implementation of RNNs can be found in numerous previous works, and has been frequently reported to achieve the best performance for emotion recognition [211, 234, 135, 30]. Particularly, in the AVEC challenge series from 2011 to 2018 [160], most of the champion systems were based on LSTM-RNN or GRU-RNN structures.

On top of that, recently, the combination of CNNs and LSTM-RNNs, firstly constructed in [196] in 2016, has been shown to be a promising approach for estimating dimensional emotions in an end-to-end manner. More specifically, a CNN subnet is applied as feature extractor with the raw signals being fed directly into it as the input. Then, the subsequent LSTM-RNN is formed to model long-term dependencies. In other words, the system jointly learns a feature learning task and an emotion inference task.

Moreover, other than building deep structures in an end-to-end fashion, the emphasis has also been given on exploring other deep structures for emotional behaviour analysis. Such attempts include, but are not limited to, Convolutional Re-

current Neural Networks (CRNNs) [124], highway networks or Residual Networks (ResNets) [107], Time-Delay Neural Networks (TDNNs) [133]. Moreover, additional learning algorithms and approaches have been investigated, such as adversarial learning, attention mechanisms, multi-task learning for emotion recognition. Particularly, in this thesis, several proposed learning strategies are closely associated with Multi-Task Learning (MTL), such as estimating the emotion and human agreement in parallel (cf. Section 3.2.2). In detail, MTL is a process of learning multiple tasks concurrently. Typically, there is one main task and one or more auxiliary tasks. By attempting to model the auxiliary tasks together with the main task, the model learns shared information across tasks, which may be beneficial to learning the main task. Mathematically, the objective function in MTL can be formatted as

$$\mathcal{J}(\theta) = \sum_{m=1}^{M} w_m L_m(\mathbf{x}, y_m; \theta_m), \tag{2.10}$$

where M denotes the number of tasks and $L_m(\cdot)$ represents the loss function of the task m, which is weighted by w_m. Also, θ_m represent the model parameters with respect to the specific task m.

Obviously, the performance of these deep structures usually heavily rely on large amounts of emotional data and considerable computational resources. Alternatively, pretrained deep networks have recently been used to learn useful representations for predicting emotions and yielded quite impressive results. Some example of such pretrained networks are AlexNet [113], FaceNet [172], GoogLeNet [191], SoundNet [8], and VGG-Face [148]. In addition, transfer learning strategy is normally requested to finetune these models for the task at hand. Alternatively, instead of manually design neural network structures by trial and error, this structure searching process in future can be facilitated by reinforcement learning, resulting in automated Machine Learning (autoML) for emotional behaviour analysis.

Another thing worthwhile to mention is that, when training these deep structures, a number of learning tricks are necessary to be considered, for instance, regularisation techniques such as early stopping, batch normalisation, dropout, and weight decay (L1 or L2 regularisation). For instance, batch normalisation is commonly applied after convolution layers to ensure that the data passing on to subsequent layers are normalised, by subtracting the batch mean and then dividing by the batch standard deviation. This process can increase the stability of a neural network, and consequently, fasten the training procedure of it. Furthermore, to overcome the overfitting problem, dropout is frequently used, by randomly dropping out units with a probability of q in the training phase. This thus can prevent the obtained model from co-adaptations too much on the training data. In general, various regularisation approaches can be combined and normally yield further performance improvement. For a full and detailed understanding of different regularisation techniques, the reader is referred to [71].

3

Contributed Methodology

In this chapter, the implementation of several deep learning-driven approaches will be discussed, aiming at tackling the corresponding challenges as aforementioned in Section 1.1. In particular, the chapter starts by presenting two techniques to address the representation learning challenge in Section 3.1, namely, a deep crossmodal latent representation learning approach and a deep bag-of-X-words approach. Thereafter, three frameworks and training strategies will be presented in Section 3.2 to tackle the deep modelling challenge with the power of deep learning. This will include strength modelling (Section 3.2.1), dynamic difficulty training (Section 3.2.2), and adversarial training (Section 3.2.3). Next, to deal with the lifelong learning challenge, a continual emotion recognition paradigm will be elaborated in Section 3.3. Finally, a novel autoencoder-based approach will be introduced in Section 3.4, aiming at dealing with the empathy emotion detection challenge.

3.1 From Hand-crafted to Data-driven Representations

Over the past few decades, massive efforts have been made to extract hand-crafted features that can capture relevant information for specific tasks in machine learning, e.g., MFCCs and SIFT features (cf. Section 2.2). In practice, nevertheless, appropriate and strong domain knowledge can hardly be attained to design a suitable feature set for a task of interest. To overcome this shortcoming, a large variety of deep learning-driven methods emerged recently, enabling to learn more generic representations directly from the raw data, such as CNNs and RNNs (cf. Section 2.4). In particular, learning meaningful and salient representations is an important task for emotion recognition where specific domain knowledge is rather vital.

To deal with the representation learning challenge, in this section, the description of a novel deep crossmodal latent representation learning approach will first be given

in Section 3.1.1. Then, it moves to introducing a deep bag-of-X-words approach with a particular focus on the generation of context-aware representations in Section 3.1.2.

3.1.1 Deep Crossmodal Latent Representations

Aiming at attaining shared representations between audio and video emotional data, a novel crossmodal emotion embedding framework called *EmoBed*, which aims to leverage the knowledge from other auxiliary modalities to improve the performance of emotion recognition, as presented by the author and her colleagues in [85, 87], will be introduced.

Albeit the notable advantages, during inference, most of these multimodal systems require the synchronous presence of all modalities that are employed in the training phase [154, 141, 83, 234, 3, 159]. This severely impedes their application in real life, since it is a common case that information from some particular modalities is missing. For example, a camera could be not always fixed in front of a user, or could not work in darkness, which results in invalid or missing visual signals. Likewise, a user could be silent although she/he is emotional, leading to the missing of audio data. The absence of any involved modalities often leads to corruption or performance degradation of a pre-trained multimodal system [223].

A straightforward solution to address this issue often makes use of the integration of an additional component, such as a voice activity detector and a face detector, in front of the multimodal recognition systems [223]. Once the absence of a particular modality is detected, the prediction process can be automatically redirected to another system that is trained via an accordingly reduced number of modalities. Nevertheless, such a system is normally inferior to the system with all modalities as aforementioned.

In contrast, to embrace the advantages and avoid the disadvantages of a multimodal system, the proposed EmoBed model is particularly innovated to enhance the performance of a monomodal emotion recognition system, by exploiting the emotion embeddings and sharing relevant information from heterogeneous data of auxiliary modalities in the training phase. Basically, it consists of two main processes, i. e., the *joint multiple modalities training* process and the *crossmodal training* process. The former process utilises the data from multiple modalities to jointly train the system, with an assumption that the knowledge from different modalities could be implicitly transferred to or fused by the system. Meanwhile, the later process takes a triplet constraint to minimise the distance of inter-class representations while maximising the inter-class ones, regardless of the modality constraints. Once the model has finished training, those auxiliary modalities are not required anymore in the evaluation phase.

As a result, the EmoBed approach differs from the training process for conventional monomodal systems that are trained merely with data from one single modality. It also differs from classic multimodal systems that often need the same

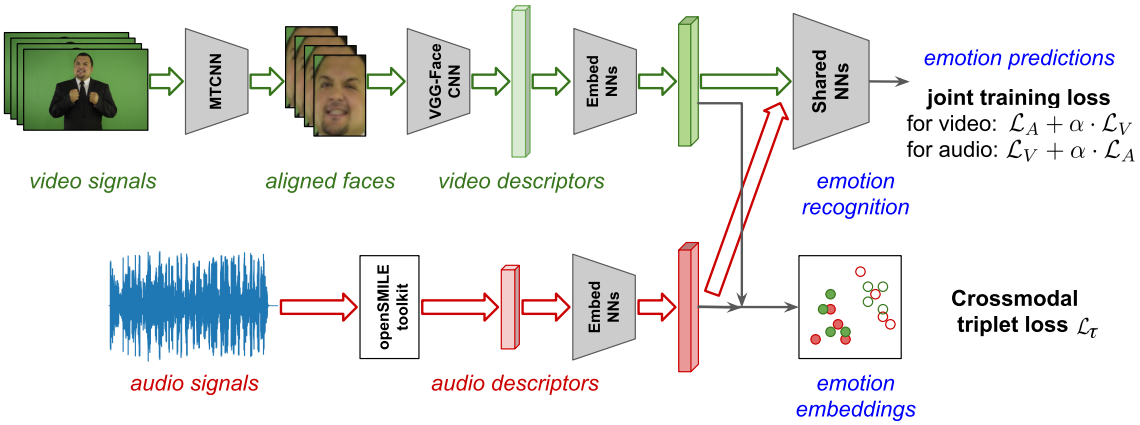


Figure 3.1: The proposed crossmodal Emotion emBedding (EmoBed) framework for monomodal emotion recognition [87].

modalities in both the training and evaluation stages. Moreover, another advantage of the proposed EmoBed is that it does not demand any data synchronisation across modalities in the training stage.

In the following, the proposed EmoBed framework will first be profiled in Section 3.1.1.1. Then, the two processes, i. e., joint multiple modalities training and crossmodal training, will be detailed in Section 3.1.1.2 and Section 3.1.1.3, respectively.

3.1.1.1 EmoBed Framework: System Overview

The framework of the proposed EmoBed approach is depicted in Figure 3.1. In particular, a shared embedding space is attained to explore the latent correlation between audio and video signals during training. Typically, after extracting audio and video descriptors via several standard and essential processing steps, two embedding networks will be jointly trained to project these multimodal descriptors into a common space, the representations of which can then be applied to predict emotions, under a monomodal scenario.

Mathematically, the two embedding networks can be expressed as two embedding functions f_A

$$f_A : \mathbb{R}^M \rightarrow \mathbb{R}^E, \tag{3.1}$$
$$\text{and } \mathbf{x}_A \mapsto \mathbf{e}, \tag{3.2}$$

and f_V

$$f_V : \mathbb{R}^N \rightarrow \mathbb{R}^E, \tag{3.3}$$
$$\text{and } \mathbf{x}_V \mapsto \mathbf{e}, \tag{3.4}$$

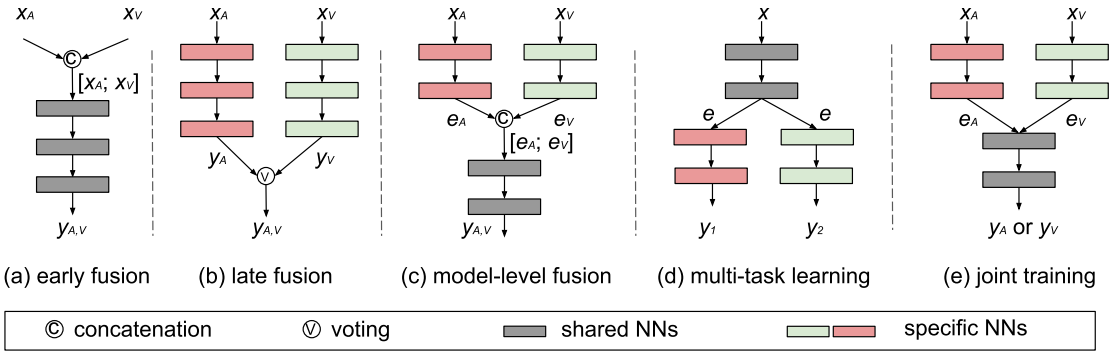


Figure 3.2: Structure comparison among the proposed joint audiovisual training (e), and other related multimodal learning frameworks (i. e., early fusion (a), late fusion (b), model-level fusion (c)), and multi-task learning (d) [87].

where $\mathbf{x}_A \in \mathbb{R}^M$ and $\mathbf{x}_V \in \mathbb{R}^N$ denote audio and visual inputs, respectively, and $\mathbf{e}$ is the corresponding embedded representations in a common coordinate space $\mathbb{R}^E$.

As can be seen in Figure 3.1, to learn such embedding functions in EmoBed, two main learning components are involved, i. e., a joint training and a crossmodal training. Both of them tend to explore the underlying semantic emotion information but with a shared recognition network or with a shared emotion embedding space, respectively. In doing this, the enhanced system trained in this way can make use of the complementary information from other modalities. Nevertheless, the presence of these auxiliary modalities is not demanded during inference. In the following, these two main components will be explained in more detail separately.

3.1.1.2 Joint Training with Audiovisual Data

Although related to other multimodal fusion techniques, the proposed joint training process is different from these methods. Therefore, before demonstrating how to learn a common embedding space with the joint training loss, let us first briefly differentiate the joint training from other structures, which was first presented by the author and her colleagues in [88], also as depicted in Figure 3.2.

Three conventional multimodal fusion paradigms are demonstrated in Figure 3.2(a)-(c). Note that, although information of multiple modalities is fused at different levels, they all contribute to multimodal emotion recognition systems. Concretely, given $\mathbf{x}_{(\cdot)}$, $\mathbf{e}_{(\cdot)}$, and $y_{(\cdot)}$ denoting the monomodal input feature, the learnt hidden-layer representation, and the output prediction for audio A and video V, respectively, the combination of audio and video knowledge is in forms of $[\mathbf{x}_A; \mathbf{x}_V]$ for feature-level fusion, weighted averaging based on y_A and y_V for decision-level fusion, or $[\mathbf{e}_A; \mathbf{e}_V]$ for model-level fusion. It should be noted that these models can be utilised, if and only if both $\mathbf{x}_A$ and $\mathbf{x}_V$ are available as inputs of the model, and there is no need for $\mathbf{e}_A$ and $\mathbf{e}_V$ to be of the same dimensions. In the proposed joint

training model, albeit the constraint of the existing of both $\mathbf{x}_A$ and $\mathbf{x}_V$ remains during training, the model can then be applied under a monomodal setting.

Besides, the joint training model is further varied from multi-task learning, which is illustrated in Figure 3.2(d). In multi-task learning, during the training phase, an auxiliary task benefits the main task by updating the parameters in the shared front-end feature-learning network. In contrast, in the proposed model, it is expected that inputs of an auxiliary modality can help improve the emotion prediction of the main modality, by optimising the parameters of the shared back-end predicting network.

Formally, when denoting an audio feature vector as $\mathbf{x}_A \in \mathbb{R}^M$ and its corresponding visual feature vector as $\mathbf{x}_V \in \mathbb{R}^N$, where M and N are the dimensions of the audio and visual vectors, respectively. As depicted in Figure 3.2(e), $\mathbf{x}_A$ and $\mathbf{x}_V$ are fed into two modality-specific embedding subnetworks, the process of which can be formulated as follows:

$$\mathbf{e}_A = f_A(\mathbf{x}_A),\ \mathbf{e}_V = f_V(\mathbf{x}_V), \tag{3.5}$$

where the function f_A and the function f_V map each input of different modalities into the same subspace, resulting in corresponding E-dimensional representations $\mathbf{e}_A$ and $\mathbf{e}_V$. After that, the following shared layers are applied to estimate the final predictions, and this process can be formulated as follows:

$$y_A = f(\mathbf{e}_A),\ y_V = f(\mathbf{e}_V), \tag{3.6}$$

where the function $f : \mathbb{R}^E \rightarrow \mathbb{R}$ estimates final predictions y_A and y_V, separately.

To aggregate the advantages of different modalities for monomodal emotion recognition (i. e., SER or FER), the model is trained with a set of audiovisual features $\{(\mathbf{x}_A, \mathbf{x}_V)\}$. When the model is applied for SER, the joint loss function $\mathcal{J}(\theta)$ is calculated by:

$$\mathcal{J}(\theta) = \mathcal{L}_A + \alpha \cdot \mathcal{L}_V, \tag{3.7}$$

where θ denotes the network parameters to be optimised, $\mathcal{L}_A$ and $\mathcal{L}_V$ stand for the loss of audio and video data, respectively, and α denotes the weight of the video prediction loss to regulate its contribution to $\mathcal{J}(\theta)$. The term $\alpha \cdot \mathcal{L}_V$ enforces the optimisation to take the auxiliary modality information into account. Likewise, for FER, the joint loss function in Equation (3.7) can be rewritten as

$$\mathcal{J}(\theta) = \mathcal{L}_V + \alpha \cdot \mathcal{L}_A. \tag{3.8}$$

Moreover, the value of α is optimised on the development set, by achieving the best performance for the selected modality.

3.1.1.3 Crossmodal Emotion Embedding

In the following, for crossmodal emotion embedding, a crossmodal triplet loss function is proposed to learn emotion-discriminative embeddings using crossmodal data.

In general, triplet loss forces to project the original descriptors into a latent space where instances with similar semantics are pulled together while instances with dissimilar semantics are pushed away. Consequently, the similarity of instances with the same semantic information is preserved in the learnt representations. Therefore, in this study, the aim is to exploit the semantic information across audio and video modalities. Before introducing the crossmodal triplet-loss-based embedding approach, let us first revisit the conventional monomodal triplet loss constraint for monomodal representation learning.

Monomodal Triplet Loss Function
Given an M-dimensional input feature $\mathbf{x} \in \mathbb{R}^M$, a DNN structure with multiple hidden layers is applied to generate a corresponding M-dimensional latent representation $\mathbf{e} \in \mathbb{R}^N$, i. e., the output from the representation layer of the network. Thus, it turns out that the effect of the network can be represented as a mapping function Φ and expressed as:

$$\Phi : \mathbb{R}^M \rightarrow \mathbb{R}^N, \tag{3.9}$$

$$\text{and } \mathbf{x} \mapsto \mathbf{e}. \tag{3.10}$$

That is, the network embeds the input $\mathbf{x}$ into a N-dimensional latent space. Particularly, in this learnt space, the semantic relationship between multiple instances should be preserved. For this purpose, during the training process, a set of *three instances* from the training set is used as a *triplet*, to enforce instances from the same class to be closer in this representation space and in the meanwhile to retain the different classes a larger distance. The overview of the latent discriminative representation learning scheme is given in Figure 3.3. In the figure, the DNN network is unfolded three times and placed in parallel for a better view and explanation.

Mathematically, given a set of triplets $\tau = \{\tau_i\}$ where $i = 1, ..., n$, each triplet τ_i (indexed by i) is an ordered set, composed of three distinct inputs and written in the following form:

$$\tau_i = (\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-), \tag{3.11}$$

where $\mathbf{x}_i$ and $\mathbf{x}_i^+$ (denoted as a *positive pair*) are from the same class, while $\mathbf{x}_i$ and $\mathbf{x}_i^-$ (denoted as a *negative pair*) belong to different categories. The target is to learn a mapping Φ to a latent representation space where $\mathbf{x}_i$ is more similar (or closer) to $\mathbf{x}_i^+$ than to $\mathbf{x}_i^-$ for all triplets. When processing instances within a selected τ_i, the latent feature representations $\mathbf{e}_i$, $\mathbf{e}_i^+$, and $\mathbf{e}_i^-$ can be obtained as

$$\mathbf{e}_i = \Phi(\mathbf{x}_i), \tag{3.12}$$

$$\mathbf{e}_i^+ = \Phi(\mathbf{x}_i^+), \tag{3.13}$$

$$\mathbf{e}_i^- = \Phi(\mathbf{x}_i^-). \tag{3.14}$$

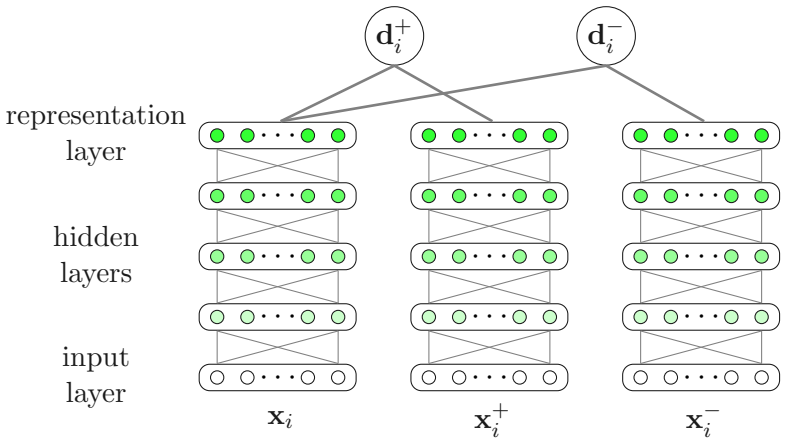


Figure 3.3: Framework for learning latent discriminative representations. For each training instance $\mathbf{x}_i$, $\mathbf{x}_i^+$ indicates a randomly selected instance within the same category as for $\mathbf{x}_i$; $\mathbf{x}_i^-$ indicates another randomly selected instance from a different category; $\mathbf{d}_i^+$ and $\mathbf{d}_i^-$ respectively denote the distances between the two latent representations learnt from the instances with the same or different categories.

Based on these, the distance of the positive pair d_i^+ and the distance of the negative pair d_i^- are computed as follows:

$$d_i^+ = \left\|\mathbf{e}_i - \mathbf{e}_i^+\right\|_2 = \left\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_i^+)\right\|_2, \tag{3.15}$$

$$d_i^- = \left\|\mathbf{e}_i - \mathbf{e}_i^-\right\|_2 = \left\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_i^-)\right\|_2, \tag{3.16}$$

where $\|\cdot\|_2$ denotes the Euclidean distance ($l2$-norm) between the two latent representations in a pair.

To learn a meaningful latent space, in the training phase, the network (parameterised by θ) is optimised to encourage instances with the same label to approach each other and instances with different labels to be apart from each other. Equivalently, the objective of the target problem can be deemed as decreasing d_i^+ and meanwhile increasing d_i^- over all triplets. For this purpose, the loss function to be minimised can be devised as

$$\mathcal{J}(\theta) = \gamma^{(d_i^+ - d_i^-)}, \tag{3.17}$$

where $(d_i^+ - d_i^-)$ indicates the discrepancy between the positive and negative pairs, γ is a predefined base of the exponential function ($\gamma > 1$ to ensure exponential growth), and θ denotes the trainable parameters of the network. Since $\mathcal{J}(\theta)$ is differentiable with respect to θ, the loss function in Equation (3.17) can be readily integrated into backpropagation in neural networks.

Note that, in contrast to taking merely the discrepancy $d = d_i^+ - d_i^-$ as the objective, in Equation (3.17) the exponential function is introduced. The underlying

rationale is that the aim is to learn more useful representations, by giving different triplets various emphasis mainly depending on the difficulty of distinguishing contents of the triplet in the learnt space. In other words, instead of paying equal attention to each triplet during training, attentions are of exponential growth with respect to the discrepancy. Mathematically, when there is a large discrepancy d, it results in an even larger gradient of $\mathcal{J}(\theta)$ to update the network; if d is small, the network updates its weights only slightly. Therefore, the emphasis is particularly placed to enforce the representation learning to facilitate difficult and ambiguous triplets, including cases when $\mathbf{e}_i^-$ is close to $\mathbf{e}_i$, when $\mathbf{e}_i^+$ is far from $\mathbf{e}_i$, or in cases when $\mathbf{e}_i$ is closer to $\mathbf{e}_i^-$ than it is to $\mathbf{e}_i^+$. As a result, the model learns to project the original features into a latent space where the intra-class feature distance is smaller comparing with the inter-class feature distance. Once the training stage is completed, a new instance $\mathbf{x}$ can be fed into the trained network to generate its corresponding latent representation $\mathbf{e}$ for further processing.

In addition, the learning scheme can be extended to further concern relative distance differences in a multi-class scenario where tuples of $k+1$ for k-class classification can be applied in place of triplets. That is, one positive pair and another $k-1$ negative pair of samples will be formed for each training sample. Then, Equation (3.17) needs to be adjusted to take all discrepancies into account. This will lead to an embedding space that may achieve even better classification performance compared with the triplet loss. However, when the number of the total categories is large, it may result in a high computational requirement. In this thesis, the analysis is focused on triplets for the sake of reducing the computational complexity.

Furthermore, it is worth noting that the proposed latent representation learning scheme can be applied to different DNN structures for specific tasks. In this thesis, for emotion recognition from audiovisual recordings, RNNs with LSTM cells are employed, as they have shown to yield good overall performance in predicting emotions [198, 30, 99, 90].

Crossmodal Triplet Loss in Crossmodal

Comparing with the monomodal triplet loss function, the triplet loss in the EmoBed framework is a crossmodal variant, aiming to supervise the crossmodal learning process.

In particular, to compute the crossmodal triplet loss $\mathcal{L}_\tau$, the audio embeddings $\mathbf{e}_A$ and the video embeddings $\mathbf{e}_V$, are jointed to form a double-sized batch of embeddings in the form of $\{\mathbf{e}_A; \mathbf{e}_V\}$. Then, a pairwise Euclidean distance matrix is obtained by computing the distance between all paired embeddings. Afterwards, for each embedding (either audio or video), another two embeddings are chosen from the same batch, to form a hard triplet. It is worth mentioning that, when generating the hardest positive or negative pair, both the intermodal and intramodal similarities are taken into consideration. That is, for a given triplet $\tau_i = (\mathbf{e}_i, \mathbf{e}_i^+, \mathbf{e}_i^-)$ with $i = 1, ..., n$, the positive (or negative) embedding $\mathbf{e}_i^+$ (or $\mathbf{e}_i^-$) could be either an audio-

based embedding or a visual one. In this manner, the learning process enforces the model to narrow the distribution gap of embeddings from different modalities, and to keep the specific emotional semantics intact in the meantime.

Supervised by the crossmodal triplet loss $\mathcal{L}_\tau$, the model is forced to minimise the optimisation objective $\mathcal{J}(\theta)$, which can be formulated as:

$$\mathcal{J}(\theta) = \mathcal{L}_{mon} + \beta \cdot \mathcal{L}_\tau, \tag{3.18}$$

where $\mathcal{L}_{mon}$ denotes the conventional monomodal discriminative loss, i. e., $\mathcal{L}_A$ for speech or $\mathcal{L}_V$ for video.

As a consequence, the overall training process of the proposed EmoBed framework is achieved, by integrating the triplet constraint into the joint training approach (cf. Section 3.1.1.2), as displayed in Figure 3.1. Generally, after extracting monomodal descriptors from standard pre-processing procedures, embedding functions f_A and f_V are estimated by two embedded neural networks, respectively, which project audio and video descriptors into a common latent space. Subsequently, the audio and visual embeddings are fed into a shared emotion recognition neural network, which is trained via a joint training loss. Concurrently, the training process is supervised by the triplet loss of the audio and visual embeddings.

Mathematically, when applying the EmoBed framework for audio emotion recognition, the objective function can be formatted as:

$$\mathcal{J}(\theta) = \mathcal{L}_A + \alpha \cdot \mathcal{L}_V + \beta \cdot \mathcal{L}_\tau + \lambda \cdot \mathcal{R}(\theta), \tag{3.19}$$

where $\mathcal{L}_A$ and $\mathcal{L}_V$ represent the discriminative loss function of audio and visual data, respectively, while $\mathcal{L}_\tau$ represents the triplet loss function of both audio and visual data. Moreover, the hyperparameters α and β are introduced to weight the contribution of the video data and the triplet loss. Furthermore, λ is applied to control the importance of the regularisation term $\mathcal{R}(\theta)$. Similarly, when training the EmoBed framework for facial emotion recognition, the objective function in Equation (3.19) can be modified by exchanging $\mathcal{L}_A$ and $\mathcal{L}_V$, i. e.,

$$\mathcal{J}(\theta) = \mathcal{L}_V + \alpha \cdot \mathcal{L}_A + \beta \cdot \mathcal{L}_\tau + \lambda \cdot \mathcal{R}(\theta). \tag{3.20}$$

After the model has been trained, the components associated with the auxiliary modality can be discarded, while the rest is retained and utilised to recognise emotional behaviours in a specific modality. It is also expected that the approach could provide a latent discriminative representation space to ameliorate the recognition performance.

3.1.2 Deep Bag-of-X-Words

In Section 3.1.1, the latent discriminative representation learning approach has been introduced where the semantic similarities among instances are exploited under either monomodal or multimodal conditions. However, with this method, the context

information is not yet taken into consideration when the representations are learnt. In the following, let us now move on to another representation learning method first proposed by the author and her colleagues [92], where context information in successive samples is utilised to produce meaningful and robust representations.

Though neural network-based representation learning performs well for emotion recognition [196, 105], the learnt representations are hard to interpret or understand. In contrast, another emerging approach, Bag-of-Audio-Words (BoAW), has been proposed for SER, to estimate a segment-level (or high-level) representation based, e.g., on MFCCs and log-energy as frame-level (or low-level) feature vectors [170]. In [170], these low-level features are quantised, and histograms are computed with a random-selected codebook as final representations which give one of the best recognition performances on the popular spontaneous emotional dataset RECOLA [164]. Moreover, BoAW has been applied successfully to several other paralinguistic information retrieval tasks, such as sound event classification [119], music genre classification [219], and copy detection [122]. While BoAW has produced meaningful and robust representations for SER, it does not take context information into consideration when creating these representations; since emotional content is involved in multiple coherent frames, context information is vital and needs to be dealt with care.

Contrary to the conventional BoAW approach, in this section, an approach to generate Bag-of-Context-Aware-Words (BoCAW) representations will be discussed. To address the second challenge as given in Section 1.1, a hierarchical architecture is applied to preserve the context information while learning the representations. More specifically, BoAW is applied twice but within different temporal scales; a small local window containing a number of context frames is first utilised, and then a global analysis window containing all frames of one instance is explored.

Such a hierarchical structure is conceptually similar to a Deep Belief Net (DBN), where features with various granularities can be extracted from each layer of the DBN [114]. In addition, BoCAW is further related to Dual-Layer Bag-of-Frames (DLBoF) proposed in [219]. The DLBoF framework attempts to model a piece of music with a two-layer structure, where frame-level characteristics and segment-level semantics can be captured and integrated together for music information retrieval tasks. In BoCAW, yet, only the segment-level features from the second layer are used. In this manner, the BoCAW approach bridges the gap between short-term frame-based features and long-term emotional segments by introducing mid-level words with context information, so as to enhance the regular BoAW approach.

In the following, the conventional BoAW method will first be introduced in Section 3.1.2.1. Then, the proposed BoCAW technique will be presented in Section 3.1.2.2.

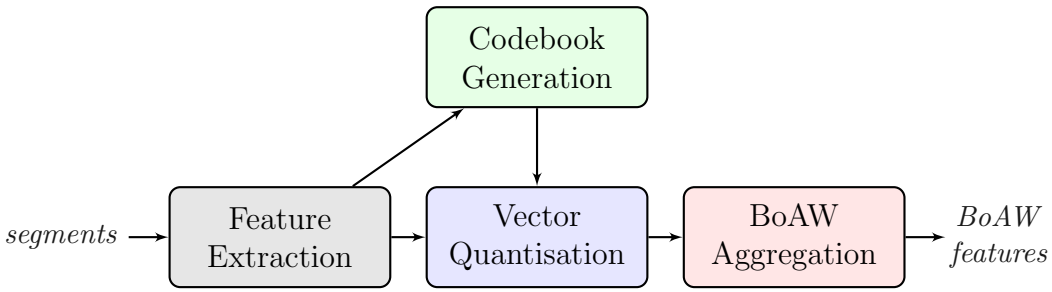


Figure 3.4: Diagram of the Bag-of-Audio-Words approach pipeline.

3.1.2.1 Bag of Audio Words

Let us denote a frame-level feature vector as $\mathbf{x}_i \in \mathbb{R}^M$ such that the index $i = 1, ..., n$, where n is the total number of frames, and M is the dimension of the vector. Thus, given one audio segment which is composed of n frames, it can be expressed as a set of features, i. e., $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. Next, once the series of $\mathcal{X}$ is extracted from all segments in the entire training set, the traditional BoAW approach can be carried out by the following three steps, as the one displayed in Figure 3.4, adapted from [92].

Codebook Generation: a codebook $\mathcal{C}$ is a set of codewords $\mathbf{c}$ learnt from the feature space $\mathcal{X}$, and $\mathcal{X}$ is obtained by extracting frame-level features for all training segments. Therefore, the codebook generation problem can be formulated as:

$$\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{K}, \mathbf{c}_k \in \mathbb{R}^M, \tag{3.21}$$

where $\mathbf{c}_k$ denotes the k-th codeword, and in total K codewords form the codebook $\mathcal{C} \in \mathbb{R}^{M \times K}$.

Normally, $\mathcal{C}$ can be built by a clustering algorithm such as k-means. In addition, random sampling has also been proposed in [158] and utilised with success for SER [170]. In this thesis, random sampling is adopted, which is much faster than k-means but delivers comparative performances in the meanwhile.

Vector Quantisation: once the codebook $\mathcal{C}$ has been generated, each frame-level feature vector $\mathbf{x}_i$ can be assigned to its closest (Euclidean distance) codeword $\mathbf{c}_k$ within $\mathcal{C}$, and be encoded as the corresponding index k. This process is referred to as the vector quantisation step. In detail, it can be formulated as a mapping function $\Phi : \mathbb{R}^M \rightarrow \mathbb{R}^K$, encoding each feature $\mathbf{x}_i$ into the codebook space. This leads to a corresponding K-dimension feature $\phi_i \in \mathbb{R}^K$, and its k-th coefficient $\phi_{i,k}$ is defined as follows:

$$\phi_{i,k} = \begin{cases} 1, & \text{if } k = \text{argmin}_k \|\mathbf{x}_i - \mathbf{c}_k\|_2 \\ 0, & \text{otherwise} \end{cases}, \tag{3.22}$$

where $k = 1, ..., K$, and $\|\cdot\|_2$ denotes the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{c}_k$.

However, it has to be noted that during the quantisation step, there might be the case that one feature vector is nearly equidistant to several codewords, and therefore single assignment is ambiguous. Hence, instead of choosing the nearest codeword, it is also possible to assign a vector to a certain number n_a of closest codewords. This variant can be referred to as multiple assignments. In this regard, Equation (3.22) can be reformulated as:

$$\phi_{i,k} = \begin{cases} 1, & \text{if } \mathbf{c}_k \in \mathcal{N}(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases}, \tag{3.23}$$

where $\mathcal{N}(\mathbf{x}_i) \subset \mathcal{C}$ is a set of n_a nearest codewords with respect to a given $\mathbf{x}_i$. In this thesis, the multiple assignments strategy is utilised as it has proven to perform better than a single assignment [170].

BoAW Aggregation: given an audio segment S spanning n_s frames, a 'bag' can then be created by simply computing a histogram of the codewords. More specifically, a histogram representation $\mathbf{h}_S$ is formed to describe the distribution of the features, i. e., how and to what extent each codeword has contributed to represent S. This process is referred to as the BoAW aggregation step, and can be formulated as $\mathbf{h}_S = g(\{\phi_i\})$, where a pooling function $g : \mathbb{R}^{K \times n_s} \rightarrow \mathbb{R}^K$ aggregates occurrences of each codeword represented by ϕ_i in all n_s frames for a given segment S. Thus, its k-th component $h_{S,k}$ can be computed as:

$$h_{S,k} = \sum_{i=1}^{n_s} \phi_{i,k}. \tag{3.24}$$

At this point, all frame-level features in S are encoded into a segment-level representation $\mathbf{h}_S$, which can be utilised for further processing, e. g., feeding it into a classifier directly for emotion prediction.

3.1.2.2 Bag of Context-Aware Words

In the conventional BoAW approach (cf. Section 3.1.2.1), the histogram representation $\mathbf{h}_S$ covering the entire instance (i. e., a very long segment) is the final high-level representation, which can be exploited for audio classification or regression tasks.

In the proposed BoCAW method, however, a set of the histogram representations $\{\mathbf{h}_{Si}\}$, with $i = 1, ..., n$, can be generated from n much shorter segments as mid-level representations first, and then applied to the BoAW approach a second time to form the final high-level representation. The concern is that typical emotion patterns may exist among a sequence of several coherent frames, and therefore features generated based on shorter segments rather than frames alone may perform better for the emotion prediction task.

The framework of the proposed BoCAW representation generation approach is depicted in Figure 3.5, consisting of two stages in a hierarchical structure. In the first stage, sub-bag (or mid-level) features are generated by applying the conventional

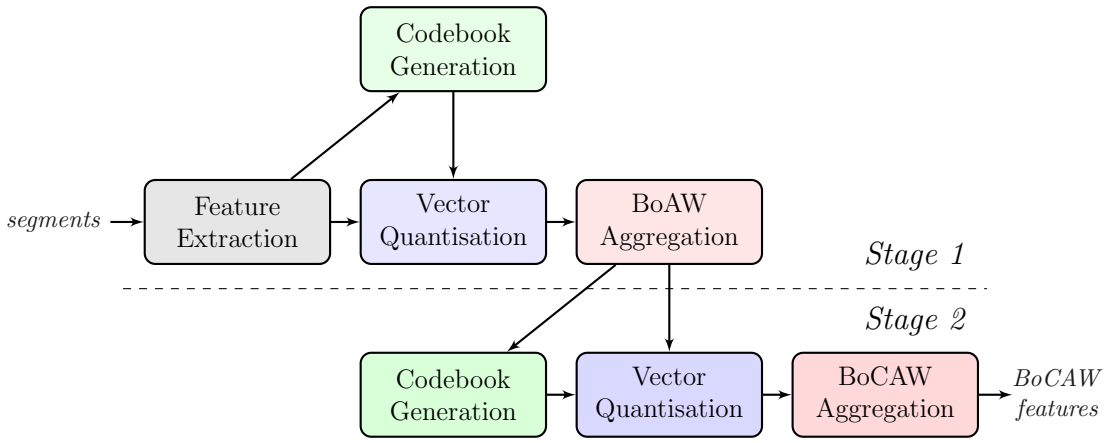


Figure 3.5: Diagram of the Bag-of-Context-Aware-Words approach pipeline.

BoAW approach to the input frame-level features within a sliding window. Note that, each sliding window contains several successive frames of input features and can be much smaller than the total length of an instance to be analysed. These BoAW features obtained from the first stage contain context information. Therefore, let us refer to these features as *context-aware words*. After that, in the second stage, these context-aware BoAW representations are further utilised to generate a final context-aware bag for each whole segment. Herein, the BoAW paradigm is again applied as in the first stage but within a global window, the length of which equals to the total length of the segment to be analysed. The whole BoCAW framework can be implemented with an open-source toolkit OPENXBOW [171], and the details of each system component are described below.

As illustrated in Figure 3.5 (adapted from [92]), in Stage 1, each given long-term segment is first sliced into n short-term windows. Herein, let us denote the i-th window as S_i with $i = 1, ..., n$, while each S_i consists of N_1 successive frames. Then, the first-stage codebook set $\mathcal{C}_1$ with K_1 codewords is constructed, and thereafter a set of BoAW representations $\{\mathbf{h}_{Si}\}$ with each $\mathbf{h}_{Si} \in \mathbb{R}^{K_1}$ is for each short-term window after assigning frames in the given window to their nearest codewords accordingly. That is, S is now represented by a set of context-aware mid-level words $\{\mathbf{h}_{Si}\}$. After that, in Stage 2, similar to the process in Stage 1, a second-stage codebook $\mathcal{C}_2$ is built, by selecting K_2 codewords over all $\{\mathbf{h}_{Si}\}$ from training samples. Subsequently, additional vector quantisation with multiple assignments and histogram aggregation processes are conducted similar as in Section 3.1.2.1. In the end, a final BoCAW representation $\mathbf{h}_S \in \mathbb{R}^{K_2}$ is gained, by counting the occurrences of corresponding second-stage codewords for all segments in it. This process is referred to as the BoCAW aggregation step.

From this, segments of variable lengths can be encoded into BoCAW representations with an equal and fixed length, and in the meanwhile short-term temporal information is preserved in them. Still, it is noteworthy that the length of the sliced

window in Stage 1 needs to be defined decently, in order to envelop a moderate amount of context information when generating the context-aware words.

3.2 From Shallow to Deep Modelling

In the previous section, the primary consideration is to learn meaningful and useful emotional representations by leveraging the power of deep learning. Additionally, beyond representation learning, deep learning-driven methods can bring benefit to formulating prediction models for different emotion classification and regression applications as well. To this end, various model structures based on deep learning will be presented in the present section, to facilitate the analysis of emotional behaviours in the wild with deep modelling.

3.2.1 Strength Modelling

In this section, a novel framework, *Strength Modelling*, will be introduced, as proposed by the author and her colleagues in [83].

In the field of affective computing, a large body of literature exists on exploring various modelling techniques for audiovisual emotion recognition [141, 163, 194]. However, when comparing the advantages offered by different models, no clear observations can be drawn as to the superiority of any of them. For instance, the work in [141] compared the performance of Support Vector Machine for Regression (SVR) and Bidirectional LSTM-RNNs (BLSTM-RNNs) on the Sensitive Artificial Listener database, and the results indicate that the latter performed better on a reduced set of 15 acoustic frame-level features. However, the opposite conclusion was drawn in [194], where SVR was shown to be superior to LSTM-RNNs on the same database with segment-level features. Other results in the literature confirm this inconsistent performance observation between SVR and diverse neural networks like (B)LSTM-RNNs and Feed-forward Neural Networks (FNNs) [163]. A possible rationale behind this is the fact that each prediction model has its pros and cons. For example, SVRs cannot explicitly model contextual dependencies, whereas LSTM-RNNs are highly sensitive to overfitting. To deal with this issue, the majority of previous studies have tended to explore the advantages (strength) of these models independently or in conventional early or late fusion strategies. However, recent results indicate that there may be significant benefits in fusing two or more, models in a hierarchical or an ordered manner [100, 126, 142].

Motivated by these initial promising results and targeting on the third challenge as mentioned in Section 1.1, a *Strength Modelling* approach is investigated, aiming at concurrently reaping the benefits of distinct learning models. In Strength Modelling, several (i. e., two in this thesis) distinct models are concatenated in a hierarchical framework. By that, the strength information of the first model, as represented by

its predictions, is joined with the original features, and this expanded feature space is then utilised as an input by the successive model.

Prior research that is of great relevance in this context is the Output Associative Relevance Vector Machine (OA-RVM) regression framework originally proposed in [142]. The OA-RVM framework attempts to incorporate the contextual relationships that exist within and between different affective dimensions and various multimodal feature spaces, by training a secondary RVM with an initial set of multi-dimensional output predictions (learnt using any prediction scheme) concatenated with the original input features spaces. Additionally, the RVM framework also attempts to capture the temporal dynamics by employing a sliding window framework that incorporates both past and future initial outputs into the new feature space. Results presented in [100] indicate that the OA-RVM framework is better suited to affect recognition problems than both conventional early and late fusion. Recently the OA-RVM model was extended in [126] to be multivariate, i.e., predicting multiple continuous output variables simultaneously.

All of these OA-RVM systems, like Strength Modelling, take original input features and output predictions into consideration to train a subsequent regression model to perform the final affective predictions. However, the strength of the OA-RVM framework is that it is underpinned by the RVM. Results in [100] indicate that the framework is not as successful when either an SVR or an SLR is used as the secondary model. Further, the OA-RVM is non-casual and requires careful tuning to find suitable window lengths in which to combine the initial outputs. This takes considerable time and effort. The proposed Strength Modelling framework, however, is designed to work with any combination of learning paradigms, which uses the initial set of predictions to help improve the accuracy of any subsequent model. Furthermore, Strength Modelling is casual. It combines the original input features and predictions on a frame-by-frame basis, which is a strong advantage over the OA-RVM in terms of deployment in real-time scenarios.

Furthermore, although it is investigated to hierarchically explore the strength of different machine learning algorithms under a monomodal setting (cf. Section 3.2.1.1), a major advantage of Strength Modelling is that it can work together with the conventional feature- and decision-level fusion strategies for multimodal affect recognition (cf. Section 3.2.1.2).

3.2.1.1 Strength Modelling in Monomodal System

The proposed Strength Modelling framework for emotion regression is depicted in Figure 3.6 (adapted from [83]). Given a M-dimensional input feature vector $\mathbf{x}_t \in \mathbb{R}^M$ at time t, and let $f_1 : \mathbb{R}^M \rightarrow \mathbb{R}$ be the function of the first model $Model_1$, it holds that the output of $Model_1$ with respect to $\mathbf{x}_t$ can be expressed as:

$$\hat{y}_t = f_1(\mathbf{x}_t). \tag{3.25}$$

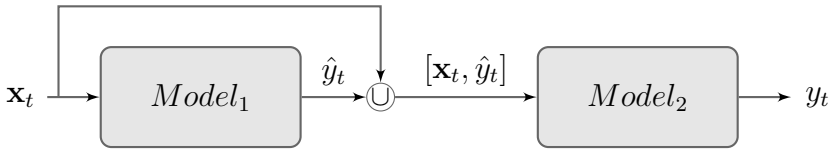


Figure 3.6: Overview of the Strength Modelling framework.

Then, $\hat{y}_t$ is concatenated with $\mathbf{x}_t$ frame-wise as the input of the second model ($Model_2$) to learn the expected prediction y_t. That is, let the function of $Model_2$ be $f_2 : \mathbb{R}^{(M+1)} \rightarrow \mathbb{R}$, y_t can be predicted in the following form:

$$y_t = f_2([\mathbf{x}_t, \hat{y}_t]) = f_2([\mathbf{x}_t, f_1(\mathbf{x}_t)]). \tag{3.26}$$

To implement the Strength Modelling for these suitable combinations of individual models, $Model_1$ and $Model_2$ are trained concurrently. In other words, $Model_2$ takes the predictive ability of $Model_1$ into account for training. The procedure is given as follows:

- first, $Model_1$ is trained with $\mathbf{x}_t$ to obtain the prediction $\hat{y}_t$.
- then, $Model_2$ is trained with $[\mathbf{x}_t, \hat{y}_t]$ to learn the expected prediction y_t.

Whilst the framework could work with any arbitrary modelling technique, here two commonly used ones, in the context of affect recognition, have been selected for initial investigations. These two models are the SVR and BLSTM-RNNs which are briefly discussed below.

SVR is extended from Support Vector Machine (SVM) to solve regression problems. It was first introduced in [51] and is one of the most dominant methods in the context of machine learning, particularly in emotion recognition [26, 163]. One of the most important advantages of SVR is the convex optimisation function, the characteristics of which gives the benefit that the global optimal solution can be obtained. Moreover, SVR is learnt by minimising an upper bound on the expected risk, as opposed to the neural networks trained by minimising the errors on all training data, which equips SVR a superior ability to generalise [80]. For a more in-depth explanation of the SVR paradigm, the reader is referred to [51].

The other model utilised in this study is BLSTM-RNN which has been successfully applied to continuous emotion prediction [159] as well as for other regression tasks, such as speech dereverberation [236] and non-linguistic vocalisations classification [151]. In general, it is composed of one input layer, one or multiple hidden layers, and one output layer [98]. The bidirectional hidden layers separately process the input sequences in a forward and a backward order and connect to the same output layer which fuses them. Such a structure grants BLSTM-RNN to learn past and future context in both short and long range. More details of this model can

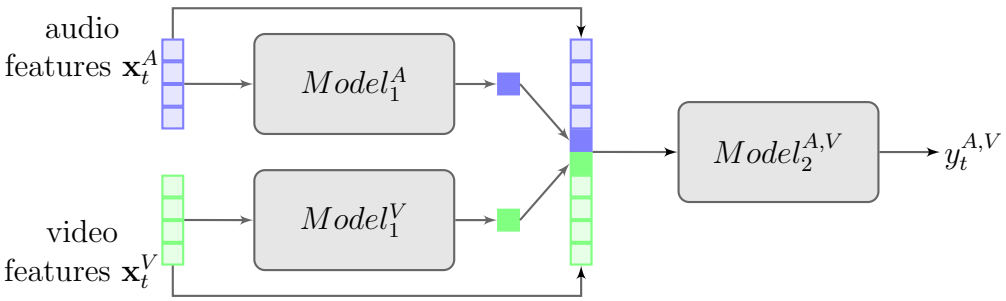


Figure 3.7: Strength Modelling with early fusion strategy.

be found in Section 2.3. It is worth noting that these two paradigms bring distinct sets of advantages and disadvantages to the framework: On the one hand, the SVR model is more likely to achieve the global optimal solution, but it is not context-sensitive [141]; on the other hand, the BLSTM-RNN model is easily trapped in a local minimum which can be hardly avoided and has a risk of over-fitting [74], while it is good at capturing the correlation between the past and the future information [141].

Based on the above considerations, $Model_1$ and $Model_2$ in Figure 3.6 could be either an SVR model or a BLSTM-RNN model, resulting in four possible permutations, i.e., SVR-SVR (denoted as S-S), SVR-BLSTM (as S-B), BLSTM-SVR (as B-S), BLSTM-BLSTM (as B-B). Note that, the B-B structure can be deemed as a variation of the neural networks in a deep structure. In addition, the S-S structure will not be considered; SVR training is achieved by solving a large margin separator, therefore it is unlikely to get any advantage in concatenating a set of SVR predictions with its feature space for subsequent SVR based regression analysis.

3.2.1.2 Strength Modelling for Multimodal System

For audiovisual emotion regression tasks, the Strength Modelling framework can further be implemented in both early and late fusion strategies.

Traditional early fusion combines multiple feature spaces into one single set. When integrating Strenght Momelling with early fusion, the initial predictions gained from models trained on the different feature sets are also concatenated to form a new feature vector. The new feature vector is then used as the basis for the final regression analysis via a subsequent model, as illustrated in Figure 3.7.

Mathematically, let $\mathbf{x}_t^A \in \mathbb{R}^M$ and $\mathbf{x}_t^V \in \mathbb{R}^N$ be an M-dimensional acoustic feature vector and an N-dimensional visual feature vector at time t, while $Model_1^A$ and $Model_1^V$ denote the first model for audio and video, respectively. Instead of the conventional early fusion, the input of the second model $\mathbf{i}_t$ in Strength Modelling can be expressed as

$$\mathbf{i}_t = [\mathbf{x}_t^A, \mathbf{x}_t^V, f_1^A(\mathbf{x}_t^A), f_1^V(\mathbf{x}_t^V)], \tag{3.27}$$

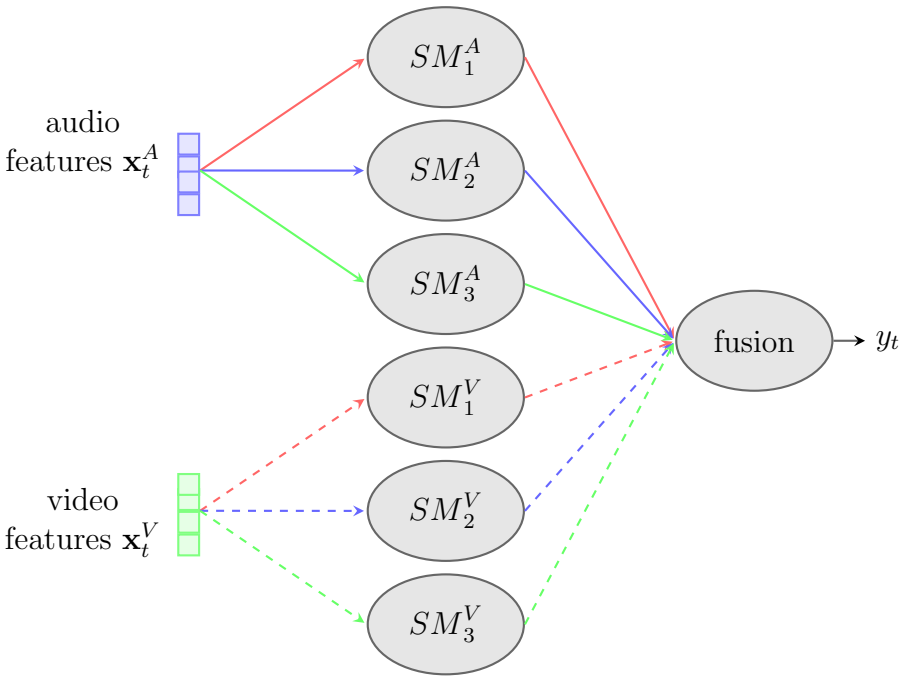


Figure 3.8: Strength Modelling (SM) with late fusion. Fused predictions are from multiple independent modalities with the same model (denoted by the red, green, or blue lines), multiple distinct models within the same modality (denoted by the solid or dashed lines), or the combination.

where f_1^A and f_1^V indicate the function of $Model_1^A$ and $Model_1^V$ to obtain the predictions accordingly. Thereafter, through the following audiovisual predicting model $Model_2^{A,V}$, the final prediction $y_t^{A,V}$ can be computed as

$$y_t^{A,V} = f_2^{A,V}(\mathbf{i}_t) = f_2^{A,V}([\mathbf{x}_t^A, \mathbf{x}_t^V, f_1^A(\mathbf{x}_t^A), f_1^V(\mathbf{x}_t^V)]), \tag{3.28}$$

where $f_2^{A,V}$ denotes the corresponding function of $Model_2^{A,V}$, holding that $f_2^{A,V} : \mathbb{R}^{M+N+2} \rightarrow \mathbb{R}^1$.

Likewise, Strength Modelling can be integrated with late fusion as well, using three different approaches, i.e., modality-based, model-based, and modality- and model-based, as demonstrated in Figure 3.8 (adapted from [83]). *Modality*-based fusion combines the decisions from multiple independent modalities with the same regression model; whilst *model*-based approach fuses the decisions from multiple different models within the same modality; and *modality- and model*-based approach is the combination of the above two approaches, regardless of which modality or model is employed. For all three techniques the fusion weights are learnt using a linear regressor:

$$y_t = \epsilon + \sum_{i=1}^{n} w_i \cdot y_{i,t}, \tag{3.29}$$

where $y_{i,t}$ denotes the original prediction at time t of the model i with $i = 1, ..., n$; ϵ and w_i are the bias and weights estimated on the development partition; and y_t is the final prediction.

3.2.2 Dynamic Difficulty Awareness Training

Having introduced the Strength Modelling framework in the previous section, where the goal is to concatenate different models to leverage the individual *strengths*, the focus of this section, in contrast, will now be laid on exploiting the *weakness* of a model in the learning process, when dealing with emotion prediction tasks.

Recently, some research has found that emotional training data can be practically learnt in different degrees [232, 78]. That is, some data can be easily learnt given a specific model, whilst some data are relatively tough. In this light, some promising approaches have been proposed in machine learning to optimise the learning procedure. For example, the most conventional approach is associated with boosting [120, 156], which dynamically updates the weights of those samples that are hard to be recognised or are even falsely recognised. Additionally, a more recent and promising approach refers to curriculum learning, which was firstly introduced in [16]. Curriculum learning presents the data from easy to hard during the training process so that the model can better avoid being caught in local minima in the presence of non-convex training criteria. Curriculum learning has become even more popular with the advance of deep learning. For emotion prediction, a handful of related studies have been reported very recently [78, 20, 123], which have shown the efficiency of curriculum learning.

However, one of the major disadvantages of these approaches is their unfriendliness to sequence-based pattern recognition tasks, such as continuous emotion prediction. That is, in the learning process, the samples, whether or not they were presented within a sequence, are considered individually and independently. The ignored context information, nevertheless, indeed plays a vital role in sequence-based pattern recognition [73]. To this end, in this section, a novel learning framework, *Dynamic Difficulty Awareness Training* (DDAT), for time-continuous emotion prediction will be presented, as investigated by the author and her colleagues in [234, 90]. In contrast to the previous approaches, such as the aforementioned boosting and curriculum learning, the proposed DDAT can be well integrated into conventional context-sensitive models (e. g., LSTM-RNNs), enabling the models to ultimately exploit the context information.

The underlying assumption of DDAT is that a model is expected to be able to deliver better performance if the model knows explicitly the learning difficulty of the samples along with time. This assumption is in line with the finding that humans normally pay more attention to the tasks that are inherently difficult so as to perform better [155, 205].

The crux in this context is how to present the learning difficulty information in an appropriate form such that the model can benefit from it. In this study, two strategies will be considered, namely, the *Reconstruction Error* (RE) of the input feature vectors or the *Perception Uncertainty* (PU) level of emotions. As a consequence, a typical DDAT framework consists of two stages: *information retrieval* and *information exploitation*. In the first stage, RE or PU is estimated as the difficulty indicator of learning specific information. Then, in the second stage, the obtained difficulty level is utilised in tandem with original features to update the model, such that it endows the models with a difficulty learning awareness. This process is also partially inspired by the awareness techniques proposed for robust speech recognition [184, 103], where the noise types are considered to be auxiliary information for acoustic modelling.

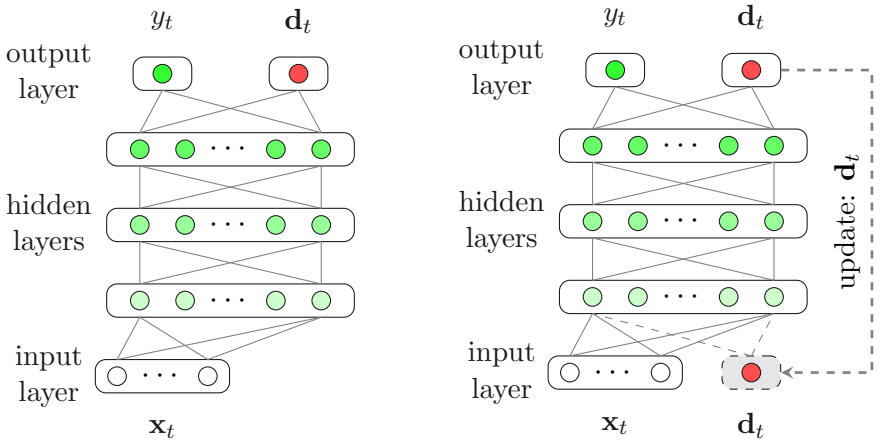
In the following, the main reasons for embracing PU and RE as difficulty indicators are discussed briefly in Section 3.2.2.1, before presenting the DDAT framework in Section 3.2.2.2. After that, the two main stages of the DDAT structure are described in more detail, i. e., the difficulty information retrieval stage in Section 3.2.2.3 and the difficulty information exploration stage in Section 3.2.2.4.

3.2.2.1 Difficulty Indicators: RE and PU

To implement the DDAT framework, the first difficulty indicator that is considered is the RE. In deep learning, RE normally serves as an objective function of an auto-encoder (AE) when extracting high-level representations. A well-designed AE is considered to reconstruct well the input from its learnt high-level representations [200]. Recently, the RE has also been exploited for other tasks, such as anomaly detection [129, 217] and classification [151]. For anomaly detection, an AE is trained on normal samples beforehand to serve as a novel event detector. When feeding a new sample into the AE, the obtained RE compared with a predefined threshold decides whether it is abnormal [129, 217]. For classification, several class-specific AEs are pre-trained separately. When feeding an unknown sample into these AEs simultaneously, the values of the corresponding RE are then interpreted as indicators of class membership [151].

Notably, all these works hypothesise that data with the same label have similar data distributions. That is, the mismatched data potentially result in higher REs than those of the matched data. This motivates the current study to employ the RE as a learning difficulty index due to the reason that, in learning process mismatched data severely promote the complexity of modelling [230]. Moreover, RE has been investigated for emotion prediction in speech with success in a previous tentative study in [90] by the author and her colleagues.

In addition, another choice of the learning difficulty indicator is the PU, when taking the subjective property of the task of interest, i. e., emotional behaviour analysis. PU is a term employed in subjective pattern recognition tasks to refer to the



(a) Difficulty information retrieving (b) Difficulty information exploiting

Figure 3.9: Dynamic Difficulty Awareness Training Frameworks with two stages. Difficulty information can be indicated by either the input reconstruction error (i. e., an error vector or the sum of all errors), or the emotion perception uncertainties. Figures are adapted from [234].

inter-annotator disagreement level when calculating a gold standard in an annotation process [48]. For emotion prediction, it has been frequently determined that the PU has a high correlation with the learning difficulty of a recognition model. For example, the reported work in [45] and [40] found that the emotion prediction systems perform better in low-uncertainty regions than in high-uncertainty regions. Likewise, the findings in [232] showed that the elimination of the samples labelled with a high uncertainty from the training set leads to a better emotion prediction model. This finding provokes this direction towards applying the PU as another learning difficulty index. It is also worth noting that the principle of PU-based strategy constrains its application to subjective pattern recognition tasks. Despite the fact that the concept of 'uncertainty' was employed in previous emotion prediction work, it was calculated among multiple predictions from variable systems [41], or merely utilised under a multi-task learning setting [91].

3.2.2.2 DDAT Framework: System Overview

In the following, let us first outline the framework of the proposed DDAT approach. Let $\mathbf{x}_t \in \mathbb{R}^M$ be an M-dimensional feature vector as the input at time t, $y_t \in \mathbb{R}$ be the corresponding label, and $\mathbf{d}_t$ denote as the difficulty indicator. Then, the DDAT framework consists of two stages, i. e., difficulty information retrieving stage, and difficulty information exploiting stage, as illustrated in Figure 3.9.

More specifically, the first stage, i. e., the difficulty information retrieving stage, is depicted in Figure 3.9 (a). In this stage, in order to extract and indicate the information related to the difficulty of the learning process, two different strategies are proposed, namely, the RE-based (or ontology-drive) strategy and the PU-based (or content-driven) strategy (cf. Section 3.2.2.3). The *ontology-driven* strategy focuses on the model itself. Specifically, the difficulty of the task is determined through the reconstruction of the input information, assuming that the RE is a proxy for its learning capability in a given moment. On the contrary, the *content-driven* strategy focuses on the data and assumes that different data can be learnt to different degrees. That is, some data can be easily learnt with a specific model, whereas other data can be difficult. This approach partially stems from curriculum learning [16], which has demonstrated that each datum cannot be equally learnt so as to be well-organised for model training. In the field of emotion prediction, a few studies have shown that the difficulty-level of the data to be learnt is closely related to its PU [45, 40], as discussed in Section 3.2.2.1. Motivated by these studies, the PU is employed to represent the difficulty and complexity of the samples.

In addition, in the second stage, i. e., the difficulty information exploiting stage, the original features $\mathbf{x}_t$ are concatenated with the difficulty vector $\mathbf{d}_t$ retrieved by one of the aforementioned two strategies, update the inputs in the form of $[\mathbf{x}_t, \mathbf{d}_t]$ to retrain the regression model for emotion prediction. This procedure is demonstrated in Figure 3.9 (b). While $\mathbf{d}_t$ varies along with time, the extended difficulty vector provides dynamic awareness when modelling $\mathbf{x}_t$ in a continuum. The pseudo-code describing the proposed algorithm is presented in Algorithm 1.

3.2.2.3 Difficulty Information Retrieval

The crux of the DDAT algorithm is the retrieval of the difficulty information, which then can be exploited dynamically to benefit the emotion prediction model. To this end, the following part will be devoted to two information retrieval strategies, i. e., RE-based (or ontology-driven) and PU-based (or content-driven), respectively.

Reconstruction Error-based Difficulty

As discussed in Section 3.2.2.1, the difficulty indicator $\mathbf{d}$ can be generated from the reconstruction process of the inputs. Thus, in the first stage, the model is trained in an MTL context, and the output includes two paths–the emotion prediction path and the AE path. The former is trained in a supervised fashion, whereas the latter is trained in an unsupervised manner. For this reason, there are two tasks to be carried out during the training phase, i. e., predicting emotions and reconstructing inputs. Specifically, given a time sequence as the input $\{\mathbf{x}_t\}$ with $t = 1, ..., T$, the network is optimised by minimising the loss function as

$$\mathcal{J}(\boldsymbol{\theta}) = w_1 \cdot L_{emt}(\cdot) + w_2 \cdot L_{re}(\cdot) + \lambda \cdot R(\boldsymbol{\theta}), \tag{3.30}$$

Algorithm 1: Dynamic Difficulty Awareness Training

Initialise:
θ: parameters of the model;
$\mathbf{x} \in \mathbb{R}^M$: input feature vector;
I, J: predefined training epochs

if *ontology-driven* **then**
  auxiliary task $\mathcal{T} \leftarrow$ reconstructing inputs;
else if *content-driven* **then**
  auxiliary task $\mathcal{T} \leftarrow$ predicting perception uncertainty;
end
% difficulty information retrieving stage
for $i = 1, ..., I$ **do**
  optimise θ via minimising a joint loss function $\mathcal{J}(\boldsymbol{\theta}) = w_1 \cdot L_{emt}(\cdot) + w_2 \cdot L_{aux}(\cdot) + \lambda \cdot R(\boldsymbol{\theta})$, where w_1 and w_2 regulate the contributions of the emotion prediction $L_{emt}(\cdot)$ and the auxiliary task prediction $L_{aux}(\cdot)$;
  evaluate the model on the development set for emotion prediction: $CCC_{dev,i}$;
  if $CCC_{dev,i} > CCC_{dev,i-1}$ **then**
    save h;
  end
end
obtain the difficulty indicator $\mathbf{d}$ based on the chosen auxiliary task $\mathcal{T}$; *% difficulty information exploiting stage*
for $j = 1, ..., J$ **do**
  update the input feature vector: $\mathbf{x}' = [\mathbf{x}, \mathbf{d}]$;
  optimise θ by minimising the loss function $\mathcal{J}(\theta)$ for emotion prediction;
  evaluate the model on the development set for emotion prediction: $CCC_{dev,j}$;
  if $CCC_{dev,j} > CCC_{dev,j-1}$ **then**
    save h;
  end
end

where $L_{emt}(\cdot)$ and $L_{re}(\cdot)$ denote the loss functions for emotion prediction and input reconstruction, respectively, w_1 and w_2 are predefined hyperparameters to regulate the contributions of $L_{emt}(\cdot)$ and $L_{re}(\cdot)$, and λ is another hyperparameter that controls the importance of the regularisation term $R(\boldsymbol{\theta})$. Furthermore, to calculate $L_{emt}(\cdot)$ and $L_{re}(\cdot)$, the Mean Square Error (MSE) is adopted for both learning paths,

i. e., for emotion prediction,

$$L_{emt}(\cdot) = \frac{1}{T}\sum_{t=1}^{T} ||\hat{y}_t - y_t||^2; \tag{3.31}$$

and for the input reconstruction,

$$L_{re}(\cdot) = \frac{1}{T}\sum_{t=1}^{T} ||\hat{\mathbf{x}}_t - \mathbf{x}_t||^2, \tag{3.32}$$

where $\mathbf{x}_t$ and y_t are a sample and its annotation at time t from an input sequence with a period of time T, respectively. $\hat{\mathbf{x}}_t$ and $\hat{y}_t$ denote the network predictions to reconstruct its input $\mathbf{x}_t$ and estimate the emotions y_t, respectively.

It is expected that $L_{re}(\cdot) \rightarrow 0$ if the model is sufficiently powerful and robust. However, empirical experiments have shown that the results are far from this expectation. Previous findings frequently indicate that a higher distribution mismatch between the given data and the entire training dataset is inclined to produce a higher RE [129, 216, 217, 238]. Therefore, the RE somewhat implies the difficulty degree of the model to learn such data or, in other words, reflects the difficulty of the data to be learnt by the model.

Once the model is trained, the difficulty indicator $\mathbf{d}_t$ can be obtained by computing the distance between the input $\mathbf{x}_t$ and its corresponding reconstruction $\hat{\mathbf{x}}_t$. The distance can be either a vector calculated by,

$$\mathbf{d}_t = \mathbf{x}_t - \hat{\mathbf{x}}_t, \tag{3.33}$$

or a scalar by summing the errors over all attributes, i. e.,

$$\mathbf{d}_t = \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2 , \tag{3.34}$$

where $\|\cdot\|_2$ denotes the Euclidean distance between the original input $\mathbf{x}_t$ and its reconstructed estimation $\hat{\mathbf{x}}_t$.

Perception Uncertainty-based Difficulty

In contrast to RE, for the content-driven difficulty retrieval strategy, PU is exploited. As mentioned earlier, PU is an indicator of the uncertainty level of the perception of an emotional state for a given observed sample. In the context of affective computing, emotion prediction is a subjective task that differs from many other objective pattern recognition tasks, such as face recognition, that hold a ground truth [176]. In order to obtain a gold standard for a subjective task, it is required that a sufficient number of raters observe the same sample and that their ratings are collected in order to eliminate as much as possible individual variations in perception and rating. In this context, a possible way to infer uncertainty is by calculating the

inter-rater disagreement level, which assumes that for each sample, the personal PU is highly correlated with the inter-rater disagreement level [131, 91].

Thus, the PU-based difficulty indicator at time t, u_t, can be represented by the standard deviation of a total of n annotations as

$$u_t = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_{t,i} - \bar{y}_t)^2}, \tag{3.35}$$

where $y_{t,i}$ is the i-th annotation with respect to x_t with $i = 1, ..., n$, and $\bar{y}_t$ denotes the mean value at time t given all n annotations:

$$\bar{y}_t = \frac{1}{n}\sum_{i=1}^{n} y_{t,i}. \tag{3.36}$$

To this end, similar as in the framework of RE-based DDAT, the difficulty indicator $\mathbf{d_t}$ can be determined by the perception uncertainty, i. e., $\mathbf{d_t} = u_t$. Therefore, the designed model includes an emotion prediction path and a PU prediction path, both of which are again jointly trained in a supervised manner. In particular, the objective function of Equation (3.30) can be reformulated as

$$\mathcal{J}(\boldsymbol{\theta}) = w_1 \cdot L_{emt}(\cdot) + w_2 \cdot L_{pu}(\cdot) + \lambda \cdot R(\boldsymbol{\theta}), \tag{3.37}$$

where $L_{pu}(\cdot)$ stands for the loss functions for PU prediction, and can be expressed by

$$L_{pu}(\cdot) = \sum_{t=1}^{T} ||\hat{u}_t - u_t||^2. \tag{3.38}$$

3.2.2.4 Difficulty Information Exploitation

As aforementioned, in order to infer the difficulty indicator $\mathbf{d}_t$, in the first stage of the DDAT framework, an MTL structure is implemented to jointly learn emotion prediction together with the reconstruction of the input features or the PU prediction.

After that, in the difficulty exploitation stage, as demonstrated in Figure 3.9 (b), the model input is updated with the new vector, i. e., $\mathbf{x}' = [\mathbf{x}_t, \mathbf{d}_t]$. When doing this under the RE-based strategy, the input feature vectors are of $2M$ or $M+1$ dimensions when feeding back an RE vector or scalar. Likewise, as for the PU-based DDAT framework, its input in the second learning stage will then be of $M + 1$ dimensions.

In addition, as mentioned already in Section 3.2.1.2, late fusion approaches can be conducted to combine the emotion predictions from different modalities, learning models, or a combination thereof. In this case, the late fusion can be performed following the linear regression approach as given in Equation (3.29). Specifically,

$y_{i,t}$ denotes the original prediction with the modality (i. e., audio or video) or model i (i. e., RE- or PU-based DDAT), ϵ and w_i are the parameters estimated on the development set, and y_t is the fused prediction at time t.

However, this conventional late fusion approach simply assumes that the predictions $y_{i,t}$ in a continuum are considered to be equally important for each prediction stream y_i. This means that the parameter of w_i remains a constant in time, given a set of y_i, and therefore, this approach ignores the changes of the reliability of the predictions along time. To address this problem, a *dynamic tuning* strategy is further proposed according to the reliability of predictions in time.

Mathematically, an additional linear regression is applied to the original prediction $y_{i,t}$ and the corresponding difficulty indicator $d_{i,t}$ at time t to generate a corresponding dynamic prediction $y'_{i,t}$:

$$y'_{i,t} = \epsilon + w_i \cdot y_{i,t} + w_d \cdot d_{i,t}, \tag{3.39}$$

where w_i and w_d are the parameters that can be determined on the development partition. Intuitively, the prediction is dynamically tuned by the difficulty information.

3.2.3 Adversarial Training

In this section, investigation on adversarial learning-driven emotion analysis techniques, as presented by the author and her colleagues in [86], will be discussed. As mentioned in Section 1.1, when deployed in real-life applications, affective computing systems face challenges such as the instability of the deep learning-driven models. Therefore, finding robust solutions to this challenge is an open and ongoing research direction.

In 2014, a novel learning algorithm called adversarial training (or adversarial learning) was first proposed by Goodfellow et al. [72], where a deep generative model can be learnt to model the data distribution of the target, while training jointly with another discriminative model as two players in a minimax game. Being a successful alternative to conventional maximum likelihood techniques, Generative Adversarial Networks (GANs) have attracted widespread research interests over the past few years across a range of machine learning domains [39, 204], including affective computing [64, 86, 157].

In line with the objectives of this thesis, and motivated by the promising results recently obtained by GANs, in this section, a conditional adversarial training framework to predict emotional states will be investigated. Specifically, this framework consists of two networks, trained in an adversarial manner: The first network tries to predict emotion from acoustic features, while the second network aims at distinguishing between the predictions provided by the first network and the emotion labels from the database using the acoustic features as conditional information.

It is noteworthy that, although this idea mainly stems from GANs, the proposed framework is different from other works where the main focus is how to best generate sufficiently realistic samples, such as images and acoustic samples [221, 233]. Instead, in this study, the proposed framework is devised for pattern recognition, especially emotion prediction.

Additionally, whereas the presented conditional adversarial training framework utilises a cascaded structure, as used in Strength Modelling (cf. Section 3.2.1) and Dynamic difficulty Awareness Training (cf. Section 3.2.2), it includes some specific advantages in comparison to those two approaches. The main idea of Strength Modelling is to take advantage of different models where predictions made by a first model are combined with the original features to learn a second model. Therefore, the two models should be diverse enough to provide complementary views and compensate for their respective weaknesses. Whereas the DDAT learning aims to explore the model weakness information that can be quantified by the difficulty level, in the assumption that the model could perform better if it is aware of its errors. Thus, the two stages should be as similar as possible, so that the weakness information extracted from the first stage can be presented for the second one. Both learning strategies are realised in an asynchronously cooperative way. That is, the well-trained first model provides additional information for the second one to assist in final decision making. The proposed conditional adversarial training framework, however, does not care about the similarity of the two models. Moreover, the two networks can be trained and optimised synchronously in a competitive way, rather than asynchronously in a cooperative way.

To this end, let us first review the structure of a basic GAN structure in Section 3.2.3.1. After that, an introduction of conditional GANs, which is the base of the proposed model, will be discussed in Section 3.2.3.2. Then, the framework of the proposed conditional adversarial training for prediction as well as its advanced version will be presented in Section 3.2.3.3 and Section 3.2.3.4, respectively.

3.2.3.1 Vanilla Generative Adversarial Networks

The conventional Generative Adversarial Networks (GANs), consist of two neural networks, namely, a generator G and a discriminator D, which contest with each other in a two-player zero-sum game, as illustrated in Figure 3.10. During this two-player game, the generator aims to capture the potential distribution of real samples and generates new samples to 'cheat' the discriminator as far as possible, whereas the discriminator, often a binary classifier, distinguishes the sources (i. e., real samples or generated samples) of the inputs as accurately as possible.

To be more specific, while D is trained to estimate the probability that a sample comes from the real data or the output of G, G learns to maximise the probability of fooling D. In other words, G and D are trained jointly in a minimax fashion.

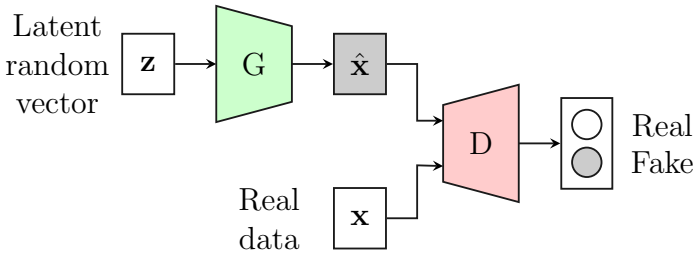


Figure 3.10: Framework of a vanilla Generative Adversarial Network (GAN) [84].

Mathematically, the minimax objective function can be formulated as:

$$\min_{\theta_g}\max_{\theta_d}\mathcal{L}(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D_{\theta_d}(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D_{\theta_d}(G_{\theta_g}(\mathbf{z})))], \tag{3.40}$$

where θ_g and θ_d denote the parameters of G and D, respectively; $\mathbf{x}$ is a real data instance following the true data distribution $p_{\text{data}}(\mathbf{x})$; whilst $\mathbf{z}$ is a vector randomly sampled following a simple distribution (e. g., Gaussian); $G_{\theta_g}(\mathbf{z})$ denotes a generated data given $\mathbf{z}$ as the input; and $D_{\theta_d}(\cdot)$ outputs the likelihood of real data given either $\mathbf{x}$ or $G_{\theta_g}(\mathbf{z})$ as the input. Note that, the likelihood is in the range of (0,1), indicating to what extent the input is probably a real data instance. Consequently, during training, θ_g is updated to minimise the objective function such that $D_{\theta_d}(G_{\theta_g}(\mathbf{z}))$ is close to 1; conversely, θ_d is optimised to maximise the objective such that $D_{\theta_d}(\mathbf{x})$ is close to 1 and $D_{\theta_d}(G_{\theta_g}(\mathbf{z}))$ is close to 0. In other words, G and D are trying to optimise a different and opposing objective function, thus pushing against each other in a zero-sum game. Hence, the strategy is named as adversarial training.

Generally, the training of G and D is done in iteratively, i. e., the corresponding neural weights θ_d, θ_g are updated in turns to compete with each other. Once training is completed, the generator is able to generate more realistic samples, while the discriminator can distinguish authentic data from fake data. More details of the basic GAN training process can be found in [72].

Since its inception, adversarial training has been frequently demonstrated to be effective in improving the quality of the simulated samples [39, 72, 204]. Furthermore, many researchers in this field have developed a bulk of variants, including but not limited to, conditional GAN [136], cycle GAN [241], and Wasserstein GAN (WGAN) [7], and Least-square GAN [128].

3.2.3.2 Conditional GANs

As aforementioned, conditional GAN (CGAN) is a variation of the vanilla GAN, which was first proposed by Mirza et al. in [136]. In the original CGAN framework, both the generator and discriminator are conditioned on certain extra information c.

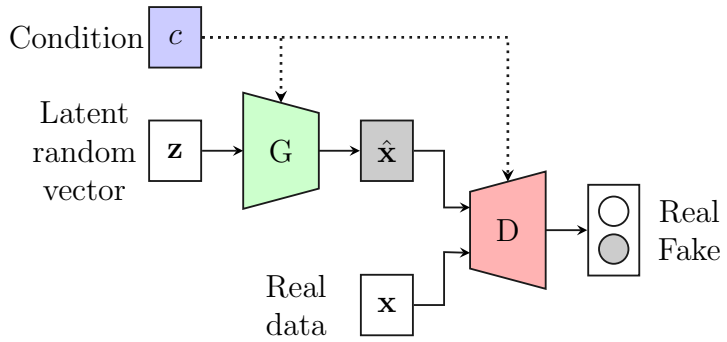


Figure 3.11: Framework of a Conditional Generative Adversarial Network (CGAN) [84].

This extra information can be any kind of auxiliary information, such as the labels or data from other modalities [136].

More specifically, for the generator G, the prior input noise variable $p_z(\mathbf{z})$ is combined with the conditional information c as a joint hidden representation. Likewise, in the discriminator, the real data $\mathbf{x}$ or the simulated data from the generator $G(\mathbf{z})$ is further extended with the conditional information c, which are then fed into the discriminator D, as shown in Figure 3.11.

In this circumstance, the minimax objective function in Equation (3.40) can be reformulated as

$$\begin{aligned}\min_{\theta_g}\max_{\theta_d}\mathcal{L}(D,G) =& \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D_{\theta_d}(x|c)]+ \\ & \mathbb{E}_{z\sim p_z(z)}[\log(1-D_{\theta_d}(G_{\theta_g}(z|c)))].\end{aligned} \tag{3.41}$$

With this structure, CGANs hold the capability to generate data within a predefined category.

3.2.3.3 Conditional GANs for Emotion Prediction

In contrast to CGAN used for data generation, a CGAN-based framework is proposed to exploit the concept of adversarial training to build a predictive model, where a predictor (or generator) and a discriminator are conditioned by obtained features from real recordings. In particular, the discriminator is employed to distinguish the joint probability distributions of features and their corresponding predictions or real annotations. In this way, the predictor is guided to modify the original predictions to achieve better performance.

The framework of the proposed conditional adversarial training is illustrated in Figure 3.12 (adapted from [86]). Whereas the structure is analogue to any CGAN with the presence of two networks, the ultimate goal of this model is to obtain an accurate pattern estimation from the generator which is guided by the discrimina-

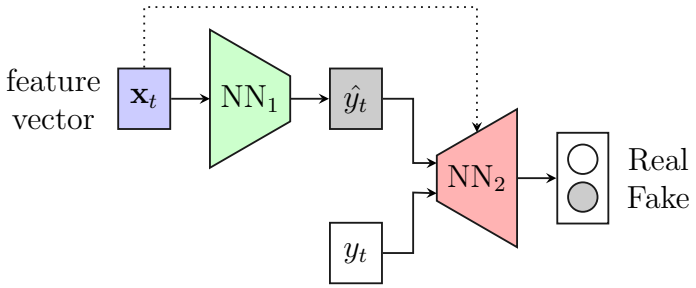


Figure 3.12: Framework of a conditional GAN framework for prediction: the first model (NN_1) predicts time-continuous labels $\hat{y}_t$ from a set of acoustic features $\mathbf{x}_t$, whereas the second model (NN_2) infers a binary decision whether the input source comes from the real data y_t or from the first model NN_1, given the context $\mathbf{x}_t$.

tor. To better demonstrate the proposed algorithm, hereinafter, the generator is presented by NN_1 and the discriminator by NN_2.

For this purpose, the algorithm of CGAN is modified. Specifically, let us ignore the prior random noise, and consider the features as the conditional information, i. e.,

$$c \leftarrow \mathbf{x}, \tag{3.42}$$

and the predicted emotional values as the generated data, i. e.,

$$G(\mathbf{z}|c) \leftarrow \hat{y}. \tag{3.43}$$

The first network (NN_1) is thus derived into a 'conventional-like' recognition model, and learns the conditional distribution $P(y_t|\mathbf{x}_t)$ given sequential features $\mathbf{x}_t$ and their labels y_t. Nevertheless, the major difference is that the NN_1 is optimised not only through its own prediction error, but also with the aid of adversarial feedback from the second network NN_2.

For NN_2, similar to CGAN, the generated one-dimensional predictions ($\hat{y}_t$) or the true labels (y_t) are extended with the auxiliary information (i. e., the original feature vectors, $\mathbf{x}_t$) to obtain a joint representation, i. e., $[\mathbf{x}_t, \hat{y}_t]$ or $[\mathbf{x}_t, y_t]$, and then employed as the inputs of NN_2. The network learns to identify whether the joint representation comes from the generation of the first network (False) or from the original labels (True). Therefore, NN_2 is trained to distinguish the joint probability distributions for features and their corresponding 'predicted' (false) labels, i. e., $P_g(\mathbf{x}_t, \hat{y}_t)$, or 'real' labels, i. e., $P_r(\mathbf{x}_t, y_t)$.

More concretely, NN_1 can be optimised by

$$\mathcal{L}_{NN_1} = \mathbb{E}(|G(\mathbf{x}_t) - y_t|^2) + \lambda \cdot \mathbb{E}(log(D(G(\mathbf{x}_t), \mathbf{x}_t))), \tag{3.44}$$

where the first item indicates the MSE between the prediction and the label, and λ denotes a hyperparameter that controls the contribution of the adversarial informa-

tion from NN_2, which can be optimised by

$$\mathcal{L}_{NN_2} = \mathbb{E}(log(D(y_t, \mathbf{x}_t))) + \mathbb{E}(log(1 - D(G(\mathbf{x}_t), \mathbf{x}_t))). \tag{3.45}$$

In this manner, intuitively, NN_1 is optimised to generate predictions as close as possible to the labels, while fooling NN_2 when fed with joint distributions composed of predictions from NN_1 and original acoustic features.

3.2.3.4 Optimising with Wasserstein Distance

Traditional generative modelling approaches rely on maximising the likelihood, or equivalently minimising the Kullback-Leibler (KL) divergence between the realistic data distribution P_r and the generated data distribution P_g [72]. One major issue this approach suffers is the vanishing gradient problem as demonstrated in [6], because the discriminator with the infinite ability to separate real from generated samples will lead to a constant Jensen-Shannon (JS) divergence between P_r and P_g when their supports have no or negligible overlap [6]. This results in an impossibility to update the generator accordingly, as the discriminator is quickly trained towards its optimality [6].

In this light, a Wasserstein (also known as Earth-Mover) distance was proposed most recently [7], so that the JS distance problem in the classic GAN can be solved by showing that the Wasserstein distance is continuous and differential almost everywhere. Motivated by this work, the NN_2 training strategy can further be updated by maximising

$$\mathcal{L}_{NN_2} = \mathbb{E}[D(\mathbf{x}_t)] - \mathbb{E}[D(G(\mathbf{x}_t))], \tag{3.46}$$

and the NN_1 training strategy by minimising

$$\mathcal{L}_{NN_1} = \mathbb{E}(|G(\mathbf{x}_t) - y_t|^2) + \lambda \cdot \mathbb{E}[D(G(\mathbf{x}_t))]. \tag{3.47}$$

In addition, a weight clipping is further applied to the NN_2 as follows

$$w_{NN_2} \leftarrow clip_by_value(w_{NN_2}, -0.01, 0.01), \tag{3.48}$$

in order to improve the stability of the training process, as suggested in [7].

3.3 From Isolated to Continual Learning

Though great progress has been achieved when training models with the aforementioned approaches, all the obtained models are still task-specific. For instance, these tasks might include but are not limit to, recognising six or eight discrete emotion categories from English speech, or predicting continuous arousal from facial images of German children. Hence, the number of models will be expanding when needing to learn more tasks. That, however, is somewhat pragmatically unfeasible to

be implemented in real-life intelligent systems due to the limited computational resources and great complexity. *Lifelong learning*, also known as *continual learning*, has approached this issue, by learning a series of tasks in a sequence with one model. In other words, it empowers machines with the capability of continually acquiring and transferring knowledge and skills throughout their lifespan, just like human beings. As a result, concepts and their relationships learnt in the past can help a machine understand and learn about a future task better, because knowledge can be transferred and shared across various tasks [31, 147].

In the following, the definition and goals of lifelong learning will first be introduced in Section 3.3.1. Then, a typical lifelong learning paradigm is presented will be presented in Section 3.3.2. After that, the procedures of how to construct and train a continual emotion recognition model are elaborated in detail in Section 3.3.3.

3.3.1 Lifelong Learning and Catastrophic Forgetting

Lifelong learning was first defined in [193] as a learning algorithm applied in a lifelong context, where a series of tasks can be learnt sequentially instead of in isolation, so that knowledge can be transferred across these tasks. In particular, like humans, knowledge obtained in previous learning tasks should be retained for future use.

Lately, the concept of *lifelong learning* was defined in [31] as follows. Given a stream of n tasks $\{T_1, T_2, ..., T_n\}$ already learnt by a model, a knowledge base B can be obtained where all previously learnt knowledge is maintained. Note that, these tasks can be of different types and from different domains [31]. Then, when a new task T_{n+1} comes, the objectives of lifelong learning are mainly twofold: on the one hand, the existing knowledge in B should be leveraged to help optimise the performance of the new task T_{n+1}; on the other hand, the knowledge obtained from T_{n+1} should be integrated into B by updating it without causing a forgetting of prior knowledge of all past tasks.

From the above description, it could be noticed that the knowledge base B plays a crucial role throughout the whole learning process. This, to some extent, can gain insight from human learning mechanisms. We, as human beings, could learn better, easier, and faster, when we have more knowledge obtained on our previous life experiences [31]. Therefore, following the line of human intelligence, the crux of learning in a lifelong manner is to endow the system the capability of performing human-like knowledge-based learning.

However, in the context of deep learning, the *catastrophic forgetting* issue associated with the knowledge base B is considerably severe in most network systems. That is, when training a deep learning system (trained on some other task already) for a new task, the new learning process tends to interfere catastrophically with the previous learning. As a consequence, the performance of the system will be degraded heavily for past tasks, which is in contrast to the human brains and human learning systems [132]. To address this issue, in recent years, a number of

lifelong learning techniques have been proposed and studied in the deep learning community, generally falling into three categories: dynamic architecture-based approaches, memory-based ones, and regularisation-based ones. For more details of various lifelong learning strategies, please refer to [31, 147].

3.3.2 Elastic Weight Consolidation

In this study, a regularisation-based lifelong learning approach called *Elastic Weight Consolidation* (EWC), which tackles the catastrophic forgetting problem by regularising the parameters in a network, is chosen for the first attempt. The approach was first proposed by Kirkpatrick et al. [111] in 2017. Recently, it has been successfully investigated in several domains and applications [32, 192]. In addition, it has been derived into several relevant variants being investigated, such as EWC++ [28], online EWC [181], and R-EWC [121].

The idea of EWC is motivated by a consolidation process of human memory, which is known as synaptic consolidation or synaptic maintenance. This process enables us to consolidate previous memories within the related synapses to handle long-term memory tasks by reducing the plasticity of these synapses, and thus, these previously consolidated memories will not be altered by a future memory [35].

Similarly, in EWC, the plasticity of the task-relevant parameters (like the synapses in nervous systems) in a previously learnt model can be altered accordingly, to avoid changing significantly on these parameters when a future unseen task arrives. Especially, this is achieved by regularising the learning process with a quadratic penalty on the difference between the parameters for the prior and current tasks, the process of which is detailed in the following paragraphs.

In general, as aforementioned, given a prior optimised configuration of parameters $\theta^*_{1:n-1}$ for the past $n-1$ tasks, the objective of a lifelong learner is to learn an updated set $\theta^*_{1:n}$ for a new task T_n. Note that, the trainable parameters in a neural network consist of weights as well as biases. In the following, an example when $n = 2$ is illustrated for an easy understanding, but EWC works when $n > 2$ as well. In this case, two datasets $\mathcal{D}_1 = (\mathcal{X}_1, \mathcal{Y}_1)$ and $\mathcal{D}_2 = (\mathcal{X}_2, \mathcal{Y}_2)$ are applied for the two tasks T_1 and T_2, respectively, with samples $x \in \mathcal{X}$ and their corresponding labels $y \in \mathcal{Y}$. Additionally, the sets of parameters Θ^*_1 and Θ^*_2 denote the configurations that deliver low loss (i. e., high performance) for θ_1 and θ_2, where θ_1 and θ_2 represent the parameters of T_1 and T_2, respectively.

In a conventional learning system where T_1 and T_2 are trained independently, our target is to find two configurations that meet the following criteria:

$$\theta^*_1 \in \Theta^*_1 \text{ or } \theta^*_2 \in \Theta^*_2, \tag{3.49}$$

where θ^*_1 and θ^*_2 denote the configurations of θ that result in a good performance for T_1 and T_2, respectively.

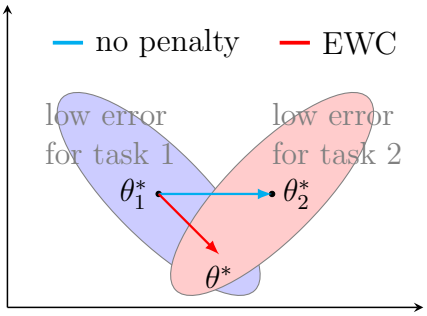


Figure 3.13: Illustration of Elastic Weight Consolidation (EWC)

Nevertheless, in EWC, as T_2 is learnt after T_1, the optimisation of θ_2 is distinguished from that in Equation (3.49), and can be given as follows:

$$\theta_2^* \in \Theta_1^* \cap \Theta_2^*, \tag{3.50}$$

where θ_2 is optimised with a constraint to stay in a low-error region of T_1 centred around θ_1^*. In this manner, a good configuration θ_2 should lie in the intersection of the low-error regions of T_1 and T_2 to prevent forgetting T_1, as depicted in Figure 3.13.

Mathematically, when considering the learning process from a probabilistic perspective, the relevance of the trainable parameters θ with respect to all the training data (i. e., $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$) can be modelled as the posterior probability distribution $p(\theta|\mathcal{D})$. Then, the logarithm value of it can be decomposed by the Bayes' theorem:

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D}). \tag{3.51}$$

Then, when splitting $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$ and employing them one after the other, the above equation can be written as:

$$\begin{aligned} \log p(\theta|\mathcal{D}) = &\log p(\mathcal{D}_2|\mathcal{D}_1, \theta) + \log p(\theta|\mathcal{D}_1) \\ &- \log p(\mathcal{D}_1|\mathcal{D}_2), \end{aligned} \tag{3.52}$$

where $\mathcal{D}_1$ denotes the data for the prior task T_1, while $\mathcal{D}_2$ is for the current task T_2.

Furthermore, let us assume that $\mathcal{D}_1$ and $\mathcal{D}_2$ are independent. In this circumstance, Equation (3.52) can be reformulated as

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_2|\theta) + \log p(\theta|\mathcal{D}_1) - \log p(\mathcal{D}_1), \tag{3.53}$$

where the posterior probability term $\log p(\theta|\mathcal{D}_1)$ contains all the knowledge related to T_1. Hence, when implementing EWC, the aim is to get information about the parameter importance from $\log p(\theta|\mathcal{D}_1)$, and then take this information into account in the succeeding learning to prevent or at least mitigate the forgetting. Unfortunately, this posterior is intractable. To handle this issue, Laplace approximation is

employed to approximate it as a Gaussian distribution with mean given by θ and a diagonal precision by the diagonal of the Fisher information matrix F as

$$F_i = \mathbb{E}_{\mathcal{D}_1}\left[\frac{\partial^2 \log p(\mathcal{D}_1|\theta)}{\partial \theta_i^2}\right]\Bigg|_{\theta=\theta_1^*}, \tag{3.54}$$

where i denotes the index of values on the diagonal of the matrix F. F_i then can be exploited to estimate the importance degree of each parameter to T_1. For instance, a great value of F_i indicates a high importance degree, and thus implies that it should not be changed much so that the performance of prior tasks can be maintained.

As a consequence, when training the network for T_2 after T_1, a penalty term with respect to T_1 is added to the objective function as

$$\mathcal{L}'(\theta) = \mathcal{L}(\theta) + \frac{1}{2}\lambda \cdot \sum_i F_i(\theta_i - \theta_{1,i}^*)^2, \tag{3.55}$$

where $\mathcal{L}'(\theta)$ represents the new loss function in EWC, and $\mathcal{L}(\theta)$ is the loss for T_2 only. In addition, λ is a pre-defined hyperparameter to regulate how important the past tasks are compared to the current one, and i is the index of each trainable parameter.

In summary, when learning for a new task, EWC penalises large changes to the most relevant parameters with respect to past tasks and suggests updating parameters for the new task mainly along with directions with low Fisher information. In our case, this training strategy enables T_2 to be learnt without suffering catastrophic forgetting of T_1.

3.3.3 Continual Emotion Recognition

As discussed in Section 3.3.2, aiming at addressing catastrophic forgetting, a network can be learnt with the assistance of EWC, in which the importance of parameters with respect to prior tasks is exploited to selectively adjust the plasticity of parameters when a new task is given.

Although several studies investigate this learning approach and its variants as aforementioned, it is believed that this is the first attempt to explore it in the field of audiovisual affective computing. In particular, in this work, cross-cultural emotion recognition is taken as a paradigm to investigate its effectiveness. Being an active research area in affective computing nowadays, the main goal of cross-cultural emotion recognition is to establish a compatible framework to handle the discrepancy across multiple cultures.

In this case, a neural network will be learnt to estimate emotion patterns, via learning sequentially from a series of databases $\{\mathcal{D}_n\}$ with $n = 1, 2, ..., N$, each consisting of plenty of emotional instances from one specific culture. After training on one database for the current culture T_n, the Fisher information can be estimated

Algorithm 2: The training process of EWC-based continual cross-cultural emotion recognition.

Initialise:
N: the total number of databases;
n: index to indicate the n-th database;
$\mathcal{D}_n = (\mathcal{X}_n, \mathcal{Y}_n)$: the n-th database;
θ: parameters of the model;
θ_0: initialised parameters of θ;

for $n = 1, ..., N$ **do**
  if *n=1* **then**
    optimise θ via minimising a conventional loss;
  else
    compute the Fisher Information for the prior $\mathcal{D}_{n-1}$;
    save the prior configuration θ^*_{n-1};
    optimise θ via minimising an EWC-based loss function;
  end
  $\theta \leftarrow \theta^*_n$;
end

as given in Equation (3.54), and in the meanwhile, the best configuration of θ^*_n is kept for future usage. Then, when given a new database of another culture T_{n+1}, the network will be optimised using Equation (3.55). These procedures can be performed repeatedly until all N cultures have been used for training the network. The pseudo-code describing these procedures is also presented in Algorithm 2.

The expectation of the approach is that after training the N-th culture, the performance of all prior $N-1$ tasks is not heavily degraded.

3.4 Synchronisation Behaviour Analysis Based on Autoencoders

Research in psychology has shown that people unconsciously mimic their counterpart in social interaction, which can be operationalised in varying ways including mimic posture, facial expressions, mannerisms, and other verbal and nonverbal expressions [27]. Moreover, the automatic detection of temporal mimicry behaviour can serve as a powerful indicator of social interaction, e. g., cooperativeness, courtship, empathy, rapport, and social judgement [77].

The previous works focus on automatically detecting mimicry behaviours particularly from head nod and smile, i. e., visual cues [17, 189]. In this present study, let us focus on the acoustic side, given that in social interaction, people mimic others not

only by body language, but also in their speech. To the best of our knowledge, this is the first time the identical behaviour is analysed from speech over different cultures in empirical research, though previous works exist where similar topics have been studied in theory [22]. As there are limited related works into this specific topic, low-level descriptors (LLDs) were first explored, such as log-energy, and pitch, and measured the similarities over each conversation turn. However, no obvious trend in these descriptors was found.

Thus, an autoencoder-based framework is proposed to harness the power of machine learning, and has been implemented by the author and her colleagues in [82]. Specifically, for each interaction, two autoencoders (AEs) are trained on the speech from two subjects A and B, respectively. Then, once the training procedure is done, the data are exchanged and fed into the two autoencoders again, i. e., A is evaluated on the AE trained by data from B while B is evaluated on the AE trained by data from A. This goes under the hypothesis that, when a subject tends to behave similarly to her counterpart, the reconstructed features from the AE trained with her counterpart's data should have a decreasing error along the time.

3.4.1 Introduction of Synchronisation Behaviour

Synchronisation behaviour, also known as mimicry behaviour, can be categorised into two different groups: *emotional mimicry* and *motor mimicry* [96]. The first describes mimicry in the underlying affective state, such as *happiness* or *sadness*, whereas the latter considers only the imitation of physical expressions, e. g., raising an eye-brow or nodding the head. As can be expected, motor mimicry is much easier to detect than emotional mimicry, given that physical expressions can be classified quite objectively by a human observer and also by automated tools. In the late 1970s, Friesen and Ekman proposed the 'Facial Action Coding System' [66] based on so-called *facial action units* (FAUs). FAUs describe 44 different activations of facial muscles, resulting in a certain facial expression, e. g., 'raising eye brow', 'wrinkling nose', or 'opening mouth'. However, several FAUs can be combined and be active at the same time. Ekman and Friesen have also shown that, there is a strong relationship between FAUs and affective states [55] and that those relationships are largely universal despite there are some differences between cultures [53]. FAUs and head movements can be robustly recognised with state-of-the-art tools, such as OpenFace [11].

Motor mimicry is a means of persuasion in human-to-human interaction, by conforming to the other's opinions and behaviour [96]. Humans are susceptible to mimic behaviours through both the audio and the visual domain [149]. Although mimicry is usually found in interactions both when subjects disagree with each other and when they do not, there are more mimicry interactions where people agree [189]. Moreover, it has been shown that there is usually a tendency to adopt gestures, postures, and behaviour of the chat partner during the conversation [27, 43].

From the methodological point of view, for the automatic detection of behavioural mimicry, a temporal regression model has been proposed to predict audio-visual features of the chat partner using a deep recurrent neural network [17]. An ensemble of models has been trained for each class (mimicry / non-mimicry) and the ensemble providing the lowest reconstruction error determined the class. *Mel-frequency cepstral coefficients* have been employed as acoustic features and *facial landmarks* as visual features.

Compared to motor mimicry, emotional mimicry has been studied much less. However, it has been found that the tendency to mimic others' behaviour is much less valid from the emotional perspective [97]. The extent of emotional mimicry highly depends on the social context, and the emotional mimicry is not present if people do not like each other or each other's opinion. Scissors et al. found out the same conclusion when analysing linguistic behaviours [182]. They observed that in a text-based chat system, within-chat mimicry (i. e., repetition of words or phrases) was much higher in chats where subjects trusted each other than in chats with a low level of trust [182]. Furthermore, it was found that linguistic mimicry has a positive effect on the outcome of negotiations [190].

3.4.2 Autoencoder-based Synchronisation Behaviour Analysis

To analyse the interpersonal sentiment and investigate the temporal behaviour patterns from speech, the 130 frame-level features within the same recordings are first standardised (zero mean and unit standard deviation) to minimise the differences between different recording conditions. This procedure makes the values of the extracted LLDs into suitable ranges, as the inputs and targets of an AE. After that, the whole LLD sequences of each recording can be divided into two groups, each including features from one subject in a pairwise chat interaction.

Following the aforementioned separation process, features from one subject can be utilised to train an AE, and features from the other subject in the same recording are then fed into the trained AE for testing. Furthermore, once all features for testing have been reconstructed with the AE, the root-mean-squared errors (RMSEs) of the reconstructed features over time were computed and examined how and to which extent the RMSE varies along time. Consequently, for each recording, two AEs are learnt based on the two subjects involved in the recording, resulting in two one-dimensional RMSE sequences calculated between the input and the output feature sequences during the testing step.

4

Experimental Evaluations

In this chapter, comprehensive experiments for automatic audiovisual emotion recognition are carried out, to verify the effectiveness and reliability of the presented methods in Chapter 3. More specifically, in this chapter, three emotional databases are first described in Section 4.1. Then, general experimental setups are presented in Section 4.2, including the applied acoustic feature sets as well as related performance measures. Afterwards, the following sections from Section 4.3 to Section 4.9 elaborate empirical performance evaluations and result discussions for each of the proposed learning algorithms.

4.1 Spontaneous and Multimodal Emotional Databases

Before discussing the experimental settings to evaluate the proposed algorithms, the following subsections briefly summarise the three audiovisual emotion databases which are utilised for the performance evaluation throughout the thesis. Note that, all these databases are publicly available for research purpose. Specifically, in Section 4.1.1 a standard database RECOLA in French is first introduced. Then, Section 4.1.2 describes another multi-cultural emotion database SEWA. These two databases are widely used for continues emotion regression. Finally, another database that is mainly designed for emotion classification tasks is presented, i. e., OMG-Emotion in Section 4.1.3.

4.1.1 RECOLA Database

RECOLA, short for REmote COLlaborative and Affective interactions, was developed by Ringeval et al. [164]. This multimodal database is widely used for audiovisual dimensional emotion estimation, and it is also a standard database which was previously used in a series of AVEC challenges since 2015 [163].

Table 4.1: Three partitions of the RECOLA database

#	**train**	**development**	**test**
female	6	5	5
male	3	4	4
French	6	7	7
Italian	2	1	2
German	1	1	0
age μ (σ)	21.2 (1.9)	21.8 (2.5)	21.2 (1.9)

RECOLA, was created to study socio-affective behaviours from multimodal data in the context of remote collaborative tasks [164]. More specifically, it consists of spontaneous and natural interactions from 46 French-speaking subjects involved in a dyadic collaborative task, and 27 of these subjects signed the consent forms to make their multimodal data accessible for research purpose. In each interaction, multimodal signals including audio, video, and peripheral physiology recordings such as ECG and Electro-Dermal Activity (EDA) were recorded continuously and synchronously over time. In this thesis, however, only audio and video recordings are utilised for the main focus of the present study.

It is worth to mention that, these subjects have different mother tongues, i.e., French, Italian, and German, which provides further diversity in the encoding of affect. In order to ensure speaker-independence, the corpus was equally divided into three partitions, i.e., training, development (validation), and test, with each partition containing nine unique recordings, by approximately balancing the gender, age, and mother tongue of the participants (cf. Table 4.1).

In addition, to obtain the annotations, time- and value-continuous dimensional affect ratings in terms of arousal and valence were performed by six annotators. Thereafter, these obtained annotations were resampled with a constant frame rate of 40 ms to align with the frame rate of the recordings. 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' [164].

4.1.2 SEWA Database

After introducing the RECOLA database, in the following, I briefly introduce another audiovisual database for continuous emotion recognition, i. e., the SEWA database [112]. This corpus was collected within the Automatic Sentiment Analysis in the Wild project[1]. In the database, a total of 197 conversations have been recorded

[1] https://sewaproject.eu/

Table 4.2: SEWA corpus: Number of conversations and subjects and total duration in minutes for each culture.

index	culture	# conversations	# subjects	total duration (min)
C1	Chinese	35	70	101
C2	Hungarian	33	66	67
C3	German	32	64	89
C4	British	33	66	94
C5	Serbian	36	72	98
C6	Greek	28	56	81
Sum		197	394	530

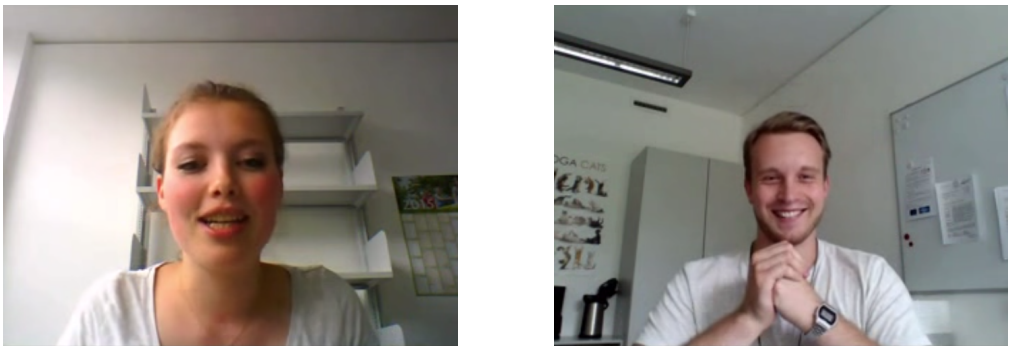


Figure 4.1: Screen-shot taken from one example recording with two subjects in the SEWA database [91].

from subjects of six different cultures (Chinese, Hungarian, German, British, Serbian, and Greek). Table 4.2 demonstrate a summary of the SEWA database. Specifically, each conversation lasted up to 3 minutes, and in each conversation two subjects from the same culture had a remote discussion talking about an advertisement they had watched beforehand on the web platform. Figure 4.1 depicts a screen-shot of one example of such a conversation.

It is noteworthy that, in SEWA, all conversations were recorded in an 'in-the-wild setting', i. e., using the subjects' personal desktop computers or notebooks and recording them either at their homes or in their offices. The chat partners always knew each other beforehand (either family, friends, or colleagues) and were balanced w. r. t. gender constellations (female-male, female-female, male-male) [112]. As a result, subjects with an age ranging from 18 to older than 60 are included in the database. The dialogues had to be held in the native language of the chat partners, but there were no restrictions concerning the exact aspects to be discussed during their chat about the advertisement [112]. Moreover, in these conversations,

a large variety of emotions and levels of agreement/disagreement or rapport were observed [112].

These aforementioned conversations were recorded at a video sampling rate between 20 and 30 fps and an audio sampling rate of either 44.1 or 48.0 KHz, depending on the recording devices used by the subjects. In particular, for all six cultures, continuous annotations were conducted first in German. That is, similar as in RECOLA, continuous dimensional affect ratings in terms of arousal and valence were performed by six German-speaking annotators for all 64 German recordings. The obtained annotations were then resampled at a constant frame rate of 100 ms, and the 'gold standard' was created with the Evaluator Weighted Estimator (EWE) [75] algorithm based on the resampled annotations.

Consequently, these annotated German video-chat recordings in the SEWA dataset has been applied as an official benchmark database in the series of AVEC challenges since 2017 [162]. In particular, the 64 recordings were divided into three speaker-independence partitions, i. e., 34 recordings for the training set, 14 for the development, and the rest 16 for the test sets. This leads to 55 072, 22 307, and 27 597 annotated frames in the training, development, and test sets, respectively, and becomes the testbed to evaluate some of the approaches proposed in the thesis. Furthermore, the dataset is available to researchers for non-commercial use via https://db.sewaproject.eu. Along with the audiovisual episodes and the annotations, hand-crafted acoustic and visual features are provided as well.

4.1.3 OMG-Emotion Database

The last database introduced in this session is the OMG-Emotion corpus, short for the One-Minute Gradual-Emotional Behavior dataset, which is created by categorical emotion recognition and utilised as a benchmark database in the OMG challenge in 2018 [13].

This dataset consists of 567 emotional monologue videos collected from Youtube, with an average length of one minute. These videos were then divided into utterance-level clips, and annotated by at least five annotators [13] via the Amazon Mechanical Turk platform. For annotation, seven categorical emotions were considered, i. e., anger, fear, disgust, happiness, sadness, surprise, and neutral. Majority voting was then applied to compute the 'gold standard' based on all annotations of the same utterance. Moreover, the dataset was split into training, development, and test sets, resulting in 2 440, 617, and 2 229 segments for each partition, respectively. Note that, in this thesis, performances will be reported only on the development set, as labels of the test set are not yet accessible.

For more details of the OMG-Emotion database and the series of OMG challenges, the reader is referred to [13].

4.2 Experimental Setup

In this section, typical experimental setups are provided. In detail, various audio and visual feature sets adopted within the thesis are introduced Section 4.2.1, respectively. Moreover, at the end of this session, performance measures to evaluate the emotion recognition performance of different models are given in Section 4.2.2.

4.2.1 Audiovisual Features

This subsection provides a general description of the audio and visual features employed in the studies within the scope of this thesis.

Acoustic Features
When predicting emotional states from the audio data, acoustic features should first be extracted from the raw audio recordings. In general, acoustic features can be categorised into two distinct groups, i. e., Low-Level Descriptors (LLDs) per frame and super-segment-level functionals. Especially, LLDs are normally derived from quite short-term frames based on short-time analysis, while super-segment-level functionals attempt to model long-term patterns by applying functionals (e. g., extremes, means, ranges, and percentiles) over LLDs of multiple successive frames. In the following subsections, detailed descriptions of the selected acoustic feature sets used in this thesis will be presented.

First, frame-level MFCCs has been exploited in a wide range of speech-related or audio-related applications, such as automatic speech recognition and acoustic scene detection. Besides, MFCCs are also common in speech emotion recognition. In this context, the open-source openSMILE [61] was used to generate 13 LLDs, i. e., MFCCs 0-12 and logarithmic energy, with a frame window size of 25 ms at a step size of 10 ms.

In addition to MFCCs, other acoustic features can also be taken into consideration to contribute to the prediction process. In this circumstance, the established 65 LLDs set from the INTERSPEECH 2013 Computational Paralinguistic ChallengeE (ComParE) [179], which were extracted with a frame window size of 20 ms or 60 ms (for different LLDs) at a step size of 10 ms. The ComParE LLD set consists of spectral (relative spectra auditory bands 1-26, spectral energy, spectral slope, spectral sharpness, spectral centroid, etc.), cepstral (Mel frequency cepstral coefficient 1-14), prosodic (loudness, root mean square energy, zero-crossing rate, F_0 via subharmonic summation, etc.), and voice quality (probability of voicing, jitter, shimmer and harmonics-to-noise ratio). In addition, the first-order derivatives (as known as deltas) can also be computed, resulting in a total of 130 LLDs for each frame. For more details about these LLDs, please refer to [179].

Other than MFCC and ComParE 2013 feature sets, the third LLD-based acoustic feature set related to the present work is based on a minimalistic expert-knowledge

based feature set, i. e., eGeMAPS [59]. Different from large-scale sets such as ComParE, its main goal is to reduce the risk of over-fitting during training. As a consequence, it contains only 23 LLDs, including but not limited to energy, spectral, cepstral, and voice quality information. Particularly, these LLDs were selected with respect to their capability to describe affective physiological changes in voice production. For more details about these LLDs, please refer to [59].

In contrast to these frame-level LLD-based feature sets where only short-term information is captured, functionals can further be applied to these LLDs to generate super-segment-level features. This is mainly due to the reason that statistic functionals can extract long-term patterns of emotional cues. For this reason, the most frequently used functionals are the arithmetic mean and the coefficient of variance. For instance, when computing means and variances over the sequential 13 LLDs within a window, it leads to 26 functional-level features.

Also, more other functionals can be applied to LLDs, such as extremes, means, moments, percentiles, peaks, temporal variables [61], resulting in more high-level features. This yields to the large ComParE 2013 standard feature set with 6373 features and the compact eGeMAPS feature set with 88 features, which will be investigated for different approaches accordingly. These feature sets were selected and explored in this research, mainly because that, they have been successfully and widely applied to and shown great performance in emotion recognition [227, 198].

Apart from LLDs and functionals, another type of super-segmental-level feature set, i. e., the Bag-of-Audio-Words (BoAW)features, can be computed with the help of the open-source toolkit openXBOW [171] based on the aforementioned LLDs. For a detailed description of the BoAW feature generation process, the reader is referred to Section 3.1.2.1.

Visual Features

To model facial emotional expressions, various types of visual features were taken into account in the thesis, i. e., facial landmarks, appearance-based features, geometric-based features, and deep learnt features. In the following, each feature type will be introduced.

First, to obtain the visual features, locations of 49-point facial landmark can be detected and used per frame, in line with the work described in [185]. The detected face regions consist of the left and right eyebrows (five points respectively), the left and right eyes (six points respectively), the nose (nine points), the inner mouth (six points), and the outer mouth (twelve points). Further, to reduce the variance of these landmark points, these points were normalised before being fed into the proposed models.

Second, the appearance-based visual features can be retrieved via the Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) algorithm [5], by splitting video recordings into spatio-temporal video volumes. In detail, each slice of the video was first convolved with a bank of 2D Gabor filters, and then divided into

blocks of different sizes. Thereafter, the Local Binary Pattern (LBP) operator was performed to each block, and then the LBP histograms from all blocks were concatenated to build appearance vectors. Last, a feature reduction was carried out by applying Principal Component Analysis (PCA). With this manner, the appearance vectors were projected into a lower-dimensional subspace to yield a compact but informative representation, resulting in fewer features per frame. For more details of the feature extraction process, please refer to [5].

Third, the geometric-based feature set was based on the 49 facial landmarks aforementioned. Specifically, after detecting these landmark locations, as features for each frame, 316 features were computed. Specifically, 196 features were obtained by calculating the difference between the coordinates of the aligned landmarks and those from the mean shape and between the aligned landmark locations in the previous and the current frame. Next, 71 features were computed by calculating the Euclidean distances (L2-norm) and the angles (in radians) between the points in three different groups. Then, another 49 features were reaped by computing the Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame. For more details on the feature extraction process, the reader is referred to [198].

It should be noted that a facial activity detector was used in conjunction with these video feature extraction procedures, and thus facial features were not extracted for the frames where no face was detected. As an alternative, the obtained sequential feature sets were interpolated by a piece-wise cubic Hermite polynomial to deal with the missing frames [198].

Moreover, similar to the super-segment-level acoustic features, functionals (the arithmetic mean and the coefficient of variance) or Bag-of-Video-Words (BoVW) can be computed over the sequential frame-level visual features within a fixed-length window, to deliver long-term representations for further processing.

Apart from the prior hand-crafted visual feature sets, a deep-model-based feature set was employed in some experiments, for a better comparison with other prior studies. In particular, a multi-task cascaded CNN [226] was first proposed for face detection and alignment on each frame. After that, frame-level intermediate deep representations of size 4 096 were extracted from the "fc-7" layer of the VGG-Face model [148], which was pre-trained on a large number of facial images. Lastly, average pooling was conducted on all frames of the same clip to deliver the final utterance-level visual representations.

4.2.2 Performance Measures

To evaluate the performance of the proposed approaches for emotion recognition, a number of measures have been proposed and utilised. In the following, a brief introduction of four frequently-used measures in the present work will be given.

Typically, these measures can be divided into two groups, two for classification tasks and another two for regression tasks.

Specifically, for emotion classification, the most frequently-used measurement is *Unweighted Average Recall* (UAR). For a given model, aiming at evaluating its general performance over all classes, UAR emphasises the average accuracy over all classes and can be computed as the sum of class-wise recall divided by the number of classes:

$$UAR = \frac{\sum_{i=1}^{k} recall_i}{k}, \tag{4.1}$$

where i denotes the i-th class of all k classes, and $recall_i$ represents the ratio of samples that are predicted as the i-th class correctly over all samples that are labelled as the i-th class. Hence, a higher UAR indicates a better performance.

Besides of UAR, another measure for emotion classification tasks used in the present work is *F1 Score* (also known as F-score or F-measure). Generally, the F1 score is the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}, \tag{4.2}$$

where *precision* is the number of samples that are predicted correctly divided by the number of all samples that are predicted as this class, while *recall* is calculated as the number of samples that are predicted correctly over the number of samples annotated as this class. Therefore, the best model can achieve an F1 score of 1 while the worst F1 score gets down to 0.

Having introduced the two performance measures for emotion classification, now let us move to the first frequently-used measure for emotion regression, i. e., *Concordance Correlation Coefficient* (CCC). In an emotion regression task, CCC measures the agreement between the gold standard and the predictions. Nowadays it is a standard evaluation metrics for time- and value-continuous emotion prediction, and exploited in several challenges such as AVEC and OMG-Emotion challenges [162, 13]. Mathematically, given two time series x and y, the CCC is calculated as follows:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{4.3}$$

where ρ is the Pearson's Correlation Coefficient (PCC) 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 variances. In contrast to the PCC, CCC takes not only the linear correlation, but also the bias between the two temporal series, i.e., $(\mu_x - \mu_y)^2$, into account. Hence, the value of CCC is within the range of $[-1, 1]$, where ± 1 represents perfect concordance and discordance while 0 means no correlation. In other words, a higher CCC demonstrates a better performance of an evaluated model.

In addition to the aforementioned CCC, another common measure to evaluate emotion regression models is *Root Mean Square Error* (RMSE). Formally, when denoting the gold standard and its correspond prediction as y_t and $\hat{y}_t$ at time t, respectively, the RMSE over the whole sequence period of T can be computed as follows:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}||\hat{y}_t - y_t||^2}. \tag{4.4}$$

As a consequence, a lower RMSE value indicates a better performance.

Finally, before moving to the next session for the experimental evaluation and results analysis with respect to each method separately, it is worthwhile to mention that, to further access the significance level of performance improvement, a statistical evaluation was carried out over the whole predictions between the proposed modal and the corresponding baseline models or other benchmark methods, by means of *Fisher's r-to-z transformation* [36]. Unless stated otherwise, a p value less than .05 indicates significance.

4.3 Emotion Prediction with Deep Crossmodal Latent Representations

In this section, the focus is to evaluate the crossmodal training approach proposed in Section 3.1.1 which aims at improving mono-modal emotion recognition via crossmodal emotion embedding. To this end, extensive experiments are carried out on two multimodal emotional datasets for two tasks, respectively. Specifically, the RECOLA dataset (cf. Section 4.1.1) was chosen for dimensional emotion regression, whereas the OMG-Emotion dataset (cf. Section 4.1.3) was selected for categorical emotion classification.

4.3.1 Experimental Evaluation

For the RECOLA dataset, the eGeMAPS set of 88 acoustic features were extracted. Meanwhile, as to the visual features, both the appearance-based and geometric-based features were employed. Next, for the OMG-Emotion corpus, the eGeMAPS feature set and VGG-Face-based deep visual features were utilised as acoustic and visual descriptors, respectively. For a detailed description of the feature sets, the reader is referred to Section 4.2.1.

Moreover, the proposed EmoBed networks were implemented with GRU-RNNs. In particular, for the RECOLA experiments, the number of hidden layers for the modality-specific subnetworks (i. e., for audio and video) and the modality-shared subnetwork were set to be two, respectively. Each hidden layer has 120 hidden nodes.

To train the network, the Adam optimisation algorithm was applied with an initial learning rate of 0.001. In addition, an L2 regularisation term with a weight decay of 10^{-4} was employed, to improve the model generality. Furthermore, to facilitate the training process, the mini-batch size was set to be four. Also note that, an online standardisation was always applied to the input data by using the means and variations of the training set.

For the OMG-Emotion experiments, the network and the training hyperparameters were kept in line with the RECOLA experiments, but only one hidden layer was implemented as the modality-specific or the modality-shared subnetworks due to the limited size of the OMG-Emotion dataset, and the mini-batch size was 64.

When training the network in a crossmodal scenario, the audio and video data were randomly selected, rather than the aligned data pair across audio and video, as the mini-batch. The advantage of this method is that it does not require the synchronous presence of both modalities in the training phase. This means that one can principally mix the audio-only and video-only databases to complete the network training process.

Moreover, an annotation delay compensation process was carried out on the RECOLA dataset to compensate for the temporal delay between the observable cues, as seen in the recordings, and the corresponding emotion reported by the annotators [130]. This delay was set to be four seconds as suggested in [100, 170], by shifting the gold standard back in time with respect to the features for both arousal and valence. Additionally, the same post-processing chain was carried out on all predictions as in [198, 170] on RECOLA, consisting of smoothing, centring, scaling, and time-shifting. All the modification parameters were optimised on the development set and then applied to the test set.

4.3.2 Performance

For the sake of fair performance comparison, result performance and related analysis will be reported on the two emotional databases with their corresponding standard testbeds of the AVEC 2016 [198] and OMG-Emotion challenges 2018 [13].

4.3.2.1 Results on RECOLA

In these experiments, two visual feature sets and one acoustic feature set were taken into account, as aforementioned in the experimental evaluation section. As a consequence, this leads to two possible feature set combinations, i. e., 'appearance + eGeMAPS' and 'geometric + eGeMAPS', as the inputs of the proposed crossmodal system. Besides, the systems were evaluated on both dimensional *arousal* and *valence* regressions as well. Moreover, for the classic monomodal systems, the training process was on either with only audio or video data. That was achieved by setting α to be 0.0 in Equation (3.7) and Equation (3.8), respectively.

Table 4.3: Performance comparison in terms of CCC for the arousal prediction among the proposed EmoBed systems, related baselines, and other state-of-the-art systems. These results pertain to the experiments conducted on the *dev*elopment and *test* partitions of the RECOLA database. Three feature sets (audio-eGeMAPS A, video-appearance V_{app}, and video-geometric V_{geo}) were employed to evaluate all approaches. The cases where EmoBed systems have a statistical significance of performance improvement over the classic monomodal systems are marked by the "$*$" symbol.

CCC	A (V_{app})		V_{app} (A)		A (V_{geo})		V_{geo} (A)	
	dev	*test*	*dev*	*test*	*dev*	*test*	*dev*	*test*
classic monomodal	.766	.605	.512	.411	.766	.605	.499	.399
joint audiovisual training	.769	.611	.520	.401	.769	.611	.515	.413
crossmodal triplet training	**.795**	.633	.541	.465	**.794**	.632	.512	.397
EmoBed	.792*	**.644***	**.549***	**.475***	.793*	**.639***	**.527***	**.417***
state of the art								
SVR [198]	.796	.648	.483	.343	.796	.648	.379	.272
Feature selection [95]	.800	—	.587	—	.800	—	.173	—
SC+CNN+RNN [19]	.846	—	.346	—	.846	—	—	—
End-to-end [197]	.786	.715	.371	.435	.786	.715	—	—
Curriculum learning [123]	.687	.591	.417	.343	.687	.591	.394	.267
Strength Modelling	.755	.666	.350	.196	.755	.666	—	—
DDAT w/ RE	.807	.694	.539	.437	.807	.694	.544	.400
DDAT w/ PU	.811	.664	.518	.438	.811	.664	.513	.397

Table 4.3 and Table 4.4 present the obtained results for arousal and valence predictions, respectively. From the two tables, it can be seen that the classic monomodal systems outperform the challenge benchmarks that used the SVR model [198] in most cases. One exception is the arousal prediction with audio signals, which probably attributes to the fact that a fixed network structure, rather than the optimised one on the arousal prediction, was employed in these experiments. These results further confirm that GRU-RNNs hold the powerful capability to capture the long-range context dependence.

Furthermore, when jointly training with both audio and video data (cf. Section 3.1.1.2), one may notice that the corresponding monomodal systems (based on either audio-eGeMAPS, video-appearance, or video-geometric) can deliver higher CCCs compared with the classic monomodal systems, in seven out of eight cases for the arousal prediction and six out of eight cases for the valence prediction, respectively. This observation implies that such a joint training process can somewhat transfer shared semantic information from other heterogeneous data to the target modality thanks to the implementations of i) a shared subnetwork and ii) a multi-task learning framework.

Table 4.4: Performance comparison in terms of CCC for the *valence* prediction among the proposed EmoBed systems, related baselines, and other state-of-the-art systems. These results pertain to the experiments conducted on the *dev*elopment and *test* partitions of the RECOLA database. Three feature sets (audio-eGeMAPS A, video-appearance V_{app}, and video-geometric V_{geo}) were employed to evaluate all approaches. The cases where EmoBed systems have a statistical significance of performance improvement over the classic monomodal systems are marked by the "$*$" symbol.

CCC	A (V_{app})		V_{app} (A)		A (V_{geo})		V_{geo} (A)	
	dev	*test*	*dev*	*test*	*dev*	*test*	*dev*	*test*
classic monomodal	.504	.381	.545	.525	.504	.381	.619	**.529**
joint audiovisual training	.505	.393	.546	.511	.513	.395	.622	.527
crossmodal triplet training	**.515**	.404	**.564**	**.517**	.512	.405	.636	.517
EmoBed	.514*	**.434***	**.564***	.516	**.521***	**.439***	**.645***	.520
state of the art								
SVR [198]	.455	.375	.474	.486	.455	.375	.612	.507
Feature selection [95]	.398	—	.441	—	.398	—	.441	—
SC+CNN+RNN [19]	.450	—	.511	—	.450	—	—	—
End-to-end [197]	.428	.369	.637	.620	.428	.369	—	—
Curriculum learning [123]	.159	.174	.446	.419	.159	.174	.300	.269

Moreover, rather than the joint audiovisual training, when performing the triplet constraint across the audio and video modalities (see Section 3.1.1.3), the obtained CCCs (results on the third lines of Table 4.3 and Table 4.4) show that the introduced enhanced monomodal systems again generally offer significant advantages over the classic monomodal systems in most cases (Fisher r-to-z transformation, $p < .05$). For instance, the obtained CCCs on the test set of the audio-based model are boosted from .605 to .633 and .632 for arousal regression, and from .381 to .404 and .405 for valence regression, when respectively integrated with video-appearance and video-geometric feature sets in the training process. This suggests that the implementation of the triplet constraint is helpful to distil emotional discriminative representations, not only in a single modality scenario as found in the latent representation learning approach Section 3.1.1, but also in a cross modalities scenario, which is consistent with previous findings in a single modality scenario [85].

Finally, when simultaneously performing the crossmodal triplet training as well as the joint audiovisual training processes, it can be seen that the EmoBed systems achieve the best performance in most cases, i. e., six out of eight cases for the arousal regression and five out of eight cases for the valence regression. For example, the obtained CCCs on the test set of the audio-based model are boosted from .605 to .644 and .639 for arousal regression, and from .381 to .434 and .439 for valence

Table 4.5: Performance of the proposed EmoBed systems, related baselines, and other reported systems in terms of F1 on the development set of the OMG-Emotion dataset. The cases where EmoBed systems have a statistical significance of performance improvement over the classic monomodal systems are marked by the "$*$" symbol.

F1 [%]	audio	video
classic monomodal	36.5	37.9
joint audiovisual training	40.2	42.1
crossmodal triplet training	40.7	41.0
EmoBed	**41.7***	**43.9***
other approaches		
SVM [13]	33.0	—
RF [13]	—	37.0

regression, when respectively integrating with video-appearance and video-geometric feature sets in the training process. Besides, the best CCCs achieved by the video-based models reach to .475 and .417 with the appearance and geometric feature sets, respectively, with absolute CCC increases of .064 and .018 compared with the classic monomodal systems for arousal regression. Although such an observation cannot be found for the video-based valence regression models on the test set, this exception possibly attributes to the distribution mismatch between the development and test partitions. Therefore, it is concluded that the proposed EmoBed can largely supply additional knowledge from audio signals to alleviate the shortage of video signals, and vice verse.

Meanwhile, as presented in Table 4.3 and Table 4.4, the EmoBed systems achieve comparable or even better performance to other state-of-the-art methods, such as the winner systems of AVEC 2015 [95] and 2016 [19], end-to-end systems [197], and curriculum learning systems [123].

4.3.2.2 Results on OMG-Emotion

For experiments on the OMG-Emotion database, seven-class categorical emotion classification tasks were conducted on audio and visual signals. Table 4.5 presents the performance of the models in terms of F1 on the development set only, since the annotations of the test set are not publicly available. From the table, one may see that on this database, the classic monomodal models outperform the other methods reported in the literature [13], i. e., Support Vector Machine (SVM) and Random Forests (RF). More specifically, the classic monomodal models yield higher F1 than SVM (36.5 % vs 33.0 %) for audio, and than RF (37.9 % vs 37.0 %) for video.

Additionally, comparing the proposed joint audiovisual training models with the classic monomodal systems, it is noticed that the former approach outperforms the latter one by a large margin, i. e., 40.2 % vs 36.5 % for audio and 42.1 % vs 37.9 % for video. These experimental results again indicate that the proposed joint audiovisual training approach is plausible to promote performances of monomodal emotion classification. Furthermore, similar results were also obtained when utilising the triplet training approach to distil the salient representations across multiple modalities. Nevertheless, the highest F1s are achieved by means of the EmoBed systems, which deliver 5.2 % and 6.0 % absolute performance gain compared with the classic monomodal systems when using audio or video signals, respectively. All these observations further confirm the findings discovered from the RECOLA database.

4.3.2.3 Visualisation of Emotion Embeddings

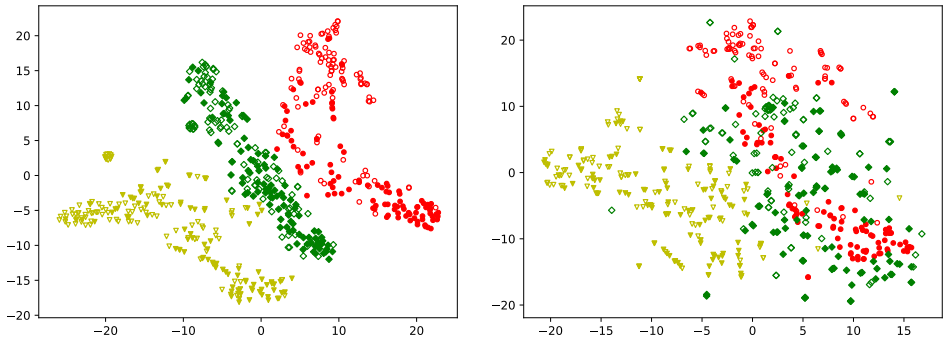
In the following, the focus is investigating how the proposed crossmodal learning framework benefit for emotion recognition. For this purpose, the learnt representations were extracted from the classic monomodal systems and the proposed EmoBed systems, respectively. Figure 4.2 illustrates the distribution of the learnt representations on the development set of the RECOLA database by means of t-Distributed Stochastic Neighbour Embedding (t-SNE). Obviously, it can be seen that with the classic monomodal systems, the learnt representations can be easily distinguished into three parts by the modalities they stem from, in either arousal (cf. Figure 4.2 (a)) or valence (cf. Figure 4.2 (c)) prediction. Specifically, the representations learnt from different modalities almost have no overlap albeit they belong to the same emotional states. In striking contrast, the representations extracted from EmoBed systems are visibly clustered together based on their emotional properties (cf. Figure 4.2 (b) and (d) for arousal and valence, respectively).

Such an observation is even more noticeable on the OMG-Emotion database (cf. Figure 4.3). Note that, for the sake of simplicity, only two emotional categories, i. e., happy and sad, are selected for visualisation. Likewise, one can find that the representations belonging to the same emotional category share almost the same latent space.

These findings indicate that the representations learnt by the proposed EmoBed are somewhat invariant to the modalities. By making use of the emotion embedding space, the emotional representations extracted from audio and video signals are able to implicitly fuse the knowledge from each other. Thus, the exploitation of mutual information possibly leads to performance improvement for a monomodal system.

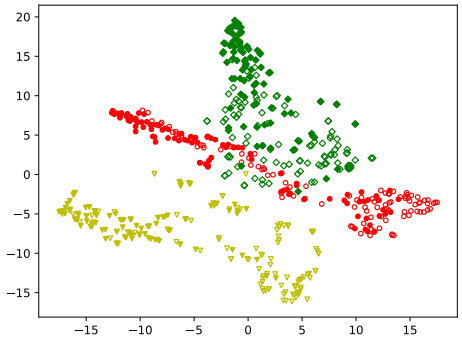
4.3.2.4 Impact of Auxiliary Modalities and Triplet Loss

To demonstrate the importance of learning from auxiliary modalities for the monomodal emotion recognition system, the impact of weight change with respect

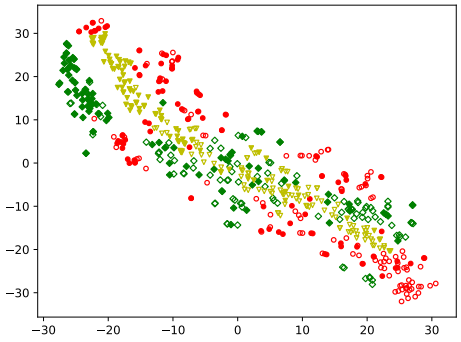


(a) arousal (classic monomodal)

(b) arousal (EmoBed)



(c) valence (classic monomodal)

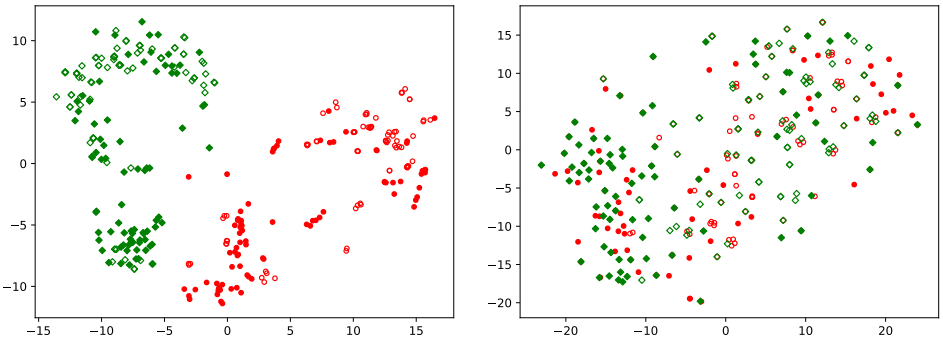


(d) valence (EmoBed)

Figure 4.2: Visualisation of the learnt representations of the development set of the RECOLA database when using the proposed EmoBed systems or the classic monomodal systems. Red, green, and yellow markers: representations from audio (eGeMAPS), video (appearance), and video (geometric) modalities; solid and hollow markers: high and low arousal/valence.

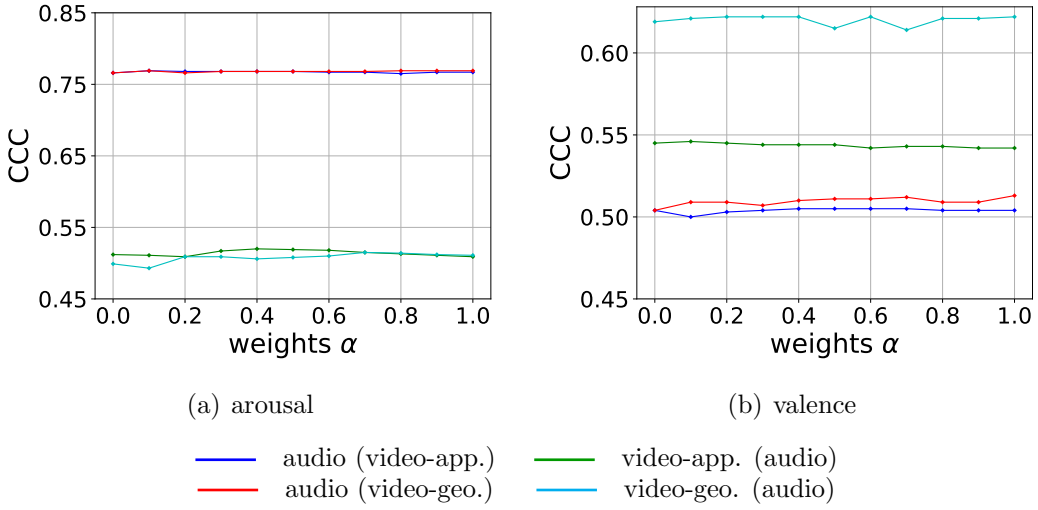
to the counterpart modality will be investigated separately in this section, when in a joint audiovisual training process or in a crossmodal triplet training process.

Figure 4.4 depicts the relationship between the obtained CCCs and the weight α (cf. Equation (3.7) and Equation (3.8)) on the RECOLA database. It is noted that the model performance is improved when the weight increases to some values for the video-based arousal regression models (green and cyan lines in Figure 4.4 (a)). Similar observations can be made as well for the audio-based valence regression models (blue and red lines in Figure 4.4 (b)). Therefore, this behaviour again indicates that learning from other modalities indeed can help the enhancement of traditional monomodal systems. Yet, it is also noted that the audio-based arousal and the video-based valence regression models almost remain without obvious performance improvement. This might suggest that transferring the information from



(a) classic monomodal (b) EmoBed

Figure 4.3: Visualisation of the learnt representations of the development set of the OMG-Emotion database when using the proposed EmoBed systems or the classic monomodal systems. Red and green markers: representations from audio and video modalities; solid and hollow markers: happy and sad categories.

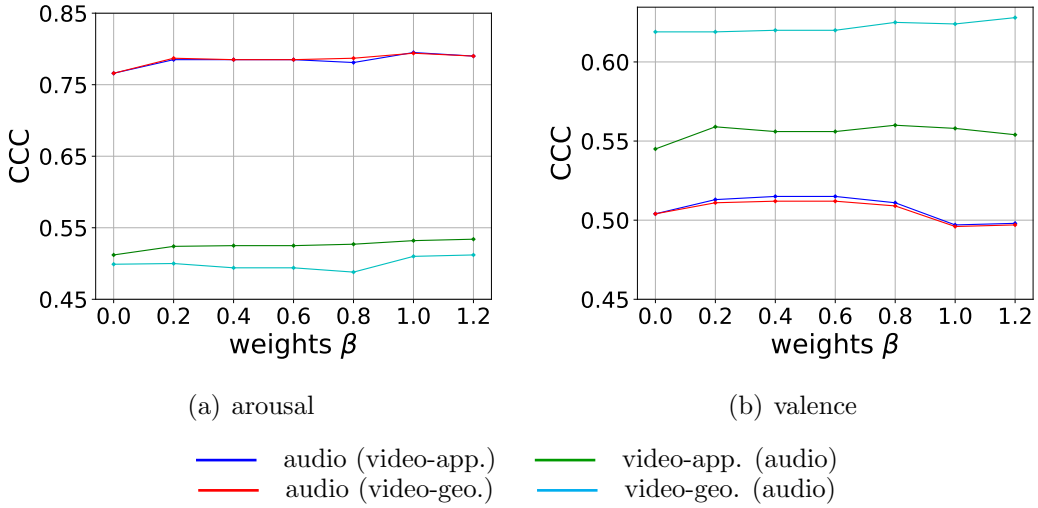


(a) arousal (b) valence

Figure 4.4: Impact of the joint auxiliary modality loss on the *joint audiovisual training* systems for either arousal (a) or valence (b) regression with the RECOLA database.

the modality with richer knowledge to the one with sparse knowledge is much easier than the other way around, as audio signals often lead to higher CCCs for arousal regression while video signals for valence regression.

Further, Figure 4.5 illustrates the relationship between the obtained CCCs and the weight β (cf. Equation (3.18)) on the RECOLA database. Obviously, it can be seen that the obtained CCCs remarkably grow with the increase of weight β in



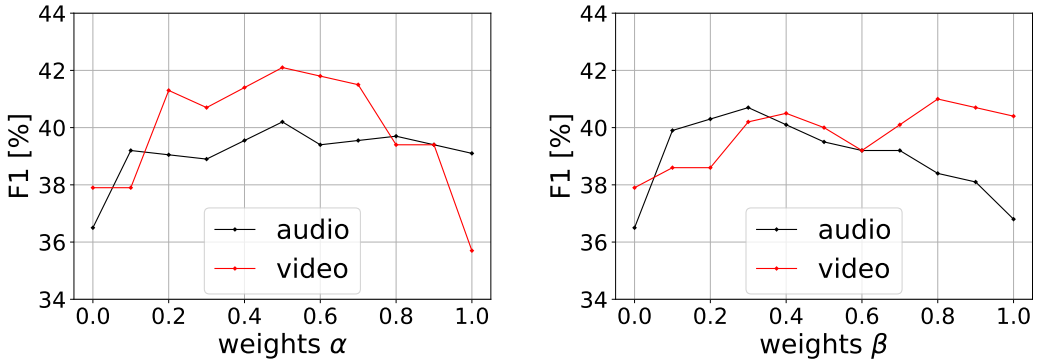
(a) arousal (b) valence

Figure 4.5: Impact of the crossmodal triplet loss on the *crossmodal triplet training* systems for either arousal (a) or valence (b) regression with the RECOLA database.

all cases for arousal (cf. Figure 4.5 (a)) and valence (cf. Figure 4.5 (b)) regression. Specifically, when $\beta = 1.0$, i. e., the triplet training contributes equally to the traditional emotion regression training, the systems yield the best CCCs in all cases for arousal regression. Nevertheless, the audio- and video-based valence regression systems deliver the best CCCs only when $\beta = 0.4$ and $\beta = 0.8/1.2$, respectively. The lower contribution from triplet loss implies that it might be more difficult to distil the valence-salient representations than the arousal-salient representations by means of triplet training.

Moreover, a similar investigation was conducted on the OMG-Emotion database for categorical emotion classification. Figure 4.6 explicitly quantifies the contributions of joint audiovisual training (a) and crossmodal triplet training (b) when in a crossmodal learning framework. For a joint audiovisual training system, when $\alpha = 0.0$, i. e., no contribution from the auxiliary modality, the model is learnt based on only the loss of each modality, separately. When α increases, i. e., the contribution of the auxiliary modality during training increases, the performance of monomodal emotion recognition (audio or video) is improved first, until a point where the contribution of the auxiliary modality might actually penalise the learning objective too much and even harm the learning of the main modality, and thus performances start to decrease. Similar observations can be found for crossmodal triplet training systems.

To this end, proper values of the weight α and β need to be identified for the tasks at hand. It can be observed from the figures that, the best performance for both audio and video emotion classification is reached when $\alpha = 0.5$ in joint audiovisual training systems; whereas the best performance for audio and video emotion



(a) joint auxiliary modality loss

(b) crossmodal triplet loss

Figure 4.6: Impact of the joint auxiliary modality loss on the *joint audiovisual training* systems (a), and impact of the crossmodal triplet loss on the *crossmodal triplet training* systems (b), with the OMG-Emotion database.

classification is achieved when $\beta = 0.3$ and $\beta = 0.8$, respectively, in crossmodal triplet training systems.

4.3.3 Summary

Different from prior related works on either the classic monomodal systems or multimodal systems, in this study, a novel monomodal emotion recognition system which exploits the information across auxiliary modalities during the training phase is presented. To implement this system with audio and visual modalities, on one hand, an emotion recognition network is shared for both audio and video signals, so that the complementary information from an auxiliary modality can be implicitly transferred to the target modality. On the other hand, a triplet constraint is applied over acoustic and visual representations to distil emotional embeddings that are less variant to the modalities. The proposed EmoBed systems were systematically evaluated on the two benchmark databases RECOLA and OMG-Emotion, and experimental results have demonstrated that it can significantly improve the prediction performance of a monomodal system, by fusing an additional modality in the training process.

Albeit the effectiveness, this framework could be further developed in the future. For example, in the triplet training process, the annotation uncertainty information could be utilised as a new distance measure between the learnt representations. Besides, it is also worth to train the model by using large-scale heterogeneous datasets from a variety of domains.

4.4 Emotion Regression Based on Deep Bag-of-X-Words

In this present section, experiments are performed to evaluate the effectiveness of the proposed deep Bag-of-X-Words in Section 3.1.2 on the RECOLA dataset for speech emotion regression. More specifically, in this study, MFCC 0-12 and the logarithmic energy are extracted from the openSMILE toolkit [61], which are exactly the same as the ones applied in the previous framework in [170].

4.4.1 Experimental Evaluation

To demonstrate the effectiveness of the proposed method for emotion regression, SVRs were implemented in the LIBLINEAR toolkit [62] with a linear kernel and trained with an L2-regularised L2-loss dual solver. Also, the complexity C was optimised on the development set in the range of $[10^{-5}, 10^{0}]$.

To optimise the model to deliver better performance, a grid-search was further conducted over the parameters of the BoCAW (cf. Section 3.1.2.2) including the local window size (W_1) and the time step size (Ts_1) in stage 1. More specifically, a best setting was determined on the best performance achieved on the development set by a grid search over $[0.01, 0.02, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6]$ for W_1 and $[0.05, 0.1, 0.2, 0.4]$ for Ts_1 when training (Ts_1 for development and test is fixed to be 40 ms to match the granularity of the annotations). Furthermore, for a fair comparison with prior findings in [170], other hyperparameters of the model were fixed, maintaining the settings for the codebook size (C_s=1000), number of assignments (N_a=20), and the global window size (W_2=8.0 s) and the time step size (Ts_2=800 ms for training to achieve a fast process or 40 ms for development and test to match the granularity of the annotations) of the second stage.

Additionally, it is noteworthy that, similar annotation delay compensation and the post-processing process chain were performed for continuous emotion regression on RECOLA, as aforementioned in Section 4.3.

4.4.2 Performance

Two regression tasks have been investigated on the RECOLA dataset, i. e., arousal and valence prediction. Table 4.6 shows result performances in terms of CCC for the proposed BoCAW features. Note that, as described in Section 3.1.2, it is of crucial importance to select the parameter settings when generating the bag-of-audio-words in the first stage for achieving the optimal performance. Therefore, experiment results are presented in the table, on both the development and test sets over different window sizes W_1 for arousal and valence, respectively. It can be seen from the table that, the best CCCs on the development set for the arousal and

Table 4.6: Performances in terms of Concordance Correlation Coefficient (CCC) of the proposed BoCAW features with various window sizes in the first stage (W_1), for both *arousal* and *valence* regressions, evaluated on the *dev*lopment and *test* partitions. Note that, for each W_1, only the best performance among four examined time step sizes (Ts_1) is reported, by calculating the averaged predictions of arousal and valence on the development set. The best results achieved are highlighted. The symbol of $*$ indicates the significance of the performance improvement over the bag-of-audio-words (BoAW) baseline method.

settings		*arousal*		*valence*	
$W_1(s)$	$Ts_1(s)$	*dev*	*test*	*dev*	*test*
BoAW [170]		.789	.738	.550	.430
0.01	0.1	.791	.746*	.557*	.432
0.02	0.1	.791	.753*	.581*	**.497***
0.05	0.1	**.800***	.750*	.572*	.463*
0.1	0.2	.797*	**.757***	**.603***	.465*
0.2	0.4	.787	.752*	.546	.455*
0.4	0.2	.780	.747*	.543	.492*
0.8	0.4	.775	.738	.540	.459*
1.6	0.4	.765	.733	.532	.423

valence dimensions are .800 and .603, respectively, and the best results obtained on the test set are .757 for arousal and .497 for valence.

Moreover, the results also imply that predictions of arousal and valence are differently influenced by the length of W_1. Therefore, to better illustrate the effect of W_1 for the prediction of emotions, the performance (in CCC) averaged over all four selected time step sizes Ts_1 for each predefined window size are computed, as shown in Figure 4.7. When W_1=0.01 s, i.e., only one frame is included in each segment on stage 1, then, the steps conducted on stage 1 equal to quantising the original features, delivering only a slight improvement. Then, when the window size increases, i.e., an increasing number of frames are contained in a segment, the performance of emotion prediction improves until a point where information of different emotional nature is contained in the window, and thus performance starts to decrease. To this end, it is essential to identify a proper analysis window size W_1 for the task at hand. Notably, one can observe from the figure that, the best window size is 0.05 s for arousal, whereas the best performance for valence is obtained with a longer window (0.1 s). This result is coherent with other findings in the literature [159, 170], and confirms that more context information is essential for valence than arousal when generating context-aware bags. Interestingly, as for human annotators, people are also slower to give valence ratings, compared to arousal [143].

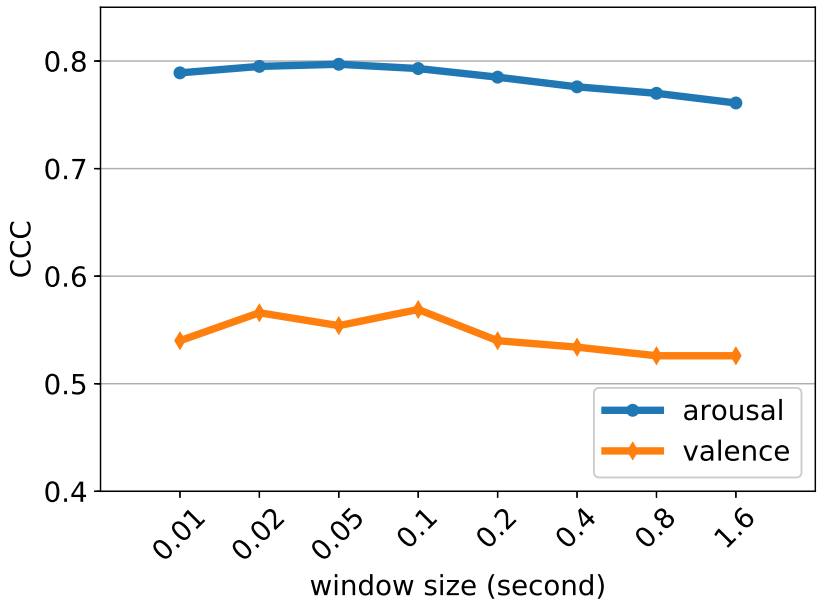


Figure 4.7: The effect of the sub-bag's window size on the performance (CCC) when predicting arousal and valence separately. Performances are averaged over all examined time step sizes on the development partition.

Additionally, to further highlight and substantiate the advantages of the BoCAW approach, the best performance it achieved on the RECOLA dataset is compared with seven other systems from the state-of-the-art. For instance, in [206], CCC was exploited as the cost function instead of standard mean squared error when training a deep model, whereas an end-to-end framework that learns representations directly from raw signals was implemented in [197]. Additionally, by compensating the weakness of a model itself or incorporating the strength of different models, the Reconstruction-Error-based (RE-based) learning framework and prediction-based learning framework were proposed in [90] and [89], respectively. More recently, an adversarial training approach was investigated for emotion regression problems in [86] for the first time. The last framework to compare with is obviously BoAW [170], which is the fundamental of this work as well. A comparison of the best performances of all aforementioned approaches and the BoCAW on the RECOLA dataset is presented in Table 4.7. It can be seen that, when using BoCAW representations, the CCC performance is significantly improved for both arousal and valence predictions compared to the original BoAW framework.

4.4.3 Summary

Summing up, the proposed Bag-of-Context-Aware-Words representation ameliorates the conventional Bag-of-Audio-Words representation with context information main-

Table 4.7: Performances in terms of CCC of the proposed method comparing with other state-of-the-art approaches on the RECOLA dataset. The best results achieved are highlighted. The symbol of $*$ indicates the significance of the performance improvement over the bag-of-audio-words (BoAW) method.

approaches	*arousal*		*valence*	
	dev	*test*	*dev*	*test*
state of the art				
CCC-objective [206]	.412	.350	.242	.199
end-to-end [197]	.752	.699	.406	.311
RE-based [90]	.785	.729	.378	.360
prediction-based [89]	.774	.744	.440	.393
adversarial training [86]	.797	.737	.501	.455
BoAW [170]	.789	.738	.550	.430
proposed				
BoCAW	.800*	**.750***	**.603***	.465*

tained on segment-level features, by making use of a hierarchical framework. In this framework, BoAW is first applied on a sequence of segments, and then, these segment-level features are fed into a second BoAW layer to extract an higher-level representation of the information captured in the first stage. Evaluations have been conducted on the RECOLA database to assess the system performance. Results show that the proposed BoCAW features obtain state-of-the-art performance for emotion regression from audio data. In future work, the proposed method can be extended to be applied to visual features for facial expression detection tasks. Further, the proposed BoCAW representations based on segments are also applicable to other pattern recognition tasks where a specific pattern lasts a period of time, e. g., laughter detection and engagement recognition.

4.5 Strength Modelling-based Emotion Recognition

In the present section, experiments will be conducted with respect to the strength modelling described in Section 3.2.1. For evaluation purposes, the RECOLA database introduced in Section 4.1.1 will be employed. In particular, the strength modelling training strategy will first be performed in audio-only and video-only settings for dimensional emotion regression. Then, the proposed method will be incorporated with early and late fusion strategies to investigate its robustness in the multimodal settings, as provided in Section 3.2.1.2.

4.5.1 Experimental Evaluation

For acoustic features, arithmetic means and the coefficient of variances of 13 LLDs are computed, resulting in 26 original features per functional window. Considering the video data, 49 features are obtained by applying PCA on the sequential 316 geometric-based features.

To demonstrate the effectiveness of the proposed strength modelling training approach, the baseline experiments were conducted, where the SVR or BLSTM-RNNs models were individually trained on the modalities of audio, video, or the combination, respectively. Specifically, the selected SVR was implemented in the Liblinear toolkit [62] with linear kernel, and trained with L2-regularised L2-loss dual solver. The tolerance value of ε was set to be 0.1, and the 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, as listed in Table 4.8. For the BLSTM-RNNs, two bidirectional LSTM hidden layers were chosen, with each layer consisting of the same number of memory blocks (nodes). The number was optimised as well by the development set for each modality and task among [40, 60, 80, 100, 120], also as shown in Table 4.8. During the network training process, gradient descent was implemented with a learning rate of 10^{-5} and a momentum of 0.9. Zero mean Gaussian noise with standard deviation 0.2 was added to the input activations in the training phase so as to improve generalisation. All weights were randomly initialised in the range from -0.1 to 0.1. Finally, the early stopping strategy was used as no improvement of the mean square error on the validation set has been observed during 20 epochs or the predefined maximum number of training epochs (150 in this case) has been executed. Furthermore, to accelerate the training process, the network weights were updated after running every mini-batch of 8 sequences for computation in parallel.

Following the naming conventions adopted in Section 3.2.1.1, the models trained with baseline approaches are referred to as *individual* models, whereas the ones associated with the proposed approaches are denoted as *strength-involved* models henceforth in the article. For the sake of fair performance comparison and computational demand reduction, those optimised parameters of individual models, i. e., SVR or BLSTM-RNN, were further applied to the corresponding strength-involved models, namely the *S*-*B*, *B*-*S*, and *B*-*B* models, respectively.

Similar as performed in Section 4.3, annotation delay compensation of four seconds was also executed to compensate for the temporal delay between the observable cues and the corresponding emotion reported by the annotators [130]. Notably, this is also suggested as well in other related work in [100, 198]. To this end, the gold standard was shifted back in time with respect to the features, in all experiments presented.

Finally, note that all fusion experiments require concurrent initial predictions from audio and visual modalities. However, in some circumstances, visual prediction

Table 4.8: The optimised complexity (C) of SVR and number (N) of hidden nodes per layer of BLSTM-RNN for different types of modality and task.

	C		N	
modality	*arousal*	*valence*	*arousal*	*valence*
audio	0.00005	0.0002	40	40
video	0.02	0.002	80	80
audio & video	0.00005	0.1	100	100

Table 4.9: Performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) on the *development* and *test* partitions from the *audio* signals. The best achieved CCCs are highlighted. The symbol of $*$ indicates the significance of the performance improvement.

	development				*test*			
	arousal		*valence*		*arousal*		*valence*	
method	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC
S	.126	.714	.149	.331	.133	.605	.165	.248
B	.142	.692	.117	.286	.155	.625	.119	.282
B-S	.127	.713	.144	.348*	.133	.606	.160	.264
S-B	.122	.753*	.113	.413*	.133	.665*	.117	.319*
B-B	.122	**.755***	.112	**.476***	.133	**.666***	.123	**.364***

cannot occur where a face has not been detected. For all fusion experiments where this took place, the initial corresponding audio predictions were replicated to fill the missing video slots.

4.5.2 Performance

The first experiment analyses the predictive power of the strength-involved models on audio signals. Table 4.9 displays the results in terms of both RMSE and CCC obtained from the strength-involved models and the individual models of SVR and BLSTM-RNN on the development and test partitions from acoustic features. The three Strength Modelling setups either matched or outperformed their corresponding individual models. This observation implies that the advantages of each model are enhanced via Strength Modelling. In particular, the performance of the BLSTM-

Table 4.10: Performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) on the *development* and *test* partitions from the *video* signals. The best achieved CCCs are highlighted. The symbol of $*$ indicates the significance of the performance improvement.

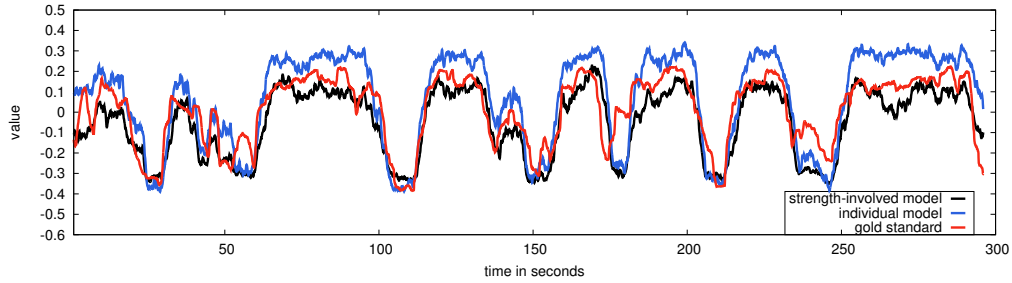
	development				*test*			
	arousal		*valence*		*arousal*		*valence*	
method	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC
S	.197	.120	.139	.456	.186	.193	.156	.381
B	.184	.287	.110	.478	.183	.193	.122	.394
B-S	.183	.292	.110	**.592***	.176	**.265***	.130	.464*
S-B	.186	**.350***	.118	.510*	.186	.196	.121	**.477***
B-B	.185	.344*	.113	.501*	.197	.184	.120	.459*

RNN model, for both arousal and valence, was significantly boosted by the inclusion of SVR predictions (S-B) on the development and test sets. This improvement could be due to the initial SVR predictions helping the subsequent RNN avoid local minima.

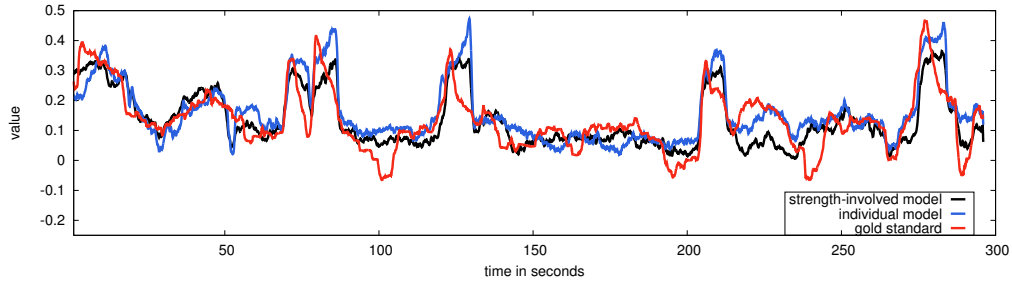
Similarly, the B-S combination brought additional performance improvement for the SVR model, although not as obvious as for the S-B model. Again, it may imply that the temporal information leveraged by the BLSTM-RNN is being exploited by the successive SVR model. The best results for both arousal and valence dimensions were achieved with the framework of B-B, which achieved relative gains of 6.5 % and 29.1 % for arousal and valence receptively on the test set when compared to the single BLSTM-RNN model (B). This indicates there are potential benefits for audio-based affect recognition by the deep structure formed by combining two BLSTM-RNNs using the Strength Modelling framework.

Next, similar experiments were also conducted on the video feature set, and results are presented in Table 4.10. As for valence, the three Strength Modelling setups either match or outperform their corresponding individual models, with the highest CCC of .477 obtained on test set achieved via the S-B model. As expected, one can observe that the models (individual or strength-involved) trained using only acoustic features is more superior for interpreting the dimension of arousal rather than valence. Whereas, the opposite observation is seen for models trained only on the visual features. This finding is in agreement with the conclusion found in the literature [159].

Additionally, Strength Modelling achieved comparable or superior performance to other state-of-the-art methods in the literature. In [100], a development set



(a) arousal predictions via audio signals



(b) valence predictions viavideo signals

Figure 4.8: Automatic prediction of arousal via audio signals (a) and valence via video signals (b) obtained with the best settings of the *strength*-involved models and *individual* models for a subject from the test partition.

arousal CCC of .689 was obtained using an OA-RVM based on audio features. Whilst in [126] a development set valence CCC of .510 was reported using an OA-RVM trained on video features. For comparison, the best development set arousal and valence CCC obtained with the proposed Strength Modelling framework were 0.755 and 0.510 for the audio and video modalities respectively.

To further highlight advantages of Strength Modelling, Figure 4.8 illustrates the automatic predictions of arousal via audio signals (a) and valence via video signals (b) obtained with the best settings of the strength-involved models and the individual models frame by frame for a single test subject from RECOLA. Note that, similar plots were observed for the other subjects in the test set. In general, the predictions generated by the proposed Strength Modelling approach are closer to the gold standard, which consequently contributes to better results in terms of CCC.

Moreover, further experiments were conducted with fused audiovisual features. Table 4.11 presents the performance of both the individual and strength-involved models integrated with the early fusion strategy. In most cases, the performance of the individual models of either SVR or BLSTM-RNN was significantly improved

Table 4.11: Performance comparison in terms of RMSE and CCC between the *strength*-involved models and the *individual* models of SVR (S) and BLSTM-RNN (B) with *early fusion strategy* on the development and test partitions. The best achieved CCCs are highlighted. The symbol of $*$ indicates the significance of the performance improvement.

	development				*test*			
	arousal		*valence*		*arousal*		*valence*	
method	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC
S	.121	**.728**	.113	.544	.132	.610	.139	.463
B	.132	.700	.109	.513	.148	.562	.114	.476
B-S	.122	.727	.118	**.549**	.132	.610	.121	**.520***
S-B	.127	.712	.096	.526	.144	.616*	.112	.473
B-B	.126	.718*	.095	.542*	.143	**.618***	.114	.499*

with the fused feature vector for both arousal and valence dimensions in comparison to the performance with the corresponding individual models trained only on the unimodal feature sets (Section 3.2.1).

For the strength modelling systems, the early fusion B-S model generally outperforms the equivalent SVR model, and the structure of S-B outperforms the equivalent BLSTM-RNN model. However, the gain obtained by Strength Modelling with the early fused features is not as obvious as that with individual models. This might be due to the higher dimensions of the fused feature sets which possibly reduce the weight of the predicted features.

Lastly, the feasibility of integrating Strength Modelling into three different late fusion strategies are investigated, i. e., modality-based, model-based, and the combination (see Section 3.2.1.2). A comparison of the performance of different fusion approaches, with or without Strength Modelling, is presented in Table 4.12. For systems without Strength Modelling, one can observe that best individual model test set performances, .625 and .394 in CCC, for arousal and valence respectively (Section 3.2.1) are boosted to .671 and .405 with the modality-based late fusion approach, and to .651 and .497 with the model-based late fusion approach. These results are further promoted to .664 and .549 when combining the modality- and model-based late fusion approaches. This result is in line with other results in the literature [163, 100], and again confirms the importance of multimodal fusion for affect recognition.

Interestingly, when incorporating Strength Modelling into late fusion one may observe significant improvements over the corresponding non-strength setups. This finding confirms the effectiveness and the robustness of the proposed method for

Table 4.12: Performance comparison in terms of RMSE and CCC between the *str*ength-involved models and the *ind*ividual models of SVR (S) and BLSTM-RNN (B) with *late fusion strategies* (i.e., modality-based, model-based, or the combination) on the *dev*elopment and *test* partitions. The best achieved CCC is highlighted. The symbol of $*$ indicates the significance of the performance improvement.

	arousal				*valence*			
	dev		*test*		*dev*		*test*	
fusion type	RMSE	CCC	RMSE	CCC	RMSE	CCC	RMSE	CCC
a. *modality*-based								
A&V; *S*	.117	.777	.134	.654	.128	.493	.149	.386
A&V; *B*	.126	.736	.134	.671	.104	.475	.113	.405
A&V; *B*-*S*	.114	**.791***	.130	.668	.090	**.664***	.105	**.542***
A&V; *S*-*B*	.117	.778	.128	**.681***	.096	.586*	.105	.495*
A&V; *B*-*B*	.117	.779*	.130	.680*	.095	.601*	.106	.506*
b. *model*-based								
A; *ind*	.119	.771	.132	.651	.112	.335	.117	.284
V; *ind*	.179	.230	.172	.184	.096	.588	.110	**.497**
A; *str*	.117	**.778***	.132	**.664***	.108	**.409***	.120	**.303***
V; *str*	.171	**.344***	.171	**.222***	.095	**.599***	.111	.477
c. *modality*- and *model*-based								
A&V; *ind*	.113	.795	.130	.664	.089	.670	.107	.549
A&V; *str*	.110	**.808***	.127	**.685***	.088	**.671**	.103	**.554**

multimodal continuous affect recognition. In particular, the best test CCCs, .685 and .554, are obtained by the strength-involved models integrated with the modality- and model-based late fusion approach. This arousal result matches performance with the AVEC 2016 affect recognition subchallenge baseline system, .682, which was obtained using a late fusion strategy involving eight separate modalities [198].

Further, the obtained results are competitive when compared with the results from the OA-RVM systems presented in [100], where the best test CCCs (.740 for arousal and .580 for valence) were achieved using four modalities and a non-casual setup. Thus, testing the suitability of other modalities, ECG and EDA in particular, in conjunction with Strength Modelling will be a key area of future research efforts.

4.5.3 Summary

In conclusion, in this study a novel framework, i. e., Strength Modelling, has been proposed and investigated for continuous audiovisual affect recognition. In a Strength Modelling-based model, the strength of an initial model, represented by its predictions, is concatenated with the original features to form a new feature set, which is then used as the basis for regression analysis in a subsequent model.

To demonstrate the suitability of the approach, the benefits from two state-of-the-art regression models, namely SVR and RNN, were explored. Results obtained on the benchmark database RECOLA indicate that Strength Modelling can offer advantages over the corresponding conventional individual models when performing emotion recognition. A further advantage of Strength Modelling is that it can be implemented as a plug-in for use in both early and late fusion stages.

In future work, there is a wide range of possible research directions associated with Strength Modelling to build on this initial set of promising results. In particular, much of the future efforts will concentrate around assessing the suitability of additional individual models for use in the framework, and exploring the advantages that Strength Modelling could bring when predicting on other modalities. Moreover, it is also of interest to further explore the promising advantages offered by Strength Modelling, by investigating its reliability and effectiveness on other affective datasets and other behavioural regression tasks.

4.6 Emotion Regression via Dynamic Difficulty Awareness Training

This section will give an in-depth analysis of the experiments conducted to evaluate the proposed Dynamic Difficulty Awareness Training (DDAT) framework, which is described in Section 3.2.2. In contrast to Strength Modelling models which integrate advantages of distinct models, the DDAT framework focuses on exploiting the weakness of the model itself. With this goal, perception uncertainties and reconstruction errors can be utilised as the difficulty indicators to improve the learning process. To demonstrate the effectiveness of this approach, experiments were performed on the RECOLA dataset.

4.6.1 Experimental Evaluation

To represent the audio data, the established eGeMAPS set of 88 acoustic features were extracted over a fixed window of 8 s with a step size of 40 ms. Next, as to the visual features, both the appearance-based and geometric-based features were considered. Following that, an online standardisation was applied to the extracted features by using the means and variations of the training set.

The implemented DDAT framework consists of a deep RNN equipped with gated recurrent units (GRUs). As an alternative to the LSTM units, GRUs can also capture the long-term dependencies in sequence-based tasks and mitigate the effects of the vanishing gradient problem [33]. Compared to LSTM units, GRUs have fewer parameters due to the fact that they do not have separate memory cells and output gates, which results in a faster training process and a less-training-data demand for achieving a good generalisation. Most importantly, many empirical evaluations [102] have indicated that GRUs perform as competitively as LSTM units.

Aiming to improve the prediction performance, the RNN structure was optimised in terms of the number of hidden layers and the number of GRUs per layer in the development phase. To this end, a search grid was applied over {1, 3, 5, 7, 9} hidden layers and {40, 80, 120} hidden units per layer. For each learning strategy, the network structure with the best performance in terms of CCC was chosen in order to alleviate the impact of the variation of network structures on the system performance. The training of the models was conducted using the Adam optimisation algorithm [110] with an initial learning rate of 0.001.

Finally, to refine the obtained prediction, the same chain of post-processing was carried out, as suggested in [198]. In detail, the filtering window size (ranging from 0.12 s to 0.44 s at a rate of 0.08 s) and the time-shifting delay (ranging from 0.04 s to 0.60 s at a step of 0.04 s) were optimised using a grid search method. All these post-processing parameters were optimised on the development set and then applied to the test set. Therefore, those post-processing parameters had various settings for different tasks.

4.6.2 Performance

Table 4.13 shows the obtained CCCs (after post-processing) of the development and test partitions for both arousal and valance predictions. The single-task learning (baseline), MTL, and the proposed DDA training systems are compared through three individual information streams, i. e., one acoustic feature set (eGeMAPS) and two visual feature sets (appearance and geometric).

For the baseline system, the obtained results are competitive to, or even better than, the benchmark of the emotion prediction subchallenge in the AVEC 2016 [198] over three information streams and two prediction tasks. These results support previous findings showing that GRUs can deliver competitive performance when compared to LSTM units [33].

When training the networks jointly with input reconstruction (RE-based MTL) or perception uncertainty prediction (PU-based MTL), one can observe that the systems slightly outperform the baseline systems in nine out of twelve cases on the test set. This indicates that there is a substantial relationship between the two jointly learnt tasks. To be more specific, the representations from the last neural network hidden layer, which are learnt synchronously from the emotion prediction

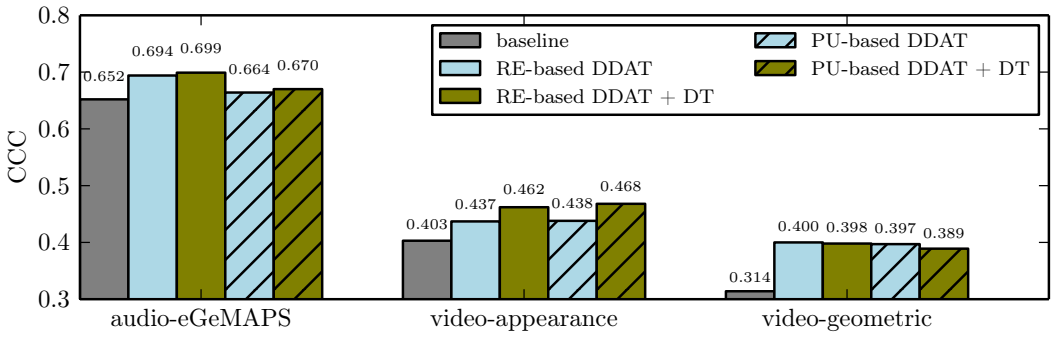
Table 4.13: System performance comparison in CCC for the conventional single-task learning (baseline) framework, the multi-task learning (MTL) framework, and the proposed Dynamic Difficulty Awareness Training (DDAT) framework using reconstruction error (RE, a vector as v or a scalar of sum as s) and perception uncertainty (PU) variants. These results pertain to the experiments conducted on the *dev*elopment and *test* partitions for both arousal and valence targets. Three feature sets (audio-eGeMAPS, video-appearance, and video-geometric) were employed to evaluate all approaches. The best results achieved on the test set are in bold. The cases where DDAT has a statistical significance of performance improvement over MTL are marked by the "$*$" symbol.

task	method	audio-eGeMAPS		video-appearance		video-geometric	
		dev	*test*	*dev*	*test*	*dev*	*test*
arousal	baseline	.783	.652	.528	.403	.523	.314
	RE-based MTL	.788	.629	.512	.425	.502	.324
	PU-based MTL	.803	.654	.502	.406	.508	.327
	DDAT w/ RE(v)	.806*	.676*	.533*	.434	.559*	.355*
	DDAT w/ RE(s)	.807*	**.694***	.539*	.437*	.544*	**.400***
	DDAT w/ PU	.811	.664*	.518*	**.438***	.513	.397*
valence	baseline	.473	.400	.493	.404	.620	.417
	RE-based MTL	.519	.331	.529	.366	.632	.488
	PU-based MTL	.506	.416	.468	.418	.643	.452
	DDAT w/ RE(v)	.517	.378*	.520	.329	.634	.473
	DDAT w/ RE(s)	.508	**.422***	.528	**.457***	.639	.471
	DDAT w/ PU	.498	.407	.514*	.431*	.632	**.501***

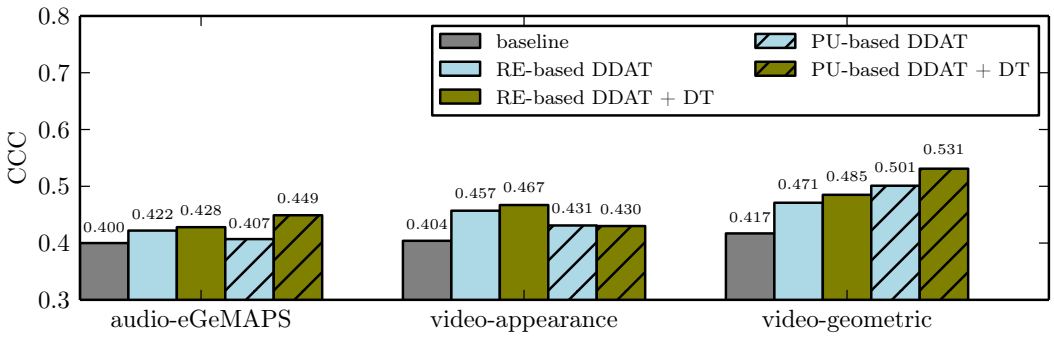
and other auxiliary tasks (i. e., reconstructing the input or predicting the perception uncertainty), potentially further benefit the emotion prediction.

Moreover, the performance of the MTL systems is further enhanced by the proposed DDAT framework, as shown in Table 4.13. In particular, the performance of the DDAT system for arousal and valence regressions respectively reaches CCC values of 0.694 and 0.422 with the audio-eGeMAPS feature set, 0.438 and 0.457 with the video-appearance feature set, and 0.400 and 0.501 with the video-geometric feature set. These results demonstrate that the DDAT systems offer significant advantages over ($p < .05$ via Fisher r-to-z transformation) the baseline method as well as the MTL approach (except in the case of valence regression with the audio-eGeMAPS feature set).

When comparing the two approaches used in the RE-based DDAT experiments, one can notice that adding the overall sum of the error leads to a better performance than adding the error vector. This is possibly attributable to the redundant dimen-



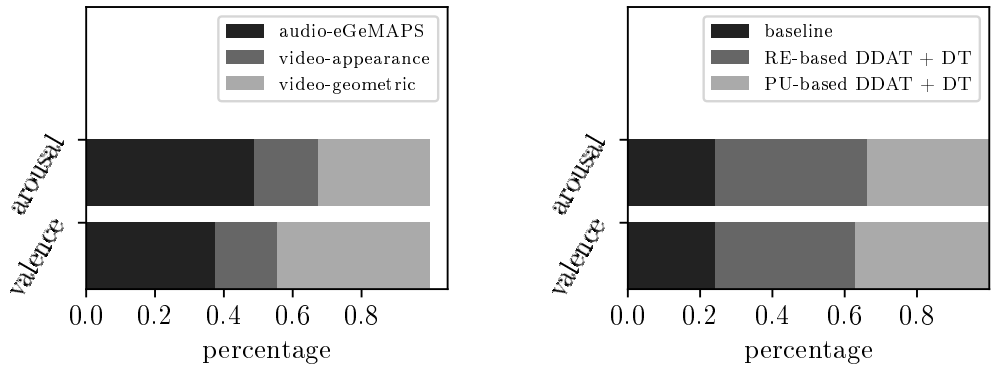
(a) arousal



(b) valence

Figure 4.9: Performance comparison between the single-task learning, the proposed dynamic difficulty awareness training approach based on reconstruction error (RE) or perception uncertainty (PU), and their dynamically-tuned (DT) versions. Results pertain to the test partition for both arousal (a) and valence (b) targets using three feature sets (audio-eGeMAPS, video-appearance, and video-geometric).

sionality of the error vector, which meanwhile yields much noise in the network training. When comparing the RE-based DDAT and the PU-based DDAT, it is noticeable that the two approaches perform similarly. This suggests that both approaches achieve the same goal but in different ways. That is, both approaches successfully explore the difficulty information in the pattern learning process, whereas the two DDAT approaches measure the difficulty information by the data reconstruction-capability and by the data perception-uncertainty, respectively. Moreover, it is worth mentioning that the RE-based DDAT approach, in contrast to the PU-based DDAT, not only fits the subjective pattern recognition tasks (e. g., emotion prediction in this work) but also holds the potential to be applied to objective tasks (e. g., phoneme prediction).



(a) information stream (b) model

Figure 4.10: Percentage of the contribution of each information stream (a) or model (b) for achieving the best arousal or valence predictions.

In addition, the dynamic tuning approach was carried out, which was introduced earlier in Section 3.2.2.4. Figure 4.9 illustrates the performances of the DDAT models with and without dynamic tuning of the predictions. Compared with the predictions without dynamic tuning, the dynamically-tuned predictions yield gains in most cases. For instance, the best achieved CCC for arousal prediction increased from 0.684 to 0.699, using the RE-based DDAT system with the audio-eGeMAPS feature set, whereas for valence prediction it increased from 0.511 to 0.531, using the PU-based DDAT system with the video-geometric feature set. The exceptions include the arousal predictions for both RE- and PU-based DDAT systems using the video-geometric feature set and the valence predictions for the PU-based DDAT system using the video-appearance feature set.

To combine predictions from different feature sets, different late fusions were applied on the individual predictions produced by using different modalities and models. As a consequence, Table 4.14 lists all scenarios (combinations) considered in the experiments as well as the respective performance. As can be seen in the table, the best performance on the test set for both arousal and valence is obtained when fusing the predictions from all *modalities* and *models*. In this context, the best results on the test set have been achieved at 0.766 for arousal and 0.660 for valence. These results beat most of the latest reported results from the same data, and they are close to the best result presented in AVEC 2016 [19] (i. e., 0.770 and 0.687 of CCCs for arousal and valence prediction), despite this system also utilising an additional modality (physiological features).

Finally, in order to analyse the importance of each modality and model, Figure 4.10 depicts their contributions to the arousal and valence predictions of the respective best performing models. For the arousal prediction, the acoustic features

Table 4.14: Late fusion performance in terms of CCC in different fusion strategies (i. e., modality-based, modality- and model-based, and dynamically-tuned modality- and model-based) for the *dev*elopment and *test* partitions of both *aro*usal and *val*ence regressions. The predictions are generated from the reconstruction-error-based DDAT framework (P_{re}) or the perception-uncertainty-based DDAT framework (P_{pu}); their dynamically-tuned versions ($P_{re,dt}$ or $P_{pu,dt}$); or the baseline model (P_{bs}). The best results achieved on the test set are in bold. Note that P_{re}, $P_{re,dt}$, P_{pu}, $P_{pu,dt}$, and P_{bs} are the fused predictions from three diverse feature sets.

various late fusion sets					*aro*		*val*	
P_{re}	$P_{re,dt}$	P_{pu}	$P_{pu,dt}$	P_{bs}	*dev*	*test*	*dev*	*test*
modality-based								
				✓	.822	.690	.705	.584
✓					.853	.763	.738	.615
		✓			.838	.715	.738	.615
modality- and model-based								
✓		✓			.860	.761	.755	.639
✓		✓		✓	.864	.752	.766	.653
modality- and model-based (dynamically-tuned)								
	✓				.853	.761	.739	.621
			✓		.819	.721	.733	.631
	✓		✓		.856	**.766**	.756	.651
	✓		✓	✓	.863	.754	.766	**.660**
state of the art								
strength modelling					.808	.685	.671	.554
end-to-end [197]					.731	.714	.502	.612
state of the art (+ physiology)								
feature selection + offset [95] [a]					.824	.747	.688	.609
SVR + offset [198] [b]					.820	.702	.682	.638
SC + CNN + LSTM [19] [c]					.862	.770	.750	.687

[a] AVEC '15 winner
[b] AVEC '16 baseline
[c] AVEC '16 winner

play a more important role than the visual features, whereas the opposite happens for the valence prediction. It is also noticed that the RE-based and PU-based DDAT systems contribute more than the baseline systems to the final predictions. Furthermore, the PU-based DDAT system is slightly more important for the valence prediction than it is for the arousal prediction. This might be due to the fact that predicting valence is much more difficult than arousal within audio modality [229].

4.6.3 Summary

In conclusion, this study provides a novel training framework which exploits the difficulty (weakness) information straightforwardly in the learning process for continuous emotion prediction. To extract the difficulty information, two strategies based on either the ontology of modelling or the content to be modelled are proposed. The two types of information separately measure the learning difficulty of a model by reconstructing its input, or the 'hardness' of the data to be learnt by predicting their perception uncertainty. This information indicated by an index can then be concatenated into the original features to update the inputs. The proposed methods were systematically evaluated on one benchmark database RECOLA. Experimental results have demonstrated that the proposed methods clearly improve the prediction performance of a model by evolving the difficulty information into its learning process.

In the future, it will be of interest to continue investigating the effectiveness of the proposed DDAT in discrete pattern predictions. Additionally, approaches for which the difficulty information could be possibly used as the prediction weights will be studied. Last but not least, the DDAT can be integrated with the Strength Modelling strategy as the two might bring benefits from different perspectives.

4.7 Emotion Prediction with Adversarial Training

To evaluate the effectiveness of the presented adversarial-training-based emotion prediction framework (cf. Section 3.2.3), in this section, experiments will be conducted on the RECOLA database for continuous emotion recognition.

4.7.1 Experimental Evaluation

In this study, to evaluate the performance of the CGAN framework for affective computing, as a first tentative work towards this research direction, speech recordings were utilised and 26 MFCC-based acoustic features were extracted from the open-source openSMILE toolkit [61].

Moreover, similar to the previous studies, an on-line standardisation was applied to the development and test sets by using the means and variations of the training partition. Meanwhile, annotation delay compensation of four seconds was performed to compensate for the temporal delay between the observable cues shown by the subjects, and the corresponding emotion reported by the annotators.

Next, the framework introduced in Section 3.2.3.3 was implemented with LSTM-RNNs. This is mainly due to that, LSTM-RNNs have been frequently examined to

be effective in capturing longer context information for sequential pattern recognition tasks, especially for continuous emotion recognition in this scenario. Moreover, the number of hidden layers was set to be two and the number of nodes per hidden layer to be 20. To accelerate the training process, the network weights were updated after running every mini-batch of eight sequences for computation in parallel.

In the network training process, NN_1 and NN_2 were alternatively trained, and this process was repeated in multiple runs. In each learning run, more training times were performed on NN_2 than on NN_1. More specially, 10 steps of training were conducted on NN_1, followed by 50 steps of training on NN_2 for arousal prediction, or 15 steps on NN_2 for valence prediction. This operation is twofold: (i) NN_2 is required to be superior enough [72], otherwise it is vulnerable to be 'cheated' by the predictions, and could not provide sufficient challenges to advance NN_1; (ii) valence prediction is normally considered as a more difficult task than arousal prediction from speech. Thus, NN_1 requests relatively more training steps for valence compared with arousal in each run. Besides, the hyperparameter λ in Equation (3.44) and Equation (3.47) was optimised via a grid search in the range of [.01, .02, .05, .1, .2, .5] on the development set.

4.7.2 Performance

To evaluate the system performance, results in terms of CCC are reported. Table 4.15 displays the performance obtained from the systems by using conditional adversarial training approaches. For comparison, conventional training approaches without conditional adversarial training were carried out as baselines. The networks with two hidden layers and four hidden layers were respectively investigated. When compared with the baselines, it can be seen that the system performance is significantly improved for both arousal and valence predictions (via a Fisher's r-to-z transformation as outlined in Section 4.2.2), when performing conditional adversarial training. Specifically, on the test set, the CCC values increase to .732 for arousal predictions, and .455 for valence predictions. The performance gain indicates that adversarial training with NN_2 brings benefit to the training of NN_1 to further ameliorate its predictions to some extent.

When implementing the Wasserstein distance into the objective function of NN_2 as detailed in Section 3.2.3.4, the obtained results are provided in the last row of Table 4.15. One may observe that the performance of the system for arousal prediction is further enhanced, i. e., from .780 to .797 in CCC on the development set, and from .732 to .737 for test; whereas for valence, one cannot get a similar observation. This implicitly suggests that for arousal, it is somewhat effortless for NN_2 to distinguish the input sources. On the other hand, it might be necessary to elaborate more to improve the objective function of NN_2 for valence.

Furthermore, one might notice that the obtained performance is comparable to or even outperforms those achieved from the state-of-the-art systems as listed in

Table 4.15: Performance in terms of Concordance Correlation Coefficient (CCC) of the proposed conditional adversarial training approaches, as well as its variation (+ Wasserstein distance), for both *arousal* and *valence* regressions, evaluated on the *dev*elopment and *test* partitions.

approaches	*arousal*		*valence*	
	dev	*test*	*dev*	*test*
baseline				
LSTM-RNN (2 layers)	.777	.718	.491	.435
LSTM-RNN (4 layers)	.761	.723	.487	.390
state of the art				
CCC-objective [206]	.412	.350	.242	.199
end-to-end [196]	.741	.686	.325	.261
Strength Modelling	.755	.666	.476	.364
reconstruction-error-based DDAT	.807	.676	.508	.422
proposed				
CGAN	.780	.732	**.501**	**.455**
CGAN w/ Wasserstein Distance	**.797**	**.737**	.474	.444

Table 4.15 for both arousal and valence prediction, which yields the best results to date on the RECOLA database from speech.

Finally, to intuitively present the system performance, the arousal and valence predictions on a randomly selected subject from the test partition are demonstrated in Figure 4.11 (a) and (b), respectively. From the figure, it is clear to observe that the predictions (blue lines) and the corresponding gold standards (red lines) have a high correlation.

4.7.3 Summary

In summary, in this study, in contrast to previous works that use adversarial training for generating realistic data, the performance of conditional adversarial training in the application of emotion recognition has been tentatively examined. To stabilise the learning process, the objective function can be modified by applying the Wasserstein distance. Result performances achieved on the benchmark database RECOLA indicate that conditional adversarial training is helpful to improve the system performance for speech emotion recognition.

Future work includes more experimental evaluations on other modalities such as video. Moreover, it is interesting to perform an end-to-end structure to automatically extract salient features for emotion prediction, rather than the hand-crafted

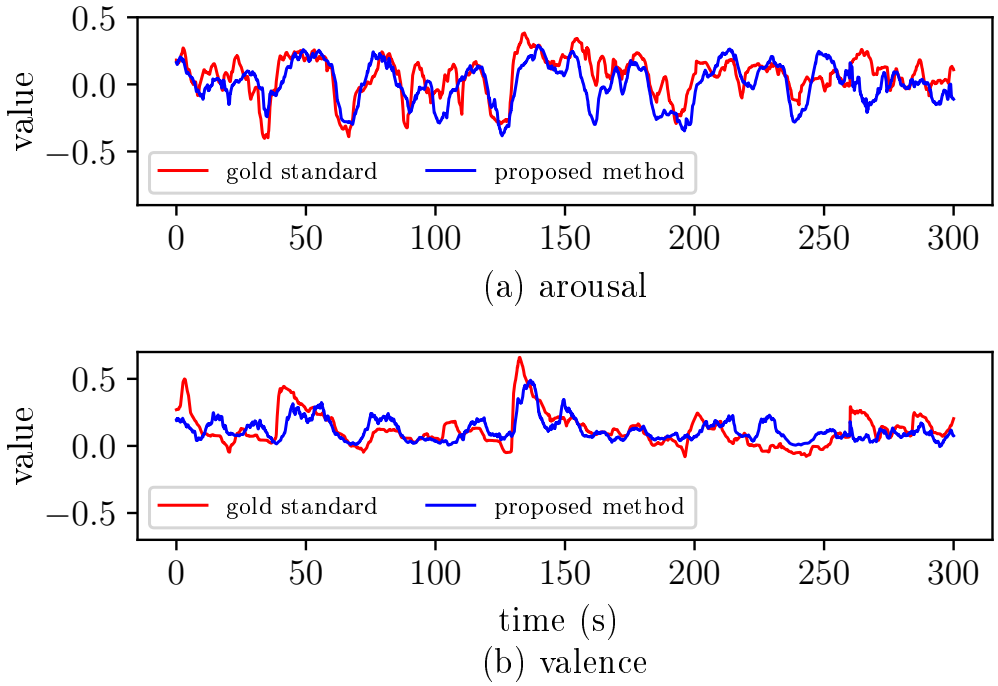


Figure 4.11: Automatic predictions of arousal (a) and valence (b) obtained by conducting conditional adversarial training, for a randomly selected subject from the test partition on RECOLA database.

features that were employed in this present framework. In the future, it would also be of considerable interest to investigate and adjust other adversarial training structures for emotional behaviour analysis, such as energy-based generative adversarial network and cycle generative adversarial network [241].

4.8 Continual Emotion Prediction via Lifelong Learning

In this section, experiments will be reported with regard to the proposed lifelong learning-based continual emotion prediction presented in Section 3.3. Specifically, aiming at addressing the catastrophic forgetting issue when modelling emotions in a cross-culture scenario, experiments were performed on the RECOLA and SEWA databases (cf. Section 4.1.1 and Section 4.1.2), to validate the feasibility and effectiveness of the introduced algorithm for French and German emotion recognition.

4.8.1 Experimental Evaluation

In this study, BoAW and BoVW representations were employed, as these features were provided in the AVEC challenges recently as established feature sets for the baselines. Thus, these features were utilised for a fair performance comparison with other related works. In particular, for audio data, the eGeMAPS set was extracted as LLDs. Then, BoAW representations were computed over a collection of successive frames for each step of 40 ms and 100 ms to match the frequency of the annotations of RECOLA and SEWA, respectively, following the same settings as provided in the AVEC challenge baseline systems [161]. More specifically, segment-level representations were computed from the LLDs with one hard assignment on a codebook of size 100. This resulted in a set of 100 acoustic BoAW features. Similarly, following the suggestions in [161], 17 Facial Action Units (FAUs) were first extracted per frame as LLDs via the open-source toolkit OpenFace [12]. The same processing chain thereafter was conducted as when generating the BoAW features. With this process, additional 100-dimensional visual representations were achieved, on both the RECOLA and SEWA sets accordingly.

Moreover, before describing the experiments, for a better understanding and a clearer view, the training with the RECOLA database is defined as *FR*, and the learning on the SEWA German database as *GE*. This led to the following two distinct sequential training schemes:

- *FR after GE, w/ EWC*, where the model is first trained on SEWA, and later trained on RECOLA, with the EWC constraint (cf. Section 3.3.2);
- *GE after FR, w/ EWC*, where RECOLA is first employed to train the model, and SEWA as the new task to be learnt.

Further, for comparison, other baselines were carried out as well. To be more specific, the following baseline systems were run:

- single task only, where isolated learning was performed on each dataset separately. In this scheme, two baselines can be obtained which are marked as *FR only* and *GE only*;
- single task only with weight regularisation, where the model was again optimised on a single dataset. However, the L2 regularisation penalty term was explored for better generalisation performance. In this scenario, another two baselines were provided, i. e., *FR only, w/ L2-norm* and *GE only, w/ L2-norm*;
- furthermore, the sequential fine-tuning was implemented, where the model is fine-tuned on the second corpus after firstly having been optimised via the first one. This is similar to the proposed training process, however, without considering any constraint. In this case, another two baselines were provided, i. e., *FR after GE, w/o EWC* and *GE after FR, w/o EWC*.

Besides of these outlined baseline systems, joint training was also investigated, i. e., training the network on all available datasets jointly, denoted hereinafter as *FR and GE*.

Moreover, within each training scheme, models were individually trained on acoustic features, visual features, or the combination of the two, for arousal and valence prediction, respectively. To this end, all models were implemented by using GRU-RNNs. Note that, for the sake of fair comparison, the same network structure was used for all training schemes, and the network settings were empirically chosen which can provide competitive performance on both databases when compared with other previous within-cultural models [161]. To be more specific, each network consists of four hidden layers with 100 units per layer. While training the network, an Adam optimiser was employed with an initial learning rate of .001. In addition, the early stopping strategy was executed if no performance improvement on the development set was observed after 20 successive epochs.

Also, it is important to note that, the development set shared the same culture type with the training set of the current task. For instance, when training a model following *FR after GE* (learn SEWA first, then RECOLA), the learning first ceased based on the performance on the SEWA development set when learning on the SEWA training set. After that, when continually learning on the RECOLA training set as a new task, the training process will be terminated by inspecting its performance on the RECOLA development set.

Furthermore, in all of the experiments, following the suggestions of the AVEC challenges [161], annotation shifting was performed to compensate annotation delay, and a post-processing chain of four stages was performed with an aim to refine the obtained predictions, similar as in all previous studies.

4.8.2 Performance

In the following, prediction results obtained with the proposed continual emotion recognition approach are provided and analysed. In particular, the proposed models are compared against other baseline systems. In addition, the effectiveness of the method is validated by visualising the effect of the elastic-weight-based penalty term and its impact on model parameters.

4.8.2.1 Cross-cultural Emotion Recognition

Table 4.16 and Table 4.17 demonstrate the results of various training strategies for arousal and valence predictions, respectively, on the RECOLA and SEWA datasets for cross-cultural emotion recognition. For a clear view, these training strategies are sorted into three categories, i. e., six baseline strategies to compare against, two lifelong training models, and a joint training strategy which can be viewed as an upper bound in this circumstance. Moreover, results are summarised in three blocks

Table 4.16: CCC performances via various training strategies for emotion regression based on *audio*, *video*, or the *combination*. Performance on the development sets and test sets of the two databases ($FR_{dev}, FR_{test}, GE_{dev}, GE_{test}$) as well as the average performance on the two test sets (μ_{test}) are reported for *arousal*, respectively.

Features	Methods	FR_{dev}	FR_{test}	GE_{dev}	GE_{test}	μ_{test}
audio	*baseline strategies*					
	FR only	.631	.552	.311	.036	.294
	GE only	-.022	.009	.388	.246	.128
	FR only, w/ $L2$-norm	.613	.543	.276	.049	.296
	GE only, w/ $L2$-norm	-.023	.014	.357	.224	.119
	FR after GE, w/o EWC	.640	.538	.334	.047	.293
	GE after FR, w/o EWC	-.057	-.006	.383	.243	.119
	proposed lifelong learning strategies					
	FR after GE, w/ EWC	.639	.538	.335	.046	.292
	GE after FR, w/ EWC	.554	.509	.377	.157	.333
	joint training (upper bound)					
	FR and GE	.498	.457	.389	.176	.317
Video	*baseline strategies*					
	FR only	.167	.180	.277	.175	.178
	GE only	.073	.014	.517	.374	.194
	FR only, w/ $L2$-norm	.181	.217	.223	.227	.222
	GE only, w/ $L2$-norm	.094	.029	.563	.404	.217
	FR after GE, w/o EWC	.195	.241	.229	.212	.227
	GE after FR, w/o EWC	.078	.024	.536	.365	.195
	proposed lifelong learning strategies					
	FR after GE, w/ EWC	.195	.241	.230	.212	.227
	GE after FR, w/ EWC	.197	.184	.601	.413	.299
	joint training (upper bound)					
	FR and GE	.168	.177	.541	.455	.316
Fusion	*baseline strategies*					
	FR only	.621	.578	.397	.102	.340
	GE only	.090	.063	.581	.401	.232
	FR only, w/ $L2$-norm	.634	.627	.351	.145	.386
	GE only, w/ $L2$-norm	.080	.050	.588	.412	.231
	FR after GE, w/o EWC	.631	.596	.436	.150	.373
	GE after FR, w/o EWC	.056	.035	.613	.424	.230
	proposed lifelong learning strategies					
	FR after GE, w/ EWC	.600	.599	.500	.224	.412
	GE after FR, w/ EWC	.551	.530	.568	.366	.448
	joint training (upper bound)					
	FR and GE	.534	.533	.601	.399	.466

Table 4.17: CCC performances via various training strategies for emotion regression based on *audio*, *video*, or the *combination*. Performance on the development sets and test sets of the two databases ($FR_{dev}, FR_{test}, GE_{dev}, GE_{test}$) as well as the average performance on the two test sets (μ_{test}) are reported for *valence*, respectively.

Features	Methods	FR_{dev}	FR_{test}	GE_{dev}	GE_{test}	μ_{test}
audio	*baseline strategies*					
	FR only	.287	.248	.085	.024	.136
	GE only	.007	-.003	.390	.245	.121
	FR only, w/ $L2$-norm	.227	.249	.140	.019	.134
	GE only, w/ $L2$-norm	.025	.046	.361	.240	.143
	FR after GE, w/o EWC	.325	.272	.072	-.011	.131
	GE after FR, w/o EWC	.027	.050	.360	.251	.151
	proposed lifelong learning strategies					
	FR after GE, w/ EWC	.232	.159	.265	.156	.158
	GE after FR, w/ EWC	.019	-.009	.406	.192	.092
	joint training (upper bound)					
	FR and GE	.293	.234	.343	.186	.210
video	*baseline strategies*					
	FR only	.413	.354	.552	.397	.376
	GE only	.406	.211	.571	.517	.364
	FR only, w/ $L2$-norm	.432	.341	.561	.413	.377
	GE only, w/ $L2$-norm	.396	.215	.581	.508	.362
	FR after GE, w/o EWC	.430	.355	.480	.289	.322
	GE after FR, w/o EWC	.368	.197	.564	.467	.332
	proposed lifelong learning strategies					
	FR after GE, w/ EWC	.479	.370	.583	.465	.418
	GE after FR, w/ EWC	.466	.271	.562	.587	.429
	joint training (upper bound)					
	FR and GE	.550	.382	.598	.600	.491
fusion	*baseline strategies*					
	FR only	.529	.436	.334	.189	.313
	GE only	.241	.134	.623	.485	.310
	FR only, w/ $L2$-norm	.511	.380	.372	.222	.301
	GE only, w/ $L2$-norm	.249	.134	.633	.536	.335
	FR after GE, w/o EWC	.528	.421	.381	.250	.336
	GE after FR, w/o EWC	.250	.150	.637	.527	.339
	proposed lifelong learning strategies					
	FR after GE, w/ EWC	.514	.412	.534	.370	.391
	GE after FR, w/ EWC	.450	.252	.617	.569	.411
	joint training (upper bound)					
	FR and GE	.523	.399	.601	.584	.492

with respect to the applied feature sets, namely, audio, video, and the combination of the two.

First, let us compare the proposed models against the six baseline strategies. From the tables, one may notice that performance is heavily degraded when there is a cultural mismatch between the training and inferring sets. Taking the arousal prediction from audio as an example, on FR_{test}, a CCC of .552 is obtained by training on the same cultural data (i. e., *FR only*), while the performance dramatically reduces to .009 if training with German data only (i. e., *GE only*). Likewise, for GE_{test}, the performance of *GE only* is remarkably superior to *FR only*. Similar observations can be drawn over all three distinct feature sets for the arousal prediction as well as for the valence prediction.

Moreover, results of another two baseline training models, *FR only, w/ L2-norm* and *GE only, w/ L2-norm*, are also provided in Table 4.16 and Table 4.17, which aims at improving the generalisation performance on new, unseen data. For instance, when comparing *FR only* and *FR only, w/ L2-norm*, the obtained CCCs on both test sets are boosted for the arousal regression via the fused audiovisual features, from .578 to .627 on FR_{test} and from .102 to .145 on GE_{test}. Nevertheless, a severe performance discrepancy still exists between the two sets, and this indicates that it is essential to construct a model to learn from data of both cultures.

Further, in order to learn from data of both sets, sequential training strategies have been evaluated in the last two baseline systems, namely *FR after GE, w/o EWC* and GE after FR, w/o EWC. These systems, without considering the EWC regularisation, suffer severely from the catastrophic forgetting. Let us take the arousal prediction from audio signals with *GE after FR, w/o EWC* as an example, where the model first learns from FR and then GE. The obtained CCC for French decreases dramatically, from .631 to $-.057$ on FR_{dev} and from .552 to $-.006$ on FR_{test}. It can be seen that, though both sets are learnt in a sequence, performance of the first task is damaged as the model adaptation to the second culture disrupts the knowledge learnt from the first one. It suggests that advanced training strategies are inevitable and essential to deal with this issue.

With the proposed continual learning approaches, one may notice that the aforementioned catastrophic forgetting problem is alleviated remarkably, by preserving the knowledge of previous tasks via the EWC regularisation during training. Again, taking for instance predicting arousal from audio signals, with EWC, the CCCs achieved on the French dataset by *GE after FR, w/ EWC* are .554 on FR_{dev} and .509 on FR_{test}, respectively, and in the meanwhile, the CCCs for German are still competitive to a German-dependent model (.377 vs .388 on GE_{dev} and .157 vs .246 on GE_{test}). Such an observation can be found in Table 4.16 and Table 4.17 in other scenarios.

Notably, when comparing *FR after GE* with *GE after FR*, it is also interesting to observe that, the performance of the latter is on average superior to the former. Given μ_{test} which is the average performance on two test sets, *GE after FR, w/ EWC*

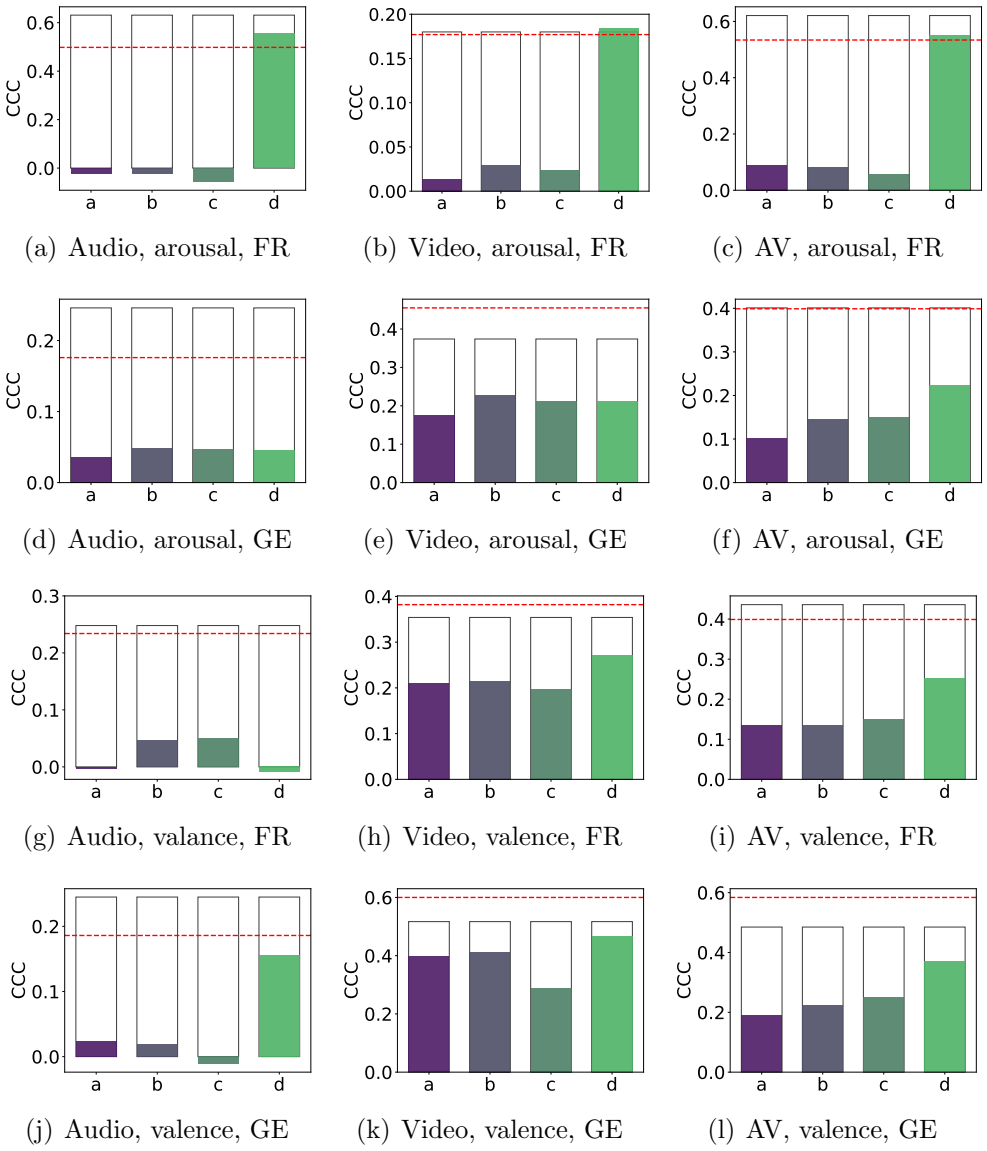
is better than *FR after GE, w/ EWC* in five out of six cases (three feature sets by two emotion dimensions) except for the valence prediction from audio signals. This may indicate that the order of the training tasks also plays a key role in a continual emotion recognition system. One may gain some insight from curriculum learning and infer that learning several tasks in a proper order might lead to improved average performance.

Furthermore, from the results, one may also observe that in most cases, when modelling emotion patterns from audio-only, the models achieve better performance in the arousal prediction than in the valence prediction. In contrast, when estimating via facial expressions, observations are found in another way around. Moreover, when combining the audio and video features via early fusion, the model performance is improved. These findings are consistent with previous studies [162]. In particular, when learning from both audio and video, the best average CCCs on the two test sets μ_{test} are obtained by joint training (*FR and GE*), reaching to .466 for arousal and .492 for the valence prediction, respectively.

Rather than the joint training paradigm, in the proposed lifelong learning-based models, the two datasets were learnt one after the other, and this reduces the high storage requirement issue one might face in joint training. With this manner, comparable performance on μ_{test} in terms of CCC is achieved with the EWC-based model *GE after FR, w/ EWC*, i. e., .448 and .411 for the arousal and valence predictions, respectively. This suggests that sequential learning is, to some degree, a potential replacement of the joint training as the system may benefit from lower storage requirement and computation load. This is extremely vital for real-life intelligent systems, where the number of given tasks and thus the amount of training data might grow rapidly.

Lastly, for a better interpretation, these results are also visualised in Figure 4.12. In this figure, the performance of the proposed continual learning systems on FR_{test} and GE_{test} are compared with their corresponding baseline systems under six distinct setting combinations of three different feature sets and two emotion dimensions separately. Moreover, the performance of two upper bound systems is depicted as well. In particular, in each subfigure, the upper bound of its matched culture-specific model (i. e., *FR only*) is presented as a white bar; while the joint training upper bound is drawn as red dotted lines. Another four coloured bars denote the corresponding performance of the mismatched culture-specific model (i. e., *GE only*), mismatched model with L2-norm (i. e., *GE only, w/ L2-norm*), sequential training model without EWC (e. g., *GE after FR, w/o EWC*), and the introduced sequential training model with EWC (i. e., *GE after FR, w/ EWC*, also the lifelong learning model), respectively. As a consequence, the performance degradation from all introduced and compared models can be visualised as the white space between two bars.

It can be seen from Figure 4.12 that, in most cases, the lifelong learning model offers advantages over other baseline models and yields less performance difference



(a) Audio, arousal, FR (b) Video, arousal, FR (c) AV, arousal, FR

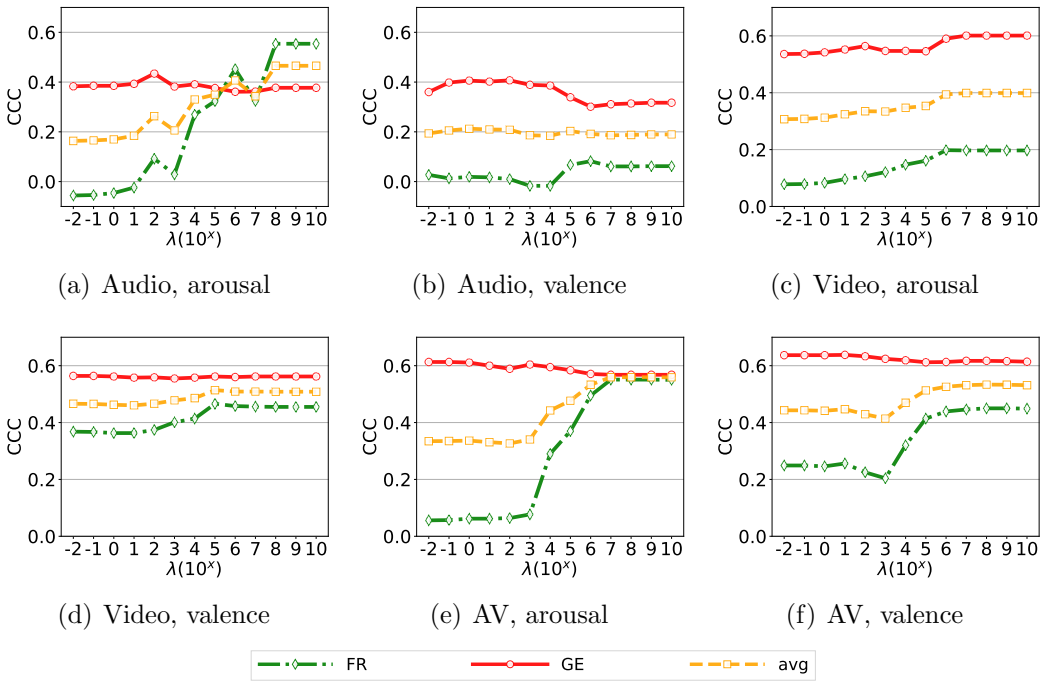
(d) Audio, arousal, GE (e) Video, arousal, GE (f) AV, arousal, GE

(g) Audio, valance, FR (h) Video, valence, FR (i) AV, valence, FR

(j) Audio, valence, GE (k) Video, valence, GE (l) AV, valence, GE

Figure 4.12: Visualisation of the performances in terms of CCC of the proposed methods comparing with other baseline approaches on the test sets of FR and GE. Results are separately shown for arousal and valence regressions via audio, video, and their combination (AV), respectively. Note that, the white bars indicate the performance of a matched culture-specific model, while the red dotted lines denote the performance of a joint training model. a: mismatched culture-specific model, b: mismatched model w/ L2-norm, c: sequential training w/o EWC, d: sequential training w/ EWC (proposed).

with its matched culture-specific model than other baseline models. In particular, when training for the arousal prediction, the *GE after FR, w/ EWC* models achieve



(a) Audio, arousal (b) Audio, valence (c) Video, arousal

(d) Video, valence (e) AV, arousal (f) AV, valence

Figure 4.13: The effect of the hyper-parameter to control the regularisation λ in the proposed *GE after FR, w/ EWC* model, when predicting arousal and valence on the development sets via three various feature types, i. e., audio, video, and audiovisual (AV). The average performance of both FR and GE is calculated and denoted as avg.

competitive or even better results than the joint training upper bounds (cf. Figure 4.12 (a), (b), and (c)). This demonstrates the great potential of implementing continual emotion recognition by investigating advanced lifelong learning training algorithms.

4.8.2.2 Hyperparameter Selection

As given in Equation (3.55), λ is utilised to regulate the contribution of previous knowledge when learning a new task. Therefore, to better demonstrate the effect of λ for the performance, various models were learnt with different λs in the range of $[10^{-2}, 10^{10}]$. Figure 4.13 shows how the performance of the lifelong model varies with respect to various λs on the development sets. Note that, in this figure, only results on *GE after FR, w/ EWC* models are demonstrated, as it generally performs better than *FR after GE, w/ EWC* (cf. Section 4.8.2.1).

As shown in Figure 4.13, when λ is small, performance on the previous task (French emotion recognition) is relatively low, compared with the second task (German emotion recognition). Then, when λ increases, the performance on FR_{dev}

improves in a large margin in most cases, until a point where it does not benefit from further increasing. Likewise, when depicting the averages of FR_{dev} and GE_{dev}, a performance improvement can also be observed along the increased λ until it becomes flat again. In this regards, a proper λ is demanded to retain the previous knowledge. However, the performance on GE_{dev} either increases (cf. Figure 4.13(c)), remains (cf. Figure 4.13 (a) and (d)), or decreases slightly (cf. Figure 4.13 (b), (e), (f)) under different settings. This may highly depend on whether the previous knowledge is beneficial or not to the current learning process.

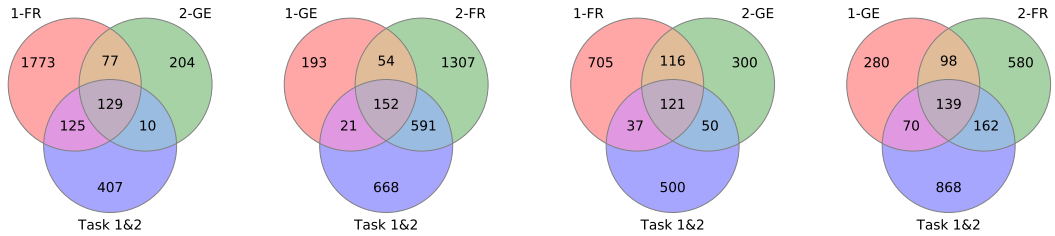
4.8.3 Effectiveness Verification

In the following, the focus further lies at validating the effectiveness of the EWC-based sequential learning approach by inspecting the impacts of EWC on the plasticity of parameters to be learnt. In particular, in the selected RNN model for emotion regression tasks, there are more than 271 K parameters to be learnt. Of these parameters, their importance with respect to a given task can be estimated by its Fisher information (cf. Section 3.3.2). On this account, given a predefined threshold 10^{-4} to only consider the parameters that have a Fisher value above it, for each model, a parameter set can be generated to incorporate these important parameters. Then, when EWC is applied, the plasticity of the parameters in the set can be decreased to tackle catastrophic forgetting when a new task comes.

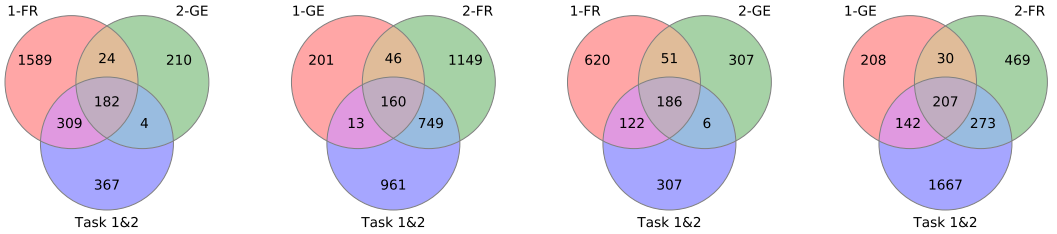
In this study, the relations of three parameter sets of this kind are investigated, i. e., one for a model trained on French only, one for German only, and one for a model trained on the two databases one after another. Then, the relations of these three sets can be demonstrated by a couple of Venn diagrams, under eight audiovisual training scenarios, as shown in Figure 4.14. In particular, the four Venn diagrams in the upper row of Figure 4.14 present four cases when carrying out sequential learning without EWC, while the remaining four in the lower row are corresponding models trained with EWC. Moreover, the values in the diagram depict the number of parameters which are vital to only one model, or two models (overlaps of every two circles), or three models (overlaps of all three circles). Hence, when investigating the intersection of a red circle (model trained only on the first task) and a purple circle (model learnt on two tasks) and comparing every two Venn diagrams in a column, one may see that the intersection in the lower diagram is always greater than (three out of four cases) or at least equal to (one case only) that from the upper diagram. This might indicate that sequential training with EWC is capable of maintaining more parameters that are important to the previous task, by reducing the plasticity of these parameters in future learning.

4.8.3.1 Discussion

In the following, some potential limitations of the current EWC-based continual emotion recognition system will be discussed. As can be seen in Figure 4.14, after



(a) aro, w/o EWC, FR (b) aro, w/o EWC, GE (c) val, w/o EWC, FR (d) val, w/o EWC, GE



(e) aro, w/ EWC, FR (f) aro, w/ EWC, GE (g) val, w/ EWC, FR (h) val, w/ EWC, GE

Figure 4.14: Venn diagrams to visualise the relations among three parameter sets by analysing the important parameters obtained in three models for audiovisual *aro*usal and *val*ence predictions, where each of them can be viewed as a circle. In particular, the red circle denotes a culture-specific model, i. e., 1-FR or 1-GE; the green circle represents another culture-specific model, i. e., 2-GE or 2-FR; and the purple circle indicates a sequential training model that learns task 1 and 2 sequentially. The values in the circles show the number of important parameters, which belong to one set only or lie at the intersection of two or even three sets.

learning two tasks, the number of important parameters is increasing to preserve knowledge from both tasks, indicating that a lower amount of parameters is able to be largely changed for future tasks. Due to this limitation, EWC is not sufficient for learning a large number of tasks, when most, if not all, parameters appear to be rather essential to keep the knowledge for all past tasks. The reason is that it might be caught in a dilemma: on the one hand, new knowledge can only be acquired by updating parameters accordingly; on the other hand, past knowledge will get destroyed if any parameter is modified.

Moreover, although the performance of the proposed model is superior to other baseline models, meaning that the catastrophic forgetting is partially addressed, much work is still needed toward closing the performance gap between it and a culture-specific model. Therefore, in future studies, other more advanced lifelong learning approaches will be investigated and applied toward general emotion perception systems. These techniques include, but are not limited to, PathNet [65],

GeppNet [68], progressive neural networks [167], and dynamically expandable networks [220].

4.8.4 Summary

In conclusion, the proposed EWC-based continual emotion recognition models outperform other related models for cross-cultural emotion recognition, by quantifying the importance of weights to previous tasks and then adjusting the plasticity of weights accordingly for following tasks. In particular, the method overcomes the limitations of the conventional isolated training approach, enabling a model to learn multiple tasks in an open-set environment, with no or limited forgetting about the knowledge obtained from previous tasks. Hence, it becomes a promising alternative of joint training to dealing with the discrepancy among multiple cultures for emotion perception. This approach was evaluated on two benchmark databases RECOLA and SEWA for emotion regression across French and German languages. Experiment results show that, the model can be adapted continually and keep on learning over time to some extent.

In future work, the goal is to investigate other more advanced lifelong learning algorithms, and compare their performance with the introduced EWC in the field of emotion recognition. Further, beyond estimating emotional states across cultures and languages, it will also be of considerable interest to develop approaches to learn and adapt continually across different modalities and tasks.

4.9 Behaviour Synchronisation Analysis

Experiments in this section are aimed at automatically analysing the mimicry behaviours from speech, which is presented in Section 3.4. For this purpose, the SEWA dataset (cf. Section 4.1.2) is exploited since it consists of recordings of conversations from six different cultures. In the following, details of the experimental setups and results are provided.

4.9.1 Experimental Evaluation

From the 197 conversations in the SEWA corpus, hand-crafted acoustic features were extracted on the frame-level from the audio of all recordings. That is, for each audio recording, 65 LLDs from the ComParE feature set were extracted, together with their corresponding first-order derivatives (deltas), resulting in a frame-level feature vector of size 130 for each step of 10 ms. Moreover, before training the AE, the LLD sequences based on the transcriptions provided in the SEWA database were segmented, where information on the start and end of each speech segment and the subject ID of the corresponding segment is given.

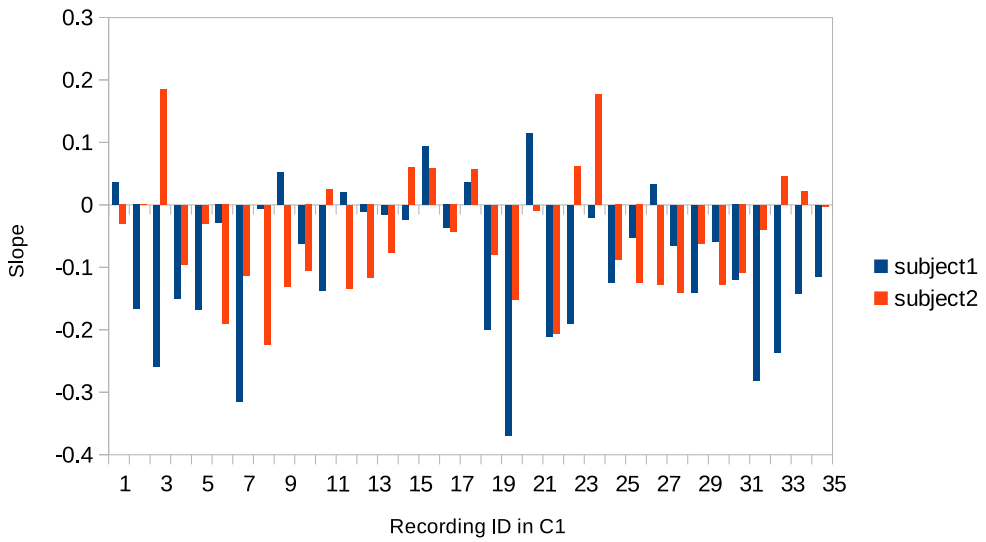


Figure 4.15: Slope of RMSE sequences of 70 Chinese subjects from 35 recordings. In each recording, there are two subjects as denoted with blue and red bars, respectively.

As a tentative study, the AE applied in this work was a three-layer encoder with a three-layer decoder. In the preliminary experiments, the number of nodes in each layer has been chosen as follows: 130-64-32-12-32-64-130, where the output dimension is exactly the same as the input dimension. After generating the reconstruction errors of the tested subject over time, the resulting sequence was exploited to perform a linear regression, with the assumption that the slope of the learnt line may indicate the changes of the behaviour patterns along time. More specifically, when the slope is negative, it may demonstrate that during the chat session, the tested subject turns to become more similar to the subject who (s)he is talking to. Thence, if the slope is positive, it may imply the opposite. Additionally, the amplitude of the slope can be an indicator to denote the level of the similarity or dissimilarity.

4.9.2 Performance

Let us first discuss the results achieved from the first culture, i. e., Chinese (C1). From all 35 recordings, the average slope of the RMSE sequences of all 70 subjects is -0.07. From Figure 4.15, one may notice that most of the slopes (54 of out 70 cases) are negative, whereas only a few (16 out of 70 cases) are positive. This indicates that, during the recordings, the acoustic LLD features of the tested subjects have a smaller reconstruction error when time passes by. Considering that the AE is trained with the other subject within the same recording, a smaller reconstruction error may reveal a higher similarity between these two subjects. To sum up, a negative slope

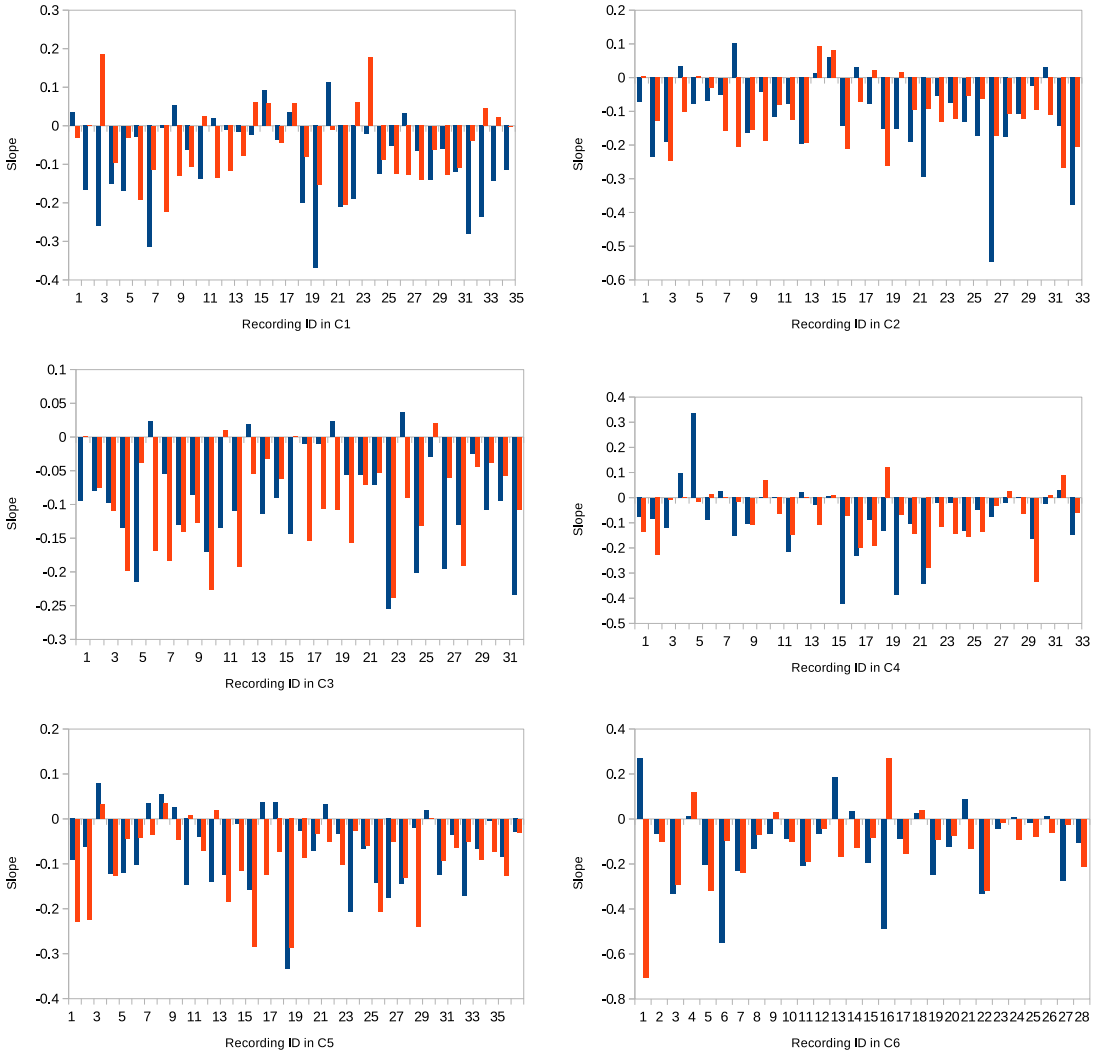


Figure 4.16: Slope of RMSE sequences of paired subjects from all recordings of six cultures (C1: Chinese, C2: Hungarian, C3: German, C4: British, C5: Serbian, C6: Greek). In each recording, there are two subjects as denoted with blue and red bars, respectively.

implies a decreasing reconstruction error along time and could indicate a similarity increasing among the speakers during the video chat. Interestingly, similar patterns have also been found in all the other five cultures. Nevertheless, the ratio of the negative slopes and the average slope are different from culture to culture. Figure 4.16 demonstrates the slope of RMSE of each subject for all recordings respectively.

Given these results, the average slopes s of all cultures were calculated separately, as well as the PCCs of two slopes obtained from all recordings within the same culture, respectively, with the aim to perceive a cultural variation in spontaneous remote conversations. Results are given in Table 4.18. Note that, a negative slope

Table 4.18: Average slope of RMSE sequences of all subjects within six different cultures is listed in the upper row, respectively; the correlation coefficient denoted as *pcc of pairs* indicates the correlation of behaviours of two subjects and is listed in the last row for each culture (C1: Chinese, C2: Hungarian, C3: German, C4: British, C5: Serbian, and C6: Greek).

	C1	C2	C3	C4	C5	C6
average slope	-0.07	-0.11	-0.10	-0.07	-0.08	-0.12
pcc of pairs	-0.03	0.34	0.15	0.39	0.39	-0.26

denotes that the subject shows a more similar speech behaviour in a conversation along time; the more similar a subject is leaking like his partner, the larger the slope is towards the negative direction.

From Table 4.18, one may notice that on average, individuals of all six cultures tend to behave more similar during the conversation, given that the average slopes are all negative. However, cultural variation remains, as the most negative slope (-0.12) is obtained for the Greek (C6) culture and the smallest slope (-0.07) is seen for Chinese (C1) and British (C4).

Moreover, taking the PCC into account, one may see the cultural variation from another view. A positive PCC value demonstrates that subjects of a culture tend to converge to a similar state, either both behave like or unlike each other, while a negative PCC may indicate that conversations are more likely to be dominated by one subject. For example, no correlation has been seen in conversations of the Chinese pairs (C1), with a PCC of -0.03 which is close to 0. However, strong linear correlations have been revealed in four cultures, either positive (Hungarian (C2), British (C4), and Serbian (C5)) or negative (Greek (C6)). Besides, a weak positive correlation can be seen in German (C3). These findings need to be verified by literature in sociology, anthropology, and in the anthropologic linguistics domain, particularly in the field of conversation analysis [188], which is, however, out of the scope of the present study. Note that, despite that the SEWA database was designed and developed with control of age and gender of the subjects, discrepancies caused by these or other aspects such as educational background, occupation, and health status cannot be avoided and might still have an impact on our observations.

4.9.3 Summary

Experiments performed in this section on the SEWA dataset have demonstrated that, an autoencoder has a great potential to recognise the spontaneous and unconscious temporal behaviour synchronisation in the social interaction, by the observation of the reconstruction error using the acoustic features extracted from the speech of a conversational partner. Also, some insights into the synchronisation of vocal

behaviour in dyadic conversations of people from six different cultures are reported. In the future, additional evaluation strategies to measure the degree of similarity between subjects will be explored, other than the slope of the reconstruction errors. Moreover, it will be of great interest to further evaluate the effectiveness of the approach for automatically detecting the behaviour synchronisation from audiovisual conversations.

5

Discussion and Outlook

Emotions are an essential part of human mental activities and of paramount importance in communication and decision-making processed. On that account, automatically detecting and interpreting emotional behaviours of human beings plays a crucial role in developing intelligent Human-Computer Interaction systems. Consequently, it has become increasingly popular in both research and industry areas, and considerable efforts have been made for this aim by exploiting various algorithms and techniques. However, most of the existing approaches for emotion recognition limited by the fact that they are still confronting many open challenges. Such challenges include evaluating systems in real-world applications rather than on constrained laboratory scenarios, learning discriminative representations for emotion modelling, tailoring established and emerging algorithms for the specific tasks, achieving continual learning in a lifelong context, and analysing the empathic behaviours.

To deal with these challenges, this thesis aims at investigating several deep learning-driven algorithms for emotional behaviour analysis, where audiovisual emotional data captured in unconstrained conditions are largely considered. In particular, this work presents and evaluates a variety of approaches to reach the following four major objectives: (1) learning advanced and meaningful representations for emotion perception, (2) improving the effectivenss and robustness of emotion recognition systems with deep learning algorithms, (3) performing lifelong learning for cross-cultural systems, and (4) conducting empathic behaviour analysis with deep learning approaches.

In the following, this chapter summarises the methods presented and the results obtained, and concludes this thesis in Section 5.1. Then, limitations of current work and open issues for future work are discussed in Section 5.2.

5.1 Contributions

In this section, the main achievements of this thesis are described in a nutshell.

Targeted at the first main research question (**RQ-1** in Section 1.2), all proposed deep-learning-based methodologies and models in this thesis have been evaluated on one or more publicly available emotion databases, and shown their effectiveness in automatic emotional behaviour analysis based on spontaneous affective audiovisual data.

Next, aiming at answering **RQ-2**, namely, to tackle the representation learning challenge within the context of emotion recognition, two approaches have been developed. First, to address the representation learning problem in a multimodal perspective, this thesis has introduced a novel crossmodal emotion embedding framework in Section 3.1.1. In this framework, a crossmodal triplet constraint is deliberated to incorporate information from heterogeneous data, especially from different but complementary modalities. Moreover, to leverage the temporal context information, this thesis has advanced the state-of-the-art bag-of-audio-word feature learning strategy with a hierarchical feature learning architecture (cf. Section 3.1.2). With this method, mid-level features with context information have been obtained, bridging the gap between frame-level features and long-term emotional segments and therefore enhancing the regular bag-of-audio-word approach.

Moreover, to approach **RQ-3**, i. e., to advance the emotion recognition performance by customising conventional deep models, three deep learning-driven approaches and frameworks have been proposed and evaluated in this thesis from Section 3.2.1 to Section 3.2.3. In particular, these training strategies include strength modelling that takes advantage of different models (cf. Section 3.2.1), dynamic difficulty awareness training which exploits the difficulty information of a certain model (cf. Section 3.2.2), and conditional adversarial training that exploits the adversarial training concept for emotion prediction (cf. Section 3.2.3). Experimental results have demonstrated that these proposed training strategies facilitate emotional behaviour analysis based on audiovisual data, yielding appealing performance improvements in comparison with the state-of-the-art audiovisual emotion predictors.

Further, another contribution of this thesis is to investigate lifelong learning for emotion recognition (cf. Section 3.3), attempting to shed light on understanding **RQ-a (side)** in Section 1.2. For the first time, a lifelong learning strategy has been applied to address the catastrophic forgetting issue in deep models when sequentially estimating emotion patterns across various cultures. The experimental results in Section 4.8 have shown that lifelong learning models outperform other conventional models and have great potential in future real-life intelligent systems.

Last but not least, aiming at answering **RQ-b (side)** where empathetic behaviours are concerned, this thesis is further dedicated to investigating empathic behaviour analysis with deep learning approaches (cf. Section 3.4). For this purpose, a deep autoencoder-based model has been proposed and presented to detect mimicry behaviours from the acoustic perspective. Extensive experiments in Section 4.9 have indicated that empathic behaviours can be detected in human social interaction from all studied cultures.

All in all, the works presented in this thesis have demonstrated that the proposed deep learning algorithms are promising methods to contribute to the development of audiovisual emotional behaviour analysis systems, delivering better recognition performance as well as more general and robust models for practical purposes.

5.2 Limitations and Future Prospects

Despite that a set of deep learning techniques have been presented in this thesis to tackle the challenges of audiovisual emotion recognition systems such as representation learning and prediction modelling, there are several potential future research directions along this line of research worth being investigated in future.

One recent trend in developing a machine learning solution for a specific task is to combine the front-end and the back-end of a deep learning framework and to jointly learn in an end-to-end manner. Therefore, the first possibility goes to incorporate a deep representation learning model (e. g., deep latent representations or crossmodal embeddings) with one proposed emotion recognition model (e. g., strength modelling or adversarial training). It will be enlightening to examine if jointly training such a combination can further increase the performance. Note that, as the combined model becomes more complicated with more parameters to learn, more data will be desired to learn a decent model, which itself remains a challenge. Moreover, beyond audio and video signals there are also other modalities where emotions can be conveyed such as text and bio-signals. Thus, it will be interesting and useful to verify the feasibility and effectiveness of the proposed algorithms in other modalities.

Moreover, apart from main challenges tackled in this thesis, investigating other highly related, yet little addressed issues might also boost the emotion perception performance in HCI systems and further provide new and valuable insight into the challenges at hand. In particular, systems modelling cultural and linguistic diversities can result in better generalisation. Also, several emerging and promising deep learning techniques have so far not been applied in audiovisual emotion recognition, such as zero-/one-shot learning [67], reinforcement learning [137], and federated learning [218], to handle related issues such as data scarcity and data privacy. Thus, it remains challenging to apply and adapt these state-of-the-art deep learning techniques for emotional behaviour analysis in the wild.

In addition, this thesis has shed light on continual emotion recognition and empathic behaviour detection, both of which are underdeveloped topics in this field. Given the encouraging and inspiring results obtained in this work, it is worth investigating more sophisticated algorithms in future to solve these two challenging tasks. To fulfil the objectives, more advanced lifelong learning approaches other than EWC could be helpful to accomplish general emotion perception systems. Likewise, further research efforts will be dedicated to applying innovative unsupervised learning algorithms to address the empathic behaviour analysis problem.

Concluding, all algorithms and approaches developed in this thesis are aligned to reach the ultimate goal, that is, to improve the effectiveness, reliability, and robustness of emotional behaviour analysis systems operated in the wild. Hopefully, the work presented here can inspire other researchers in this community and expedite the pace at which deep learning approaches strengthen automatic audiovisual emotion recognition and cultivate its application in real-life products.

Acronyms

AE . . . . . auto-encoder
autoML . . . . . automated Machine Learning
AVEC . . . . . Audio Video Emotion Challenge
BLSTM . . . . . Bidirectional Long Short-Term Memory
BN . . . . . Bayesian Network
BoAW . . . . . Bag-of-Audio-Words
BoCAW . . . . . Bag-of-Context-Aware-Words
BoVW . . . . . Bag-of-Video-Words
CCC . . . . . Concordance Correlation Coefficient
CGAN . . . . . Conditional Generative Adversarial Network
CNN . . . . . Convolutional Neural Network
ComParE . . . . . Computational Paralinguistic ChallengeE
DBN . . . . . Deep Belief Net
DDAT . . . . . Dynamic Difficulty Awareness Training
DLBoF . . . . . Dual-Layer Bag-of-Frames
DNN . . . . . Deep Neural Network
DSSM . . . . . Deep Structured Semantic Model
ECG . . . . . Electrocardiogram
EEG . . . . . Electroencephalogram
EMG . . . . . Electromyogram
EmotiW . . . . . Emotion Recognition in the Wild Challenge

EWC . . . . . Elastic Weight Consolidation
EWE . . . . . Evaluator Weighted Estimator
FAU . . . . . Facial Action Unit
FER . . . . . Facial Emotion Recognition
FNN . . . . . Feed-forward Neural Network
GAN . . . . . Generative Adversarial Network
GRU . . . . . Gated Recurrent Unit
HCI . . . . . Human-Computer Interaction
HMM . . . . . Hidden Markov Model
HOG . . . . . Histogram of Oriented Gradients
JS . . . . . Jensen-Shannon
KL . . . . . Kullback-Leibler
LBP . . . . . Local Binary Pattern
LBP-TOP . . . . . Local Binary Patterns from Three Orthogonal Planes
LDA . . . . . Linear Discriminant Analysis
LDR . . . . . Latent Discriminative Representation
LGBP-TOP . . . . Local Gabor Binary Patterns from Three Orthogonal Planes
LLD . . . . . Low-Level Descriptor
LPCC . . . . . Linear Prediction Cepstral Coefficient
LPQ . . . . . Local Phase Quantisation
LSTM . . . . . Long Short-Term Memory
MEC . . . . . Multimodal Emotion Recognition Challenge
MFCC . . . . . Mel-Frequency Cepstral Coefficient
MSE . . . . . Mean Squared Error
MTL . . . . . Multi-Task Learning
OA-RVM . . . . . Output Associative Relevance Vector Machine
OMG . . . . . One-Minute Gradual Emotion Recognition
PCA . . . . . Principle Component Analysis
PCC . . . . . Pearson's Correlation Coefficient
PU . . . . . Perceptio Uncertainty

RE . . . . . Reconstruction Error
RELOCA . . . . . Remote COLlaborative and Affective interactions dataset
RF . . . . . Random Forests
RMSE . . . . . Root Mean Square Error
RNN . . . . . Recurrent Neural Network
SER . . . . . Speech Emotion Recognition
SEWA . . . . . Automatic Sentiment Analysis in the Wild database
SIFT . . . . . Scale-Invariant Feature Transform
SVM . . . . . Support Vector Machine
SVR . . . . . Support Vector Machine for Regression
t-SNE . . . . . t-Distributed Stochastic Neighbor Embedding
UAR . . . . . Unweighted Average Recall
WGAN . . . . . Wasserstein Generative Adversarial Network

List of Symbols

n ..... number of samples/ tasks

k ..... number of classes

Φ ..... mapping function

$\mathbf{x}$ ..... feature vector

$\mathbf{x}_i$ ..... feature vector, indexed by i

$\mathbf{x}_t$ ..... feature vector, at time t

$\mathbf{x}_i^+$ ..... feature vector that has the same label as $\mathbf{x}_i$

$\mathbf{x}_i^-$ ..... feature vector that has a different label with $\mathbf{x}_i$

$\mathbf{e}$ ..... latent representation vector, i. e., embedding

d ..... distance between two feature vectors

γ ..... predefined base of exponential function

$\mathbb{R}^n$ ..... n-dimensional vector space of real numbers

M, N ..... number of dimensions of features

E ..... number of dimensions of embeddings

τ ..... triplet

$\mathcal{J}(\theta)$ ..... objective function

$\mathcal{X}$ ..... a set of feature vectors

$\mathcal{C}$ ..... codebook

$\mathbf{c}$ ..... codeword

$\mathbf{c}_k$ ..... codeword, indexed by k

K ..... number of codewords in a codebook

ϕ . . . . . feature vector obtained via vector quantisation
n_a . . . . . number of multiple assignments
$\mathcal{N}$ . . . . . a set of codewords
S . . . . . segment
$\mathbf{h}_S$ . . . . . histogram representation with respect to the segment S
n_s . . . . . number of frames in a segment S
f . . . . . predicting or mapping function
y_t . . . . . prediction at time t
$y_{i,t}$ . . . . . prediction by the i-th model at time t
$\mathbf{x}_t$ . . . . . feature vector, at time t
$\mathbf{x}^A$ . . . . . audio feature vector
$\mathbf{x}^V$ . . . . . visual feature vector
ϵ . . . . . bias
w_i . . . . . weight of the i-th model
$\mathbf{d}_t$ . . . . . difficulty indicator at time t
L_{emt} . . . . . loss function of emotion prediction
L_{re} . . . . . loss function of input reconstruction
R_θ . . . . . regularisation term with respect to parameters θ
T . . . . . period of time
T_i . . . . . the i-th task
$\hat{\mathbf{x}}_t$ . . . . . estimation of $\mathbf{x}_t$
$\hat{y}_t$ . . . . . estimation of y_t
I, J . . . . . number of epochs during training
L_{pu} . . . . . loss function of perception uncertainty prediction
u_t . . . . . perception uncertainty of $\mathbf{x}_t$
$y_{t,i}$ . . . . . the i-th annotation of $\mathbf{x}_t$
$\bar{y}_t$ . . . . . mean value of all $y_{t,i}$ at time t
$\mathbf{x}'$ . . . . . updated input feature vector
$d_{i,t}$ . . . . . dynamic difficulty indicator of the i-th model at time t
$y'_{i,t}$ . . . . . dynamic prediction by the i-th model at time t

w_d ..... contribution of the difficulty indicator to generate a dynamic prediction

G ..... function of a generator

D ..... function of a discriminator

$\mathcal{L}$ ..... loss function

$\mathcal{L}'$ ..... new loss function in EWC

p_{data} ..... probability of real data

p_z ..... probability of latent random vector

$\mathbf{z}$ ..... latent random vector

θ ..... set of tuning parameters

θ_g ..... parameters of a generator

θ_d ..... parameters of a discriminator

θ_i^* ..... configurations of θ which lead to a good performance of the i-th task

Θ_i^* ..... a set of θ_i^*

$\mathcal{D}$ ..... data set

$\mathcal{D}_i$ ..... the i-th data subset

F ..... the Fisher information matrix

c ..... conditional information

$P(\mathbf{x}, \hat{y})$ ..... joint probability distribution

α, β, λ ..... hyperparameter that needs to be predefined

σ_x ..... standard deviation of time series x

μ_x ..... mean of variables of time series x

Bibliography

[1] M. Abdelwahab and C. Busso. Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2423–2435, Dec. 2018.

[2] S. M. Alarcao and M. J. Fonseca. Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing*, June 2017. 20 pages, in press.

[3] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear. Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, 9(4):478–490, Oct. 2018.

[4] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proc. Conference on Human Language Technology (HLT) and Empirical Methods in Natural Language Processing (EMNLP)*, pages 579–586, Vancouver, Canada, 2005.

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

[6] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. *CoRR*, abs/1701.04862, Jan. 2017.

[7] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proc. the 34th International Conference on Machine Learning (ICML)*, pages 214–223, Sydney, Australia, 2017.

[8] Y. Aytar, C. Vondrick, and A. Torralba. SoundNet: Learning sound representations from unlabeled video. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 892–900, Barcelona, Spain, 2016.

[9] K. Bahreini, R. Nadolski, and W. Westera. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments*, 24(3):590–605, Apr. 2016.

[10] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, Feb. 2019.

[11] T. Baltrušaitis, P. Robinson, and L.-P. Morency. OpenFace: An open source facial behavior analysis toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, Lake Placid, NY, 2016.

[12] T. Baltrušaitis, P. Robinson, and L.-P. Morency. OpenFace: An open source facial behavior analysis toolkit. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, Lake Placid, NY, 2016.

[13] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter. The OMG-emotion behavior dataset. In *Proc. International Joint Conference on Neural Networks (IJCNN)*, pages 1408–1412, Rio, Brazil, 2018.

[14] A. Beatty. Anthropology and emotion. *Journal of the Royal Anthropological Institute*, 20(3):545–563, May 2014.

[15] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug. 2013.

[16] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proc. 26th Annual International Conference on Machine Learning (ICML)*, pages 41–48, Montreal, Canada, 2009.

[17] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual detection of behavioural mimicry. In *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 123–128, Geneva, Switzerland, 2013.

[18] K. Boehner, R. DePaula, P. Dourish, and P. Sengers. Affect: From information to interaction. In *Proc. 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, pages 59–68, Aarhus, Denmark, 2005.

[19] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. M. Campbell, C. K. Dagli, and T. S. Huang. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proc. 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 97–104, Amsterdam, The Netherlands, 2016.

[20] S. Braun, D. Neil, and S. Liu. A curriculum learning method for improved noise robustness in automatic speech recognition. In *Proc. 25th European Signal Processing Conference (EUSIPCO)*, pages 548–552, Kos, Greece, 2017.

[21] S. Brave, C. Nass, and K. Hutchinson. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62(2):161–178, Feb. 2005.

[22] J. K. Burgoon and A. E. Hubbard. *Cross-cultural and intercultural applications of expectancy violations theory and interaction adaptation theory*, pages 149–171. Sage Thousand Oaks, CA, US, 2005.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of German emotional speech. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1517–1520, Lisbon, Portugal, 2005.

[24] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proc. International Conference on Multimodal Interfaces (ICMI)*, pages 205–211, State College, PA, 2004.

[25] J. M. Carroll and J. A. Russell. Facial expressions in Hollywood's protrayal of emotion. *Journal of Personality and Social Psychology*, 72(1):164, Jan. 1997.

[26] C.-Y. Chang, C.-W. Chang, J.-Y. Zheng, and P.-C. Chung. Physiological emotion analysis using support vector regression. *Neurocomputing*, 122:79–87, Dec. 2013.

[27] T. L. Chartrand and J. A. Bargh. The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, June 1999.

[28] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proc. the European Conference on Computer Vision (ECCV)*, pages 532–547, Munich, Germany, 2018.

[29] M. Chen, X. He, J. Yang, and H. Zhang. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, Oct. 2018.

[30] S. Chen, Q. Jin, J. Zhao, and S. Wang. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proc. the 7th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 19–26, San Francisco, CA, 2017.

[31] Z. Chen and B. Liu. *Lifelong Machine Learning*. Morgan & Claypool, San Rafael, CA, 2018.

[32] Z. Chen, N. Ma, and B. Liu. Lifelong learning for sentiment classification. In *Proc. the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL)*, pages 750–756, Beijing, China, 2015.

[33] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proc. Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, pages 103–111, Doha, Qatar, 2014.

[34] C. Clavel and Z. Callejas. Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7(1):74–93, Jan. 2016.

[35] C. Clopath. Synaptic consolidation: An approach to long-term learning. *Cognitive Neurodynamics*, 6(3):251–257, June 2012.

[36] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, Abingdon, UK, 2013.

[37] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1548–1568, Aug. 2016.

[38] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, Jan. 2001.

[39] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, Jan. 2018.

[40] T. Dang, V. Sethu, J. Epps, and E. Ambikairajah. An investigation of emotion prediction uncertainty using gaussian mixture regression. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1248–1252, Stockholm, Sweden, 2017.

[41] 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 *Proc. 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 27–35, Mountain View, CA, 2017.

[42] L. C. De Silva, T. Miyasato, and R. Nakatsu. Facial emotion recognition using multi-modal information. In *Proc. International Conference on Information, Communications and Signal Processing (ICICS)*, pages 397–401, Singapore, Singapore, 1997.

[43] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, July 2012.

[44] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, and B. Schuller. Speech-based diagnosis of autism spectrum condition by generative adversarial network representations. In *Proc. International Conference on Digital Health*, pages 53–57, London, UK, 2017.

[45] J. Deng, W. Han, and B. Schuller. Confidence measures for speech emotion recognition: A start. In *Proc. the 10th ITG Conference on Speech Communication*, pages 1–4, Braunschweig, Germany, 2012.

[46] J. Deng, Z. Zhang, F. Eyben, and B. Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072, Sep. 2014.

[47] J. Deng, Z. Zhang, E. Marchi, and B. Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 511–516, Geneva, Swizerland, 2013.

[48] L. Devillers, L. Vidrascu, and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407 – 422, May 2005.

[49] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proc. 15th ACM on International Conference on Multimodal Interaction (ICMI)*, pages 509–516, Sydney, Australia, 2013.

[50] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages 653–656, Boulder, CO, 2018.

[51] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. *Proc. Advances in Neural Information Processing Systems (NIPS)*, 9:155–161, 1997.

[52] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proc. ACM on International Conference on Multimodal Interaction (ICMI)*, pages 467–474, Seattle, WA, 2015.

[53] P. Ekman. Universals and cultural differences in facial expressions of emotion. *California Mental Health Research Digest*, 8(4):151–158, Sep. 1970.

[54] P. Ekman. *Basic Emotions*, chapter 3, pages 45–60. John Wiley & Sons, Ltd, 1999.

[55] P. Ekman and W. V. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, Los Altos, CA, USA, 1st edition, 2003.

[56] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, Mar. 2011.

[57] R. El Kaliouby, R. Picard, and S. Baron-Cohen. Affective computing and autism. *Annals of the New York Academy of Sciences*, 1093(1):228–248, Feb. 2007.

[58] F. Eyben. *Real-time speech and music classification by large audio feature space extraction*. Springer, Basel, Switzerland, 1st edition, 2016.

[59] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, Apr. 2016.

[60] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proc. ACM International Conference on Multimedia (MM)*, pages 835–838, Barcelona, Spain, 2013.

[61] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – the Munich versatile and fast open-source audio feature extractor. In *Proc. ACM International Conference on Multimedia (MM)*, pages 1459–1462, Florence, Italy, 2010.

[62] 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, June 2008.

[63] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1):259–275, Jan. 2003.

[64] W. Fedus, I. Goodfellow, and A. M. Dai. MaskGAN: Better text generation via filling in the _. In *Proc. International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[65] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, Jan. 2017.

[66] E. Friesen and P. Ekman. *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[67] Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, and S. Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, Jan. 2018.

[68] A. Gepperth and C. Karaoguz. A bio-inspired incremental learning architecture for applied perceptual problems. *Cognitive Computation*, 8(5):924–934, Oct. 2016.

[69] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer. Representation learning for speech emotion recognition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3603–3607, San Francisco, CA, 2016.

[70] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer. Technique for automatic emotion recognition by body gesture analysis. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–6, Anchorage, AK, 2008.

[71] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT press, 2016.

[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. the 27th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680, Montreal, Canada, 2014.

[73] A. Graves. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer, Heidelberg, Germany, 2012.

[74] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610, Jul 2005.

[75] M. Grimm and K. Kroschel. Evaluation of natural emotions using self assessment manikins. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 381–385, Cancún, Mexico, 2005.

[76] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, May 2010.

[77] N. Gueguen, C. Jacob, and A. Martin. Mimicry in social interaction: Its effect on human judgment and behavior. *European Journal of Social Sciences*, 8(2):253–259, 2009.

[78] L. Gui, T. Baltrušaitis, and L. P. Morency. Curriculum learning for facial expression recognition. In *Proc. 12th IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 505–511, Washington, DC, 2017.

[79] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proc. International Conference on Automatic Face and Gesture Recognition (FG)*, pages 827–834, Santa Barbara, CA, 2011.

[80] S. R. Gunn. Support vector machines for classification and regression. Technical Report 14, School of Electronics and Computer Science, University of Southampton, Southampton, England, May 1998.

[81] B.-J. Han, S. Rho, S. Jun, and E. Hwang. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460, May 2010.

[82] J. Han, M. Schmitt, and B. Schuller. You sound like your counterpart: Interpersonal speech analysis. In *Proc. 20th International Conference of Speech and Computer (SPECOM)*, pages 188–197, Leipzig, Germany, 2018.

[83] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller. Strength modelling for real-world automatic continuous affect recognition from audiovisual signals. *Image and Vision Computing*, 65:76–86, Sep. 2017.

[84] J. Han, Z. Zhang, N. Cummins, and B. Schuller. Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives. *IEEE Computational Intelligence Magazine*, 14(2):68–81, May 2019.

[85] J. Han, Z. Zhang, G. Keren, and B. Schuller. Emotion recognition in speech with latent discriminative representations learning. *Acta Acustica united with Acustica*, 104(5):737–740, Sep. 2018.

[86] J. Han, Z. Zhang, Z. Ren, F. Ringeval, and B. Schuller. Towards conditional adversarial training for predicting emotions from speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6822–6826, Calgary, Canada, 2018.

[87] J. Han, Z. Zhang, Z. Ren, and B. Schuller. EmoBed: Strengthening Monomodal EmotionRecognition via Training with CrossmodalEmotion Embeddings. *IEEE Transactions on Affective Computing*, 10, 2019. 12 pages, to appear.

[88] J. Han, Z. Zhang, Z. Ren, and B. Schuller. Implicit fusion by joint audiovisual training for emotion recognition in mono modality. In *Proc. 44th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5861–5865, Brighton, UK, 2019.

[89] J. Han, Z. Zhang, F. Ringeval, and B. Schuller. Prediction-based learning for continuous emotion recognition in speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5005–5009, New Orleans, LA, 2017.

[90] J. Han, Z. Zhang, F. Ringeval, and B. Schuller. Reconstruction-error-based learning for continuous emotion recognition in speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2367–2371, New Orleans, LA, 2017.

[91] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proc. ACM International Conference on Multimedia (MM)*, pages 890–897, Mountain View, CA, 2017.

[92] J. Han, Z. Zhang, M. Schmitt, Z. Ren, F. Ringeval, and B. Schuller. Bags in bag: Generating context-aware bags for tracking emotions from speech.

In *Proc. 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3082–3086, Hyderabad, India, 2018.

[93] S. Hantke, F. Eyben, T. Appel, and B. Schuller. iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 891–897, Xi'an, China, 2015.

[94] S. L. Happy and A. Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE Transactions on Affective Computing*, 6(1):1–12, Jan. 2015.

[95] 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 *Proc. 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 73–80, Brisbane, Australia, 2015.

[96] U. Hess and A. Fischer. Emotional mimicry as social regulation. *Personality and Social Psychology Review*, 17(2):142–157, May 2013.

[97] U. Hess and A. Fischer. Emotional mimicry: Why and when we mimic emotions. *Social and Personality Psychology Compass*, 8(2):45–57, 2014.

[98] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997.

[99] 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 *Proc. the 7th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 11–18, San Francisco, CA, 2017.

[100] 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 *Proc. 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 41–48, Brisbane, Australia, 2015.

[101] Z. Huang, M. Dong, Q. Mao, and Y. Zhan. Speech emotion recognition using cnn. In *Proc. 22nd ACM International Conference on Multimedia (MM)*, pages 801–804, Orlando, FL, 2014.

[102] R. Jozefowicz, W. Zaremba, and I. Sutskever. An empirical exploration of recurrent network architectures. In *Proc. International Conference on Machine Learning (ICML)*, pages 2342–2350, Lille, France, 2015.

[103] P. Karanasou, Y. Wang, M. J. Gales, and P. C. Woodland. Adaptation of deep neural network acoustic models using factorised i-vectors. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2180–2184, Singapore, Singapore, 2014.

[104] G. Keren and B. Schuller. Convolutional RNN: An enhanced model for extracting features from sequential data. In *Proc. International Joint Conference on Neural Networks (IJCNN)*, pages 3412–3419, Vancouver, Canada, 2016.

[105] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis, and E. M. Provost. Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition. In *Proc. INTERSPPECH*, pages 1253–1257, Stockholm, Sweden, 2017.

[106] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, Dec. 2008.

[107] J. Kim, G. Englebienne, K. P. Truong, and V. Evers. Deep temporal models using identity skip-connections for speech emotion recognition. In *Proc. 25th ACM International Conference on Multimedia (MM)*, pages 1006–1013, 2017.

[108] J. Kim, G. Englebienne, K. P. Truong, and V. Evers. Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1113–1117, Stockholm, Sweden, 2017.

[109] Y. Kim, H. Lee, and E. M. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3687–3691, Vancouver, BC, 2013.

[110] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015. 15 pages.

[111] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, Mar. 2017.

[112] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, E. Hajiyev, and M. Pantic. SEWA DB: A rich database for audio-visual emotion and sentiment research in the

wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. to appear.

[113] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, Lake Tahoe, NV, 2012.

[114] H. Lee, P. Pham, Y. Largman, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1096–1104, Vancouver, Canada, 2009.

[115] J. Lee and I. Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany, 2015.

[116] I. Leite, A. Pereira, S. Mascarenhas, C. Martinho, R. Prada, and A. Paiva. The influence of empathy in human–Robot relations. *International journal of human-computer studies*, 71(3):250–260, Mar. 2013.

[117] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia. MEC 2016: The multimodal emotion recognition challenge of CCPR 2016. In *Proc. Chinese Conference on Pattern Recognition (CCPR)*, pages 667–678, Chengdu, China, 2016.

[118] Y. Li, S. Wang, Q. Tian, and X. Ding. A survey of recent advances in visual feature detection. *Neurocomputing*, 149:736–751, Feb. 2015.

[119] H. Lim, M. J. Kim, and H. Kim. Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3325–3329, Dresden, Germany, 2015.

[120] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1805–1812, Columbus, OH, 2014.

[121] X. Liu, M. Masana, L. Herranz, J. Van de Weijer, A. M. Lopez, and A. D. Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *Proc. 24th International Conference on Pattern Recognition (ICPR)*, pages 2262–2268, Beijing, China, 2018.

[122] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu. Coherent bag-of audio words model for efficient large-scale video copy detection. In *Proc.*

ACM International Conference on Image and Video Retrieval (CIVR), pages 89–96, Xi'an, China, 2010.

[123] R. Lotfian and C. Busso. Curriculum learning for speech emotion recognition from crowdsourced labels. *arXiv preprint arXiv:1805.10339*, May 2018.

[124] D. Luo, Y. Zou, and D. Huang. Investigation on joint representation learning for robust feature extraction in speech emotion recognition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 152–156, Hyderabad, India, 2018.

[125] A. Majid. Current emotion research in the language sciences. *Emotion Review*, 4(4):432–443, July 2012.

[126] A. Manandhar, K. D. Morton, P. A. Torrione, and L. M. Collins. Multivariate output-associative RVM for multi-dimensional affect predictions. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 10(3):408–415, Jan. 2016.

[127] Q. Mao, M. Dong, Z. Huang, and Y. Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213, Dec. 2014.

[128] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang, and S. Smolley. Least squares generative adversarial networks. *arXiv preprint ArXiv:1611.04076*, 2016.

[129] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller. A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks. In *Proc. IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, pages 1996–2000, Brisbane, Australia, 2015.

[130] S. Mariooryad and C. Busso. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108, Apr. 2015.

[131] I. B. Mauss and M. D. Robinson. Measures of emotion: A review. *Cognition and Emotion*, 23(2):209–237, Feb. 2009.

[132] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Jan. 1989.

[133] 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, Apr. 2016.

[134] A. Metallinou, S. Lee, and S. Narayanan. Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2462–2465, Dallas, TX, 2010.

[135] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198, Jan. 2012.

[136] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, June 2014.

[137] T. M. Moerland, J. Broekens, and C. M. Jonker. Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, 107(2):443–480, Feb. 2018.

[138] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *Proc. IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10, Lake Placid, NY, 2016.

[139] R. Morris, D. McDuff, and R. Calvo. *Crowdsourcing techniques for affective computing*, pages 384–394. Oxford Univ. Press, 2014.

[140] V.-e. Neagoe, A. Barar, N. Sebe, and P. Robitu. A deep learning approach for subject independent emotion recognition from facial expressions. In *Proc. International Conference on Image Processing and Pattern Recognition (IPPR)*, pages 93–98, Budapest, Hungary, 2013.

[141] M. 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, Apr. 2011.

[142] M. Nicolaou, H. Gunes, and M. Pantic. Output-associative RVM regression for dimensional and continuous emotion prediction. *Image and Vision Computing*, 30(3):186–196, Mar. 2012.

[143] A. Nicolle and V. Goel. Differential impact of beliefs on valence and arousal. *Cognition & Emotion*, 27(2):263–272, Feb. 2013.

[144] A. Ortony, G. L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge university press, 1990.

[145] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, C. Zoll, and L. Hall. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proc. International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 194–201, New York, NY, 2004.

[146] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, Sep. 2003.

[147] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, May 2019.

[148] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. British Machine Vision Conference (BMVC)*, pages 1–12, Swansea, UK, 2015.

[149] F. Parrill and I. Kimbara. Seeing and hearing double: The influence of mimicry in speech and gesture on observers. *Journal of Nonverbal Behavior*, 30(4):157, 2006.

[150] A. Pereira, I. Leite, S. Mascarenhas, C. Martinho, and A. Paiva. Using empathy to improve human-robot relationships. In *Proc. International Conference on Human-Robot Personal Relationship (HRPR)*, pages 130–138, Leiden, Netherlands, 2010.

[151] S. Petridis and M. Pantic. Prediction-based audiovisual fusion for classification of non-linguistic vocalisations. *IEEE Transactions on Affective Computing*, 7(1):45–58, Jan. 2016.

[152] R. W. Picard. *Affective Computing*. MIT press, Cambridge, MA, 1997.

[153] G. Pons and D. Masip. Supervised committee of convolutional neural networks in automated facial expression analysis. *IEEE Transactions on Affective Computing*, 9(3):343–350, July 2018.

[154] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, Jan. 2016.

[155] M. I. Posner and S. E. Petersen. The attention system of the human brain. *Annual Review of Neuroscience*, 13(1):25–42, 1990.

[156] L. L. Presti and M. L. Cascia. Boosting hankel matrices for face emotion recognition and pain detection. *Computer Vision and Image Understanding*, 156:19–33, Mar. 2017.

[157] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proc. International Conference on Learning Representation (ICLR)*, San Juan, PR, 2016. no pagination.

[158] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze. Robust audio-codebooks for large-scale event detection in consumer videos. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2929–2933, Lyon, France, 2013.

[159] F. Ringeval, F. Eyben, E. Kroupi, A. Yüce, J. 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, Nov. 2015.

[160] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Ciftçi, H. Güleç, A. Salah, and M. Pantic. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proc. 8th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 3–13, Seoul, South Korea, 2018.

[161] F. Ringeval, B. Schuller, M. Valstar, et al. AVEC 2019 workshop and challenge: State-of-mind, depression with ai, and cross-cultural affect recognition. In *Proc. 9th AudioVisual Emotion Challenge (AVEC) associated with ACM Multimedia*, Nice, France, 2019. 10 pages.

[162] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmi, and M. Pantic. AVEC 2017–Real-life depression, and affect recognition workshop and challenge. In *Proc. 7th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 3–10, Mountain View, CA, 2017.

[163] 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 *Proc. 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 3–8, Brisbane, Australia, 2015.

[164] F. Ringeval, A. Sonderegger, J. S. Sauer, and D. Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interac-

tions. In *Proc. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, Shanghai, China, 2013.

[165] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, Dec. 1980.

[166] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, Jan. 2003.

[167] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, Sep. 2016.

[168] O. C. Santos, M. Saneiro, J. G. Boticario, and M. C. Rodriguez-Sanchez. Toward interactive context-aware affective educational recommendations in computer-assisted language learning. *New Review of Hypermedia and Multimedia*, 22(1-2):27–57, Jan. 2016.

[169] 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.

[170] M. Schmitt, F. Ringeval, and B. Schuller. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 495–499, San Francisco, CA, 2016.

[171] M. Schmitt and B. Schuller. openXBOW—-Introducing the Passau open-source crossmodal bag-of-words toolkit. *Journal of Machine Learning Research*, 18(96):1–5, Oct. 2017.

[172] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, Boston, MA, 2015.

[173] B. Schuller. *Intelligent Audio Analysis*. Signals and Communication Technology. Springer, 2013.

[174] B. Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, May 2018.

[175] B. Schuller. The handbook of multimodal-multisensor interfaces. chapter Multimodal User State and Trait Recognition: An Overview, pages 129–165. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA, 2019.

[176] B. Schuller and A. Batliner. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing.* John Wiley & Sons, Hoboken, NJ, 2013.

[177] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger. The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, parkinson & eating condition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 478–482, Dresden, Germany, 2015. ISCA.

[178] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2001–2005, San Francisco, CA, 2016.

[179] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 148–152, Lyon, France, 2013.

[180] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011 – The first international audio/visual emotion challenge. In *Proc. the 4th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 415–424, Memphis, TN, 2011.

[181] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell. Progress & compress: A scalable framework for continual learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 4535–4544, Stockholm, Sweden, 2018.

[182] L. E. Scissors, A. J. Gill, and D. Gergle. Linguistic mimicry and trust in text-based cmc. In *Proc. the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 277–280, San Diego, CA, 2008.

[183] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Emotion recognition based on joint visual and audio cues. In *Proc. 18th International Conference on Pattern Recognition (ICPR)*, pages 1136–1139, Hong Kong, China, 2006.

[184] M. L. Seltzer, D. Yu, and Y. Wang. An investigation of deep neural networks for noise robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7398–7402, Vancouver, Canada, 2013.

[185] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proc. International Conference on Computer Vision Workshop (ICCVW)*, pages 1003–1011, Santiago, Chile, 2015.

[186] 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*, 7(1):17–28, Jan. 2016.

[187] M. Song, J. Bu, C. Chen, and N. Li. Audio-visual based emotion recognition – a new approach. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, Washington, DC, 2004.

[188] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, and S. C. Levinson. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, June 2009.

[189] X. Sun, A. Nijholt, K. P. Truong, and M. Pantic. Automatic visual mimicry expression analysis in interpersonal interaction. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 40–46, Colorado Springs, CO, 2011.

[190] R. I. Swaab, W. W. Maddux, and M. Sinaceur. Early words that work: When and how virtual linguistic mimicry facilitates negotiation outcomes. *Journal of Experimental Social Psychology*, 47(3):616–621, 2011.

[191] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1–9, Boston, MA, 2015.

[192] B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proc. the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2062–2068, Minneapolis, Minnesota, 2019.

[193] S. Thrun and T. M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15(1-2):25–46, Jan. 1995.

[194] L. Tian, J. D. Moore, and C. Lai. Emotion recognition in spontaneous and acted dialogues. In *Proc. Affective Computing and Intelligent Interaction (ACII)*, pages 698–704, Xi'an, China, 2015.

[195] M. Tkalčič, B. De Carolis, M. De Gemmis, A. Odić, and A. Košir. *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*. Springer, Basel, Switzerland, 2016.

[196] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, Shanghai, China, 2016.

[197] P. Tzirakis, G. Trigeorgis, M. Nicolaou, B. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, Dec. 2017.

[198] 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 *Proc. the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 3–10, Amsterdam, The Netherlands, 2016.

[199] G. K. Verma and U. S. Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, 102:162–172, Nov. 2014.

[200] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proc. International Conference on Machine Learning (ICML)*, pages 1096–1103, Helsinki, Finland, 2008.

[201] T. Vogt and E. André. Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conference (LREC)*, pages 1123–1126, Genoa, Italy, 2006.

[202] J. Wagner, J. Kim, and E. André. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pages 940–943, Amsterdam, Netherlands, 2005.

[203] J. Wagner, D. Schiller, A. Seiderer, and E. André. Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant? In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 147–151, Hyderabad, India, 2018.

[204] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang. Generative adversarial networks: Introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598, Sep. 2017.

[205] D. A. Washburn and R. Putney. Attention and task difficulty: When is performance facilitated? *Learning and Motivation*, 32(1):36–47, Feb. 2001.

[206] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller. Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2196–2202, New York, NY, 2016.

[207] J. M. Wilce. *Language and Emotion*. Cambridge University Press, Cambridge, UK, 2009.

[208] J. R. Wilson. Towards an affective robot capable of being a long-term companion. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 754–759, Xi'an, China, 2015.

[209] M. Wimmer, B. Schuller, D. Arsic, B. Radig, and G. Rigoll. Low-level fusion of audio and video feature for multi-modal emotion recognition. In *Proc. International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 145–151, Funchal, Portugal, 2008.

[210] L. Wisp. *History of the Concept of Empathy*, pages 17–37. Cambridge University Press, 1987.

[211] 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 *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 597–600, Brisbane, Australia, 2008.

[212] 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.

[213] C.-H. Wu, J.-C. Lin, and W.-L. Wei. Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing*, 3(12):1–18, Nov. 2014.

[214] R. Xia, J. Deng, B. Schuller, and Y. Liu. Modeling gender information for emotion recognition using denoising autoencoder. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 990–994, Florence, Italy, 2014.

[215] R. Xia and Y. Liu. Using denoising autoencoder for emotion recognition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2886–2889, Lyon, France, 2013.

[216] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1511–1519, Santiago, Chile, 2015.

[217] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 4633–4641, Santiago, Chile, 2015.

[218] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, Feb. 2019.

[219] C.-C. M. Yeh, L. Su, and Y.-H. Yang. Dual-layer bag-of-frames model for music genre classification. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250, Vancouver, Canada, 2013.

[220] J. Yoon, E. Yang, J. Lee, and S. J. Hwang. Lifelong learning with dynamically expandable networks. In *Proc. International Conference on Learning Representations (ICLR)*, New Orleans, LA, 2018. 11 pages.

[221] L. Yu, W. Zhang, J. Wang, and Y. Yu. SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, pages 2852–2858, San Francisco, CA, 2017.

[222] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages 435–442, Seattle, WA, 2015.

[223] 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.

[224] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. S. Huang, and S. Levinson. Audio-visual affect recognition through multi-stream fused HMM for HCI. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 967–972, San Diego, CA, 2005.

[225] B. Zhang, E. M. Provost, and G. Essl. Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences. *IEEE Transactions on Affective Computing*, 10(1):85–99, Jan. 2019.

[226] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct. 2016.

[227] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller. Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4990–4994, New Orleans, LA, 2017.

[228] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller. Distributing recognition in computational paralinguistics. *IEEE Transactions on Affective Computing*, 5(4):406–417, Oct. 2014.

[229] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(1):115–126, Jan. 2015.

[230] Z. Zhang, N. Cummins, and B. Schuller. Advanced data exploitation for speech analysis – An overview. *IEEE Signal Processing Magazine*, 34(4):107–129, July 2017.

[231] Z. Zhang, J. Deng, E. Marchi, and B. Schuller. Active learning by label uncertainty for acoustic emotion recognition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2856–2860, Lyon, France, 2013.

[232] Z. Zhang, F. Eyben, J. Deng, and B. Schuller. An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena. In *Proc. 5th International Workshop on Emotion Social Signals, Sentiment & Linked Open Data, satellite of LREC*, pages 21–26, Reykjavik, Iceland, 2014.

[233] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller. Snore-GANs: Improving automatic snore sound classification with synthesised data. *IEEE Journal of Biomedical and Health Informatics*, 2019. to appear.

[234] Z. Zhang, J. Han, and B. Schuller. Dynamic difficulty awareness training for continuous emotion prediction. *IEEE Transactions on Multimedia*, PP, Sep. 2018. 13 pages.

[235] Z. Zhang, J. Han, X. Xu, J. Deng, F. Ringeval, and B. Schuller. Leveraging unlabelled data for emotion recognition with enhanced collaborative semi-supervised learning. *IEEE Access*, 6(1):22196–22209, Dec. 2018.

[236] Z. Zhang, J. Pinto, C. Plahl, B. Schuller, and D. Willett. Channel mapping using bidirectional long short-term memory for dereverberation in hand-free voice controlled devices. *IEEE Transactions on Consumer Electronics*, 60(3):525–533, Aug. 2014.

[237] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller. Enhanced semi-supervised learning for multimodal emotion recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5185–5189, Shanghai, China, 2016.

[238] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller. Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3593–3597, San Francisco, CA, 2016.

[239] J. Zhao, R. Li, S. Chen, and Q. Jin. Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In *Proc. 8th International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 65–72, Seoul, South Korea, 2018.

[240] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, pages 730–738, New Orleans, LA, 2018.

[241] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, Mar. 2017.