Modelling Person-specific and Multi-scale Facial Dynamics for Automatic Personality and Depression Analysis

Submitted November 2020, in partial fulfillment of
the conditions for the award of the degree **PhD Computer Science.**

Siyang Song
4285836

Supervised by
Prof. Michel Valstar
Prof. Alan Johnston

School of Computer Science
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature Siyang Song

Date 30 / 11 / 2020

I hereby declare that I have all necessary rights and consents to publicly distribute this dissertation via the University of Nottingham's e-dissertation archive.

Public access to this dissertation is restricted until: DD/MM/YYYY

Abstract

'To know oneself is true progress'. While one's identity is difficult to be fully described, a key part of it is one's personality. Accurately understanding personality can benefit various aspects of human's life. There is convergent evidence suggesting that personality traits are marked by non-verbal facial expressions of emotions, which in theory means that automatic personality assessment is possible from facial behaviours. Thus, this thesis aims to develop video-based automatic personality analysis approaches. Specifically, two video-level dynamic facial behaviour representations are proposed for automatic personality traits estimation, namely person-specific representation and spectral representation, which focus on addressing three issues that have been frequently occurred in existing automatic personality analysis approaches: 1. attempting to use super short video segments or even a single frame to infer personality traits; 2. lack of proper way to retain multi-scale long-term temporal information; 3. lack of methods to encode person-specific facial dynamics that are relatively stable over time but differ across individuals.

This thesis starts with extending the dynamic image algorithm to modeling preceding and succeeding short-term face dynamics of each frame in a video, which achieved good performance in estimating valence/arousal intensities, showing good dynamic encoding ability of such dynamic representation. This thesis then proposes a novel Rank Loss, aiming to train a network that produces similar dynamic representation per-frame but only from a still image. This way, the network can learn generic facial dynamics from unlabelled face videos in a self-supervised manner. Based on such an approach, the person-specific representation encoding approach is proposed. It firstly freezes the well-trained generic network, and incorporates a set of intermediate filters, which are trained again but with only person-specific videos based on the same self-supervised learning approach. As a result, the learned filters' weights are person-specific, and can be concatenated as a 1-D video-level person-specific representation. Meanwhile, this thesis also proposes a spectral analysis approach to retain multi-scale video-level facial dynamics. This approach uses automatically detected human behaviour primitives as the low-dimensional descriptor for each frame, and converts long and variable-length time-series behaviour signals to small and length-independent spectral representations to represent video-level multi-scale temporal dynamics of expressive behaviours. Consequently, the combination of two representations, which contains not only multi-scale video-level facial dynamics but also person-specific video-level facial dynamics, can be applied to automatic personality estimation.

This thesis conducts a series of experiments to validate the proposed approaches: 1. the arousal/valence intensity estimation is conducted on both a controlled face video dataset (SEMAINE) and a wild face video dataset (Affwild-2), to evaluate the dynamic encoding capability of the proposed Rank Loss; 2. the proposed automatic personality traits recognition systems (spectral representation and person-specific representation) are evaluated on face video datasets that labelled with either ’Big-Five’ apparent personality traits (ChaLearn) or self-reported personality traits (VHQ); 3. the depression studies are also evaluated on the VHQ dataset that is labelled with PHQ-9 depression scores. The experimental results on automatic personality traits and depression severity estimation tasks show the person-specific representation’s good performance in personality task and spectral vector’s superior performance in depression task. In particular, the proposed person-specific approach achieved a similar performance to the state-of-the-art method in apparent personality traits recognition task and achieved at least 15% PCC improvements over other approaches in self-reported personality traits recognition task. Meanwhile, the proposed spectral representation shows better performance than the person-specific approach in depression severity estimation task. In addition, this thesis also found that adding personality traits labels/predictions into behaviour descriptors improved depression severity estimation results.

Acknowledgements

This thesis would not have been possible without my supervisor, Prof. Michel Valstar. I would like to sincerely thank him for his continuous and kind support for both my PhD research and life in the past four years. On the one hand, he has given me invaluable inspiration in the field of vision-based automatic human facial behaviour understanding. Most importantly, he has given me a lot of freedom and even encouraged me to investigate some ideas that seem ’strange’, making me always feel energetic and hunger about the research. On the other hand, he is always very kind to me and encourages me to enjoy my life in my spare time. I would say, Michel is one of the most decent gentlemen I have ever met in my life and it was my great honor to study under his supevision. Many thanks to Prof. Alan Johnston, my second supervisor, for always supporting me for not only the research but also my future career.

Prof. Linlin Shen is my external supervisor who works for Shenzhen University. Prof. Shen has helped me with my research since 2015. Ever since, we’ve kept in touch and he’s been an incredibly valuable source of information. Especially regarding the problem of facial analysis, he also helped me a lot with my paper writing.

Dr. Erin Solovey was my supervisor of my internship in Boston, USA. During the summer of 2018, I was her guest at Worcester polytechnic institute. Erin is a very knowledgeable researcher in HCI and has wide connections. It was a remarkable experience, as Ruixue and I were eventually collaborated with researchers at MIT and had some good results.

2020 is a challenging year that many people lost their lives because of Covid-19. This also stopped me from meeting my family at the end of this year. Nevertheless, I know that you are always supporting me. I wouldn’t have reached this point without your support.

And finally I would like to thank all the people in the college with whom we’ve always had interesting acaedemic discussions and leisure drinks. Thank you Shashank, Aaron, Keerthy, Tom, Mani, Doratha, Jing, Feng, Zhonglin, Bowen, Kike, Ioanna, Dimitrios and other friendly colleagues. I would also like to thank all the friends, I have met in UK. It was a tough but super wonderful four years!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Who am I? One's identity is a puzzling thing, and is something that we only learn about slowly throughout our life. 'Know thyself!' is a commandment common in many cultures. What exactly constitutes one's identity is perhaps not entirely known, but we do know that part of it is one's personality. As suggested by psychological studies, human personality information can help improving the understanding of human behaviour (Roberts & Jackson, 2008; Lane & Manner, 2011; Barnett, Pearson, Pearson, & Kellermanns, 2015), cognition and emotional processes (Komulainen et al., 2014; Revelle & Scherer, 2009) as well as health conditions (Huang et al., 2017; Kendler, Gardner, Gatz, & Pedersen, 2007; Hettema, Neale, Myers, Prescott, & Kendler, 2006). For example, many biological and psychological studies (Chioqueta & Stiles, 2005; De Moor, Beem, Stubbe, Boomsma, & De Geus, 2006; Kendler et al., 2007; Hettema et al., 2006) claimed that neurotic people tend to experience depressed feelings more, while extrovert individuals are less likely to suffer from depression. Given these findings, it is interesting and impactful to explore how to develop an objective and accurate automatic self-reflection tool based on one's own expressive behaviours - a mirror of one's personality.

Specifically, human personality can be defined as the characteristic set of behaviours, cognition, and emotional patterns that evolve from biological and environmental factors (Hogan, Johnson, Johnson, & Briggs, 1997), which can be reflected by the coherent patterning of affect and behaviours over time and space (Revelle & Scherer, 2009), displaying the integration over time of feelings, actions, desires and other components (Ortony, Norman, & Revelle, 2005). Since the last century, the trait-based models such as Three Factor Model (Eysenck & Eysenck, 1965), Five Factor Model (McCrae & Costa, 1987), etc., have been frequently employed to measure complex and implicit human personality. They primarily focus on evaluating the aspects of personality that are relatively stable over time but differ across individuals (Kassin, 2003). In general, personality recognition tasks can be categorized into two types (Vinciarelli & Mohammadi, 2014): 1. self-reported personality, which is reflected by a person's observable behaviours; 2. apparent (perceived) personality, which is defined as an observer's perception of the individual based on one or more cues.

While the standard approaches to evaluate personality traits are subjective as they use questionnaires based on verbal behaviour descriptors, e.g., Big Five Inventory (Cavallera, Passerini, & Pepe, 2013), multiple psychological studies (Qin, Gao, Xu, & Hu, 2018; DePaulo, 1992; Borkenau & Liebler, 1992) suggested that non-verbal behaviours also contain vital information of a human's implicit dispositions and internal states, including both static and dynamic elements of physical appearance, and these cues are differentially associated with personality traits (Naumann, Vazire, Rentfrow, & Gosling, 2009).

In particular, there is convergent evidence that personality is marked by certain facial display of emotions (John, 1990), i.e. people with different personalities may behave differently in response to the same stimulus. For example, Keltner et al. (Keltner, 1996) found that both emotion and personality traits are expressed in distinct, observable behaviours that evoke responses in others, and can interpret human nature based on a certain number of universal characters (P. Ekman, 1992b; Izard, 1977). Since facial display and behaviours are quick, reliable, and seemingly universal signals of an individual's emotions and emotions can be expressed by an ordered set of facial displays (P. Ekman, 1992a), there is converging evidence that the face can provide useful information regarding a person's emotion and personality (Knapp, Hall, & Horgan, 2013). As a result, a large part of personality prediction works (Borkenau, Brecke, Möttig, & Paelecke, 2009; DeBruine et al., 2006; Shevlin, Walker, Davies, Banyard, & Lewis, 2003; Zebrowitz, Hall, Murphy, & Rhodes, 2002) are building on headshots or above-the-waist photographs. Some studies

(J. A. Hall, Andrzejewski, Murphy, Mast, & Feinstein, 2008; Kenny, 1994; Knapp et al., 2013) showed that when targets were photographed with a spontaneous pose and facial displays, observers' judgments were accurate for almost all the traits.

Since the year (2012) when AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) was proposed, the rapid development of deep learning techniques boosts the increased number of automatic facial expression analysis systems (Jaiswal & Valstar, 2016; Chen, Chen, Chi, & Fu, 2018; K. Zhang, Huang, Du, & Wang, 2017; Martinez, Valstar, Jiang, & Pantic, 2017; K. Zhang et al., 2017; Nicolle, Bailly, & Chetouani, 2015), etc., allowing accurate predictions of individuals expressive facial behaviours to be made, which can be further utilised to reflect individuals' high-level mental status such as emotion (Song, Sánchez-Lozano, Kumar Tellamekala, et al., 2019; Song, Sánchez-Lozano, Shen, Johnston, & Valstar, 2019; Nicolle, Rapp, Bailly, Prevost, & Chetouani, 2012; Kossaifi, Tzimiropoulos, Todorovic, & Pantic, 2017; Kollias et al., 2019a; Komulainen et al., 2014), pain feeling (J. O. Egede et al., 2020; J. Egede, Valstar, & Martinez, 2017), mental health status (L. Yang, Jiang, & Sahli, 2018; Pampouchidou et al., 2016; Song, Shen, & Valstar, 2018; Song et al., 2020; Jaiswal, Song, & Valstar, 2019), and personality traits (Celiktutan & Gunes, 2015; Güçlütürk, Güçlü, van Gerven, & van Lier, 2016; L. Zhang, Peng, & Winkler, 2019; Kampman, Barezi, Bertero, & Fung, 2018; Bekhouche, Dornaika, Ouafi, & Taleb-Ahmed, 2017). However, to the best of the author's knowledge, all existing automatic personality analysis approaches (Celiktutan & Gunes, 2015; Güçlütürk et al., 2016; L. Zhang et al., 2019; Kampman et al., 2018; Bekhouche et al., 2017) failed to model person-specific properties that are stable over time for each person but differ between individuals. Additionally, the produced machine representations cannot retain multi-scale temporal dynamics. Moreover, some approaches attempted to model personality traits from a frame or very short segment, which contradicts the definition of personality traits. Specifically, while a person's frame/segment-level facial status can change rapidly in a short period, video-level (event) labels will remain unchanged for each video, as personality doesn't change on that time scale. Therefore, a brief segment (a single frame or couple of frames) of expressive behaviours may not carry sufficient information to encode a personality trait-specific pattern.

Motivated by these problems, this thesis firstly focuses on developing face video-based automatic personality analysis systems by extending the current state-of-the-art machine learning techniques. Specifically, it aims to deal with the methodological shortcomings summarized above, which have been frequently occurred in existing personality analysis systems (Sec.1.1). This can be achieved by applying the proposed two long-term facial behaviour dynamic modelling approaches, where the person-specific descriptor encodes facial person-specific dynamics for the given individual while spectral approach encodes multi-scale facial behaviour dynamics, to personality analysis task.

As discussed above, it is well known that personality impacts on many behaviours, including some important clinically relevant behaviours such as depression (Vukasović & Bratko, 2015; Lo et al., 2017; DENNIS, CHARNEY, NELSON, QUINLAN, et al., 1981; Matsudaira & Kitamura, 2006). Particularly, this thesis will also explore the effect of personality on behaviour that is indicative of the level of depression severity. The motivation is that the standard clinical depression assessment is subjective because these depend almost entirely on the health professional's own understanding of the individual's verbal psychological report, e.g. clinical interview and questionnaires completed by patients or caregivers (Cohn et al., 2009). Additionally, this is often a lengthy procedure which hinders access to early treatment. In the UK it has been reported that more than half of the patients have to wait at least 3 months before receiving talking treatment (*People with mental health problems still waiting over a year for talking treatments*, 2013). Sometimes the relevant patient information or mental health experts may not be accessible, which results in many patients missing the best chance for preventing or treating their depression at early stages of depression. This is problematic, because correct early diagnosis is an important factor in the treatment of depression. To improve this, automatic objective assessment approaches to aid monitoring and diagnosis need to be developed. While there is a lot of automatic depression analysis approaches have been proposed previously, to the best of the author's knowledge, none of them systematically investigated the way to apply personality descriptors to help automatic depression analysis system.

Since there is solid psychological evidences (Ellgring, 2007; Girard, Cohn, Mahoor, Mavadati, & Rosenwald, 2013; Stuhrmann, Suslow, & Dannlowski, 2011; Girard et al., 2014; Chentsova-Dutton, Tsai, & Gotlib, 2010; Gehricke & Shapiro, 2000; Renneberg, Heyn, Gebhard, & Bachmann, 2005; Sloan, Strauss, & Wisner, 2001) that depression status are also marked by non-verbal objective cues related to facial attributes (S. Scherer, Stratou, & Morency, 2013; Goldstein, 1964) and associated with personality, this thesis is also interested in extending the proposed approaches to the automatic depression analysis domain. In particular, it explores multiple ways to apply the personality traits descriptors to help automatic depression analysis as well as investigating how much improvement can be achieved.

1.1 Research questions and motivations

1.1.1 Long and subjective traditional clinical assessment

Standard clinical personality/depression assessment techniques can be subjective because these depend almost entirely on the professional's own understanding of the individual's verbal psychological report such as interviews and questionnaires completed by individuals or others' perception on them. Sometimes, the relevant personal information or experts may not be accessible. **Then, how to develop accurate and expert-independent automatic personality/depression analysis systems based on face videos, which should not only provide the objective predictions but also speed up the assessment process, is the first research question**.

1.1.2 Lack of proper way to extract personality-related descriptor from a frame/short-segment

Many face video-based automatic personality analysis approaches attempt to infer personality traits from very short video segments or even single frames (Bekhouche et al., 2017; Aran & Gatica-Perez, 2013b; Fang, Achard, & Dubuisson, 2016; Nguyen, Marcos-Ramiro, Marrón Romera, & Gatica-Perez, 2013), rather than long-term behaviour. The problem is that people with different personalities can display the exact same configuration of facial behaviours in a single frame or very short segment (e.g., facial display, head pose, etc.). Thus, a training strategy that learns relations between short segments and personality labels would lead to an ill-posed machine learning task, i.e., the same input pattern has multiple labels, making it practically impossible to learn a good hypothesis. Although a person's frame/segment-level facial status can change rapidly in a short period, video-level (event) labels will remain unchanged for each video, as personality doesn't change on that time scale. Therefore, a brief segment (a single frame or very short segment) of expressive behaviours may not carry sufficient information to encode a personality trait-specific pattern. **Then, the second research question is how to properly extract a single multi-dimensional personality-related descriptor from a corresponding brief segment or frame without using video-level personality label.** In this thesis, the short-term and mid-term refer to the time duration that is less than 1.5 seconds and from 1.5 seconds to 5 seconds, respectively. Meanwhile, the long-term represents the time duration that is similar to the length of an entire personality/depression video.

1.1.3 Lack of proper person-specific and multi-scale dynamic long-term representation

Personality traits can be defined as the aspects of personality that are relatively stable over time but differ across individuals (Kassin, 2003). However, to the best of the author's knowledge, there are no reports of attempts to specifically construct a computer-vision descriptor that reflects long-term person-specific behaviours. Meanwhile, as the optimal temporal scales of facial behaviours for inferring personality is still unclear, existing automatic personality analysis approaches generally fail to retain multi-scale temporal dynamics in their long-term descriptor. **To this end, the third research question is that how to properly extract a descriptor that reflect the behaviours of an entire video, which should: 1. encode person-specific spatio-temporal information; 2. encode multi-scale facial temporal dynamics from all available frames; 3. have a fixed-length representation regardless of the length of the**

video.

1.1.4 How to apply personality information to help automatic depression detection

In addition, though many psychological studies (Ellgring, 2007; Girard et al., 2013; Stuhrmann et al., 2011; Girard et al., 2014; Chentsova-Dutton et al., 2010) claim that personality traits are associated with depression status, to the best of the author's knowledge, no previous study has systematically investigated proper ways to apply personality descriptors to help automatic personality analysis systems. **Therefore, the last research question that this thesis will investigate is whether the personality information can enhance the performance of automatic depression analysis system, and if so what are suitable ways of doing so.**

1.2 Description of the work

This thesis proposes a video-level person-specific facial dynamic encoding approach and a video-level multi-scale facial dynamic encoding approach for automatic personality/depression analysis. The proposed approaches focus on addressing the main problems of existing automatic personality traits analysis systems (summarized in Sec.1.1). In addition, it also investigates proper ways to apply personality descriptor to automatic depression analysis.

1.2.1 Self-supervised learning video-level person-specific facial dynamics

The person-specific representation approach explores a novel way to encode video-level facial dynamics directly from the face images. Firstly, most existing approaches (Kollias et al., 2019a; Ringeval et al., 2019) model short-term facial dynamics at the feature level using sequential latent models such as Recurrent Neural Networks (RNN). However, learning dynamics from extracted features may result in the loss of significant dynamics in original face images. To keep such information, this thesis firstly extends the dynamic image algorithm (Bilen et al., 2017) to summarize short-term facial dynamics directly from face images. It starts with producing per-frame aligned static facial images based on the automatically detected facial landmarks, removing the background noises. Then, it extends the dynamic image algorithm (Bilen et al., 2017) to produce a dynamic facial representation based on the preceding and succeeding static facial images of the given frame. Particularly, it encodes the past and future facial temporal evolution of the given frame, i.e. how was the current facial status evolved from the past and how would the current facial status changes in the future, as well as its spatial information into a single and fixed-size 9-channel matrix. This way, the produced frame-wise representations provide spatio-temporal patterns to the back-end models rather than solely input static facial image as most methods did, which only contains spatial patterns.

As it can be observed, the sequence-based dynamic encoding approach described above achieved better performance than static face images for the task of dimensional affect estimation, which shows its strong ability in encoding facial dynamics. Thus. this thesis further proposes to generate a similar dynamic representation (DR) that summarizes motions surround the given face image, but that can be inferred from still images. It can be treated as an image-to-image translation approach where a network is trained to generate a Dynamic Representation (DR) that has the same size as the input face image. During the training, a large pool of unlabelled preceding and succeeding frames for each face image is given, from which the temporal evolution of adjacent frames can be learned in a self-supervised manner without using target representations. The network is then trained to generate a representation that, when projected onto each adjacent frame within a given time-window, is capable of sorting them in time. This way, the well-trained network can infer generic short-term facial dynamics from any previous unseen face image. After that, this generic model is frozen, and a set of intermediate person-specific adaptation layers (PALs) are incorporated into this architecture. The self-supervised learning is then resumed with only person-specific videos. As are result, the learned adaptation layers' weights will be person-specific, making them a valuable source of person-specific dynamic modelling. This approach (Song, Jaiswal, et al., 2021) then concatenates the weights of the learned adaptation layers as a video-level person-specific representation, which can be directly used to predict the personality traits without needing other parts of the network.

1.2.2 Video-level multi-scale facial dynamics modelling

Besides the person-specific video-level representation, this thesis also proposes a multi-scale video-level facial dynamic modelling approach called spectral approach. The spectral approach consists of concatenating descriptors of all frames in a video as a multi-channel time-series signal, describing the visual expressive facial behaviours. To produce a multi-scale, length-independent video-level representation, Fourier Transform is employed to encode the multi-channel time-series behaviour or latent signal of the entire video. This approach further employs two frequency alignment methods to create spectral representations of equal size and frequency coverage, regardless of variation in the length of input videos. The produced spectral representations contain long-term (video-level) facial temporal information in the frequency domain, where each frequency component stands for a unique scale of dynamics. In short, such a representation encode multi-scale facial dynamics of the video.

1.2.3 Applying person-specific and multi-scale facial behaviour representation to automatic personality traits analysis

Both person-specific representation and spectral representation can encode the facial dynamics from variable-length videos into fixed-size representations. Therefore, this thesis proposes to combine them. The combined representation contains not only video-level multi-scale facial temporal information but also video-level person-specific facial dynamics. After that, this thesis presents how to apply such representations to automatically predict personality traits. In this thesis, two datasets, e.g., VHQ dataset (Jaiswal, Song, & Valstar, 2019; Jaiswal, Valstar, Kusumam, & Greenhalgh, 2019) and ChaLearn dataset (Ponce-López et al., 2016) were used to evaluate the performance of the proposed approaches on both self-reported and apparent personality traits analysis. Several back-end models, e.g., Artificial Neural Network (ANN), Support Vector Machine Regressor (SVR), Convolutional Neural Networks (CNN), etc., are employed to generate predictions. This contribution is detailed explained in Chapter 5.

1.2.4 Personality traits-guided automatic depression analysis

As suggested by (Klein, Kotov, & Bufferd, 2011; Kendler, Gatz, Gardner, & Pedersen, 2006; Kendler et al., 2007; Hettema et al., 2006), people with certain personality traits are more likely to be affected by depression at some point in their life. Hence, personality traits can act as strong priors for predicting depression. In this sense, the final part of this thesis devotes to investigate a superior way to apply personality descriptors to automatic depression analysis. In particular, this study firstly employs and extends the proposed long-term approaches as the baselines to automatically estimate depression severity. Then, multiple strategies are proposed to apply different personality traits descriptors (including labels, predictions, latent features, etc.) as the additional information to depression analysis. This contribution is detailed explained in Chapter 7.

1.3 Contributions

1. This thesis proposes a Dynamic Facial Representation (DFR) encoding algorithms that returns a length-independent representation from an image sequence, which integrates the facial temporal and spatial information, summarising the variation over time. This representation dismisses non-face related attributes.

2. This thesis proposes a novel Rank Loss to train an image-to-image translation network in a self-supervised manner, tasked with inferring a Dynamic Representation (DR) from a given face image. The DR is formulated as a kernel that, when projected onto the adjacent frames, can sort them in time. In particular, the proposed rank loss enforces the generated DR to rank both preceding and proceeding frames according to their relative temporal distance to the input frame. This way, the network not only learns to map an image to a corresponding DR, but also contributes

to define it. Importantly, this approach allows the network learning generic facial dynamics from unlabelled videos.

3. Based on the proposed Rank Loss, this thesis also proposes to train a set of person-specific adaptation layers (PALs)' individually, and proposes to concatenate the learned weights of the PALs as the video-level person-specific facial dynamic representation. This representation can encode person-specific facial dynamics that are stable over time for the given person but differ from other individuals, and is shown to be useful for the automatic personality analysis tasks.

4. This thesis proposes a novel spectral approach that can convert long and variable length time-series data to short and fixed-size spectral representations. The produced spectral representations can encode multi-scale temporal dynamics of the given time-series data, and can be easily processed by standard Machine Learning models. This approach is shown to be useful for automatic depression analysis. Specifically, This approach can take any time-series such as behaviour primitives (e.g. AU, gazes, etc) as the input, and consequently the advance in developing front-end detector (e.g. AU, gazes, etc) could also benefit the task performance of this spectral approach.

5. This thesis proposes and evaluates multiple ways to combine the personality descriptors with visual behavioral descriptors for automatic depression severity estimation, showing that additional personality information can improve the depression severity estimation performance across two proposed video-level representation approaches.

6. This thesis investigates the influence of multiple factors on the performance of automatic personality and depression analysis, by evaluating interview tasks, types of behaviour primitives, the length/task contents of videos, fusion strategies, etc., and found that AUs have different impacts on personality and depression analysis performance and differences in task contents and fusion strategies will also resulted in different analysis result.

1.4 Publications

This thesis has led to a number of peer-reviewed published papers:

1. **S. Song**, L. Shen and M. Valstar, "Human Behaviour-Based Automatic Depression Analysis Using Hand-Crafted Statistics and Deep Learned Spectral Features," 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), Xi'an, 2018, pp. 158-165.

2. **S. Song**, S. Zhang, B. W. Schuller, L. Shen and M. Valstar, "Noise Invariant Frame Selection: A Simple Method to Address the Background Noise Problem for Text-independent Speaker Verification," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1-8.

3. S. Jaiswal, **S. Song** and M. Valstar, "Automatic prediction of Depression and Anxiety from behaviour and personality attributes," 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, United Kingdom, 2019, pp. 1-7.

4. **S. Song**, E. Sánchez-Lozano, M. K. Tellamekala, L. Shen, A. Johnston and M. Valstar, "Dynamic Facial Models for Video-Based Dimensional Affect Estimation," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 1608-1617.

5. F. Ringeval, B.W. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.M. Messner, **S. Song** , et al. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. InProceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop 2019 Oct pp. 3-12.

6. **S. Song**, S. Jaiswal, L. Shen and M. Valstar, "Spectral Representation of Behaviour Primitives for Depression Analysis," in IEEE Transactions on Affective Computing. (2020)

7. Egede, J. O., **S. Song**, T. A. Olugbade, C. Wang, A. Williams, H. Meng, M. Aung, N. D. Lane, M. Valstar, and N. Bianchi-Berthouze. ”EMOPAIN Challenge 2020: Multimodal Pain Evaluation from Facial and Bodily Expressions.” In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG), pp. 499-506.

8. **S. Song**, E. Sánchez, L. Shen, and M. Valstar, “Self-supervised Learning of Dynamic Representations for Static Images,” In 2021 25th IEEE International Conference on Pattern Recognition (ICPR 2021)

9. R. Liu, B. Reimer, **S. Song**, B. Mehler, and E. Solovey. ”Unsupervised fNIRS feature extraction with CAE and ESN autoencoder for driver cognitive load classification.” in Journal of Neural Engineering 18, no. 3 (2021). p.036002.

10. **S. Song**, S. Jaiswal, E. Sánchez-Lozano, G. Tzimiropoulos, L. Shen, and M. Valstar, ”Self-supervised Learning of Person-specific Facial Dynamics for Automatic Personality Recognition,” in IEEE Transactions on Affective Computing. (2021)

1.5 Thesis structure

The overall structure of this thesis is comprised of seven chapters. The rest of this thesis is organized as follows.

1. Chapter 2 reviews the psychological backgrounds of personality and depression, especially the relationship between personality, depression and facial behaviours. This would provide a theoretical basis for the researches conducted in this thesis, i.e., it is feasible to infer personality and depression from non-verbal facial behaviours.

2. Chapter 3 reviews the existing automatic personality analysis and automatic depression analysis approaches, including the typical methods of pre-processing, feature extraction, information fusion and classification/regression models that have been used for such tasks. This chapter also specifically summaries the temporal models used in the reviewed approaches.

3. Chapter 4 proposes a person-specific adaptation approach. This chapter starts with extending the dynamic image algorithm to model short-term facial dynamics, providing the theoretical basis for the proposed novel self-supervised learning algorithm, which can learn short-term generic facial dynamics. Both approaches encode bidirectional facial spatio-temporal patterns of the input image in the context of the face, providing richer and more discriminate facial patterns compared to the original static face images. Based on such self-supervised learning algorithm, person-specific adaptation approach is also proposed, which encodes a person-specific video-level facial dynamic representation for each individual.

4. Chapter 5 proposes a multi-scale video-level facial dynamics encoding approach: spectral analysis approach. This approach encodes the video-level representation from multi-channel time-series signals that produced by features of all frames in each video. This approach can generate fixed-size representation regardless of the length of videos, and encode multi-scale long-term facial behaviour dynamics.

5. Chapter 6 firstly explains all experimental settings including datasets, training details, evaluation metrics, etc. Then, it specifically evaluates the influence of different settings on the proposed person-specific adaptation approach and spectral analysis approach on the task of self-reported and apparent personality traits intensity estimation as well as compares them to the state-of-the-art approaches. In addition, this chapter also conducts a series of ablation studies on frame ranking and dimensional affect estimation tasks to evaluate the proposed self-supervised learning algorithms in modelling short-term facial dynamics.

6. Chapter 7 firstly introduces a novel audio-visual VHQ datasets that annotated both personality and depression labels for each participants. This chapter applies the proposed automatic personality analysis approaches as the baselines to automatic depression severity estimation task. In addition, it explores multiple ways of combining the personality descriptors with facial behaviour descriptors to benefit the automatic depression analysis performance.

7. The last chapter devotes to the conclusions and future work, including a briefly review of the proposed approaches, a summary of the problems that have been addressed in this thesis, and some conclusions of the experimental results on automatic personality and depression analysis. This chapter also discusses the limitations of this thesis and the potential future works in detail.

Chapter 2

Background

Human personality is defined as the characteristic set of behaviours, cognition, and emotional patterns that evolve from biological and environmental factors (Hogan et al., 1997), which can be reflected by the coherent patterning of affect and behaviours over time and space (Revelle & Scherer, 2009). Recognizing personality can help a better understanding of human behaviour (Roberts & Jackson, 2008), emotional processes (Komulainen et al., 2014) and health conditions (Huang et al., 2017; Jaiswal, Song, & Valstar, 2019). Traditionally, personality analysis can be categorized into two types: 1. self-reported personality, which is reflected by a person's observable behaviours; 2. apparent (perceived) personality, which is defined as observers' perception about the person through various cues.

Recently, the trait-based models such as Three Factor Model (Eysenck & Eysenck, 1965), Five Factor Model (McCrae & Costa, 1987), etc., have become the most common way to measure complex and implicit human personality. They primarily focus on evaluating the aspects of personality that are relatively stable over time but differ across individuals (Kassin, 2003). While a standard approach of evaluating personality traits is to use questionnaires based on verbal behaviour descriptors, such as the Big Five Inventory (Cavallera et al., 2013), previous psychological studies frequently (Qin et al., 2018; DePaulo, 1992; Borkenau & Liebler, 1992) suggested that non-verbal behaviours also contain vital cues of a human's implicit dispositions and internal states. Since the human face contains rich details regarding personal identity, such as gender, ethnicity, emotional state, age, etc., facial appearance and behaviours can be used as a primary and salient non-verbal source of information to infer various personalized attributes (Joo, Steen, & Zhu, 2015).

The purpose of this chapter is reviewing the related psychological studies that provide a theoretical basis to support facial behaviour-based automatic personality/depression analysis. It firstly reviews the personality trait theory in Sec.2.1, and the impact of personality in Sec.2.2, emphasizing the importance of the study conducted in this thesis. The facial expression theory is briefly reviewed in Sec.2.3. Then, the relationship between facial expressions of emotions and personality are also explored in Sec.2.4, from the perspective of psychology, which provides the psychological evidence that personality is marked by human facial displays and behaviours. Since this thesis is also interested in applying personality information to help automatic depression analysis, the Sec.2.5 also reviews the psychological background of the depression. The relationship between depression and personality traits/facial behaviours are reviewed in Sec.2.6 and Sec.2.7, respectively, showing that depression status is associated with facial behaviours and personality traits.

2.1 Personality traits models

2.1.1 Three factor model

Three Factor Model, also called PEN model, was proposed by Eysenck et al. (Eysenck & Eysenck, 1965). This model is based on the assumption that both internal factors (e.g., biological factors) and external factors (e.g., environmental factors) would impact on a individual's personality traits. The PEN model comprises three traits dimensions: extraversion-introversion, neuroticism-emotional stability and psychoticism-normality. Specifically, this theory (Eysenck & Eysenck, 1965) claimed that extraversion-introversion trait correlates to the levels of brain activity, or cortical arousal. For example, extrovert individuals generally have lower levels of cortical arousal than introvert people, and thus they are tended to seeking arousal from external stimuli. Meanwhile, higher internal arousal levels lead introvert people avoiding experiencing more arousal from outside. These can be reflected by that extrovert people are more talkative and outgoing than introvert individuals. For the neuroticism-emotional stability dimension, individuals who have high neuroticism scores are more likely to experience stress and anxiety but their emotions are more stable in comparison to emotional people. Finally, high psychoticism score usually achieved by individuals that are creative (Eysenck & Furnham, 1993). However, this types of people are more likely to be involved in irresponsible and miscalculated events compared to normal people. The detailed description of PEN model is shown in Table 2.1.1. To measure these dimensions, a verbal-based Eysenck Personality Questionnaire (EPQ) was proposed in (Eysenck & Eysenck, 1965).

Table 2.1: Three Factor Model (PEN Model)

Trait Name	High	Low
Psychoticism-Normality	1. Engaging in irresponsible or miscalculated behaviours 2. Contravene accepted social norms 3. Regardless of consequences 4. Creative, aggressive, cold and impulsive	1. Responsible behaviours 2. Less likely to commit crimes
Extraversion-Introversion	1. Engaging more in social activities 2. Enjoying being the focus of attention 3. Accumulating a larger social network of friends and associates 4. Talkative and outgoing	1. Shying away from large social gatherings 2. Feeling uncomfortable when engaging with strangers 3. Enjoying contemplative exercises 4. Individual's psychic energy
Neuroticism-Emotional Stability	1. Experiencing higher levels of stress and anxiety 2. Worrying about relatively insignificant matters 3. Focusing on negative aspects of a situation 4. Feeling jealous of others 5. Feeling dissatisfied, angry or frustrated with others	1. Experiencing more emotional stability 2. Able to cope with stressful events and set 3. Less stringent demands of themselves 4. More tolerant of the failings of others and remain more calm in demanding situations

2.1.2 Five factor model

Five Factor Model (FFM) (McCrae & Costa, 1987), also known as the 'Big Five', or OCEAN model, is a typical lexical approach for personality assessment. In recent years, it has been frequently employed in various studies and applications. The FFM model is made up of five relatively independent personality dimensions: Openness to experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism.

Specifically, Extraversion is frequently associated with gregarious, assertive, talkative, active and sociable personality (Barrick & Mount, 1991). This dimension is also claimed to have two sub-dimensions: ambition and sociability (Hogan, 1982). Neuroticism (also called Emotional stability) relates to the level of anxiety, anger, worry, embarrassment, insecurity and emotion. The third dimension has generally been interpreted as the Conscientiousness, which is related to conformity, will to achieve (Digman, 1989; Smith, 1967), dependability (organized, planful and responsible) (Noller, Law, & Comrey, 1987; Fiske, 1949; Botwin & Buss, 1989), and volition (hardworking, achievement-oriented, persevering) (Smith, 1967; Lei & Skinner, 1982; Digman & Takemoto-Chock, 1981; Bernstein, Garbin, & McClellan, 1983). Agreeableness, also called likability (Noller et al., 1987; Norman, 1963), is found to be associated with being flexible, forgiving, tolerant, courteous, good-natured and soft-hearted. Finally, Openness reflects a person's intellectual curiosity, creativity and a preference for novelty and variety. Individuals who score high in openness are more likely to lose focus and engage in risky behaviour (Ambridge, 2014). The robustness of this model has been empirically validated under various situations (Conley, 1985; Costa Jr & McCrae, 1988; McCrae & Costa, 1989; Noller et al., 1987; Bond, Nakazato, & Shiraishi, 1975; Digman, 1989; Digman & Inouye, 1986; Digman, 1990). The detailed description of FFM model is shown in Table 2.1.1.

Table 2.2: Five Factor Model (OCEAN Model)

Trait Name	Behaviours	Facial expressions of emotion
Neuroticism	1. Emotional 2. Worrying about things 3. Easy to be disturbed 4. Often feeling blue	1. Increased facial displays of negative emotion 2. Embarrassing face with various emotions 3. Negatively correlated with Duchenne smiles
Conscientious	1. Organized and dependable 2. Self-discipline and acting dutifully 3. Perceived as being stubborn and focused	1. Controlled smile 2. Gaze aversion and head movements down 3. Facial muscle action that inhibited the smile
Extraversion	1. Energetic and sociable 2. Talkative and more dominant 3. Perceived as attention-seeking and domineering	1. Facial displays that encourage positive social contacts
Agreeableness	1. Compassionate and cooperative to others 2. Naive or submissive 3. Be more naturally altruistic 4. Feeling concern for their community	1. Frequently displaying duchenne laughter or sympathy 2. Negatively correlated with facial displays of anger and disgust, displayed by the oblique eyebrows of sadness and a head movement forward
Openness	1. Unpredictability and lack of focus 2. Seeking new experiences and adventures 3. Preferring variety and diversity 4. Creative and open-minded	1. Keeping the face open and exposed 2. Showing curious facial displays

2.2 The impacts of personality

Personality traits model is a bio-psycho-social system that reflects individual differences (Roberts & Jackson, 2008). The main motivation of developing personality analysis systems is that personality traits have great impacts on many aspects of people's behaviours such as emotion, health condition, social behaviour, cognition process, etc. Thus, an accurate personality traits recognition system could help a better understanding of human behaviours.

Firstly, personality traits are related to individuals' social behaviours. According to (Barnett et al., 2015), conscientiousness people usually prefer to learn and use new technologies, while neurotic people tend to refuse to perceive and use them. In addition, individuals who have high extraversion scores showed a better ability to adapt to new technologies. As claimed by (Lane & Manner, 2011), different personalities would result in people having different ownership and use of smartphones. As explored in (Tsao & Chang, 2010) and (Heinström, 2000), the variability in personality also leads the different shopping behaviours. For example, people who have high scores in neuroticism, agreeableness or openness prefer shopping online, while people who have high degrees of neuroticism, extraversion, and openness are more likely to shop for fun (hedonic purchase motivation). Besides, personality difference is found to be strongly correlated to an individual's working behaviours (Zimmerman, 2008; Jenkins, 1993). The emotional stability of employers determines their intentions to give up the task or even quit the job. In contrast, conscientious and agreeable people are less likely to turn over their decisions, which means they tend to finish what they have decided. Besides, personality also relates to human behaviours such as teenagers' illegal drug usage (Brook et al., 2001), information-seeking behaviour (Al-Samarraie, Eldenfria, & Dawoud, 2017), politics decisions (Greenstein, 1967), etc.

Apart from personal behaviours, many studies suggested that personality traits have significant impacts on human cognition and emotional processes (Komulainen et al., 2014). As found by Revelle et al. (Revelle & Scherer, 2009), individuals that have certain personality traits are more likely to experience the corresponding emotions than others. This research claimed that trait differences in emotionality would increase the odds of experiencing trait-congruent emotions. For example, trait surprise may result in a higher chance of experiencing anxiety while trait pleasure may lead to reduced despair. Revelle et al. (Revelle & Scherer, 2009) concluded that neurotic people usually have larger affect variability. They are also more likely to have higher negative and lower positive affect status than others, resulting in a less positive attitude to daily incidents. In contrast, conscientiousness and agreeableness people tend to have the opposite emotional processes. These people would generally have lower variability of sadness, and more likely to have a positive feeling of daily incidents. Similarly, individuals who score high in extraversion would normally have higher positive affect and more positive subjective evaluations of daily events; Although there is no significant relationship that has been found between openness trait and human affect, it predicted higher reactivity to daily stressors.

Finally, personality traits are strongly associated with health conditions. A large number of previous studies (Lo et al., 2017; Kendler et al., 2006, 2007; Hettema et al., 2006; DENNIS et al., 1981; Jaiswal, Song, & Valstar, 2019; Matsudaira & Kitamura, 2006) found that personality traits are correlated to the mental sickness, e.g. depression (please see Sec.2.7 for details). Meanwhile, as claimed by (Huang et al., 2017), personal characteristics, such as sense of coherence, neuroticism and optimism, affect people's psychological health-related quality of life (HRQOL). Specifically, higher scores in extraversion, agreeableness, openness or conscientiousness usually predict better HRQOL conditions for corresponding individuals. On the other hand, neurotic people and people who have negative affectivity personalities are more likely to experience poorer HRQOL conditions.

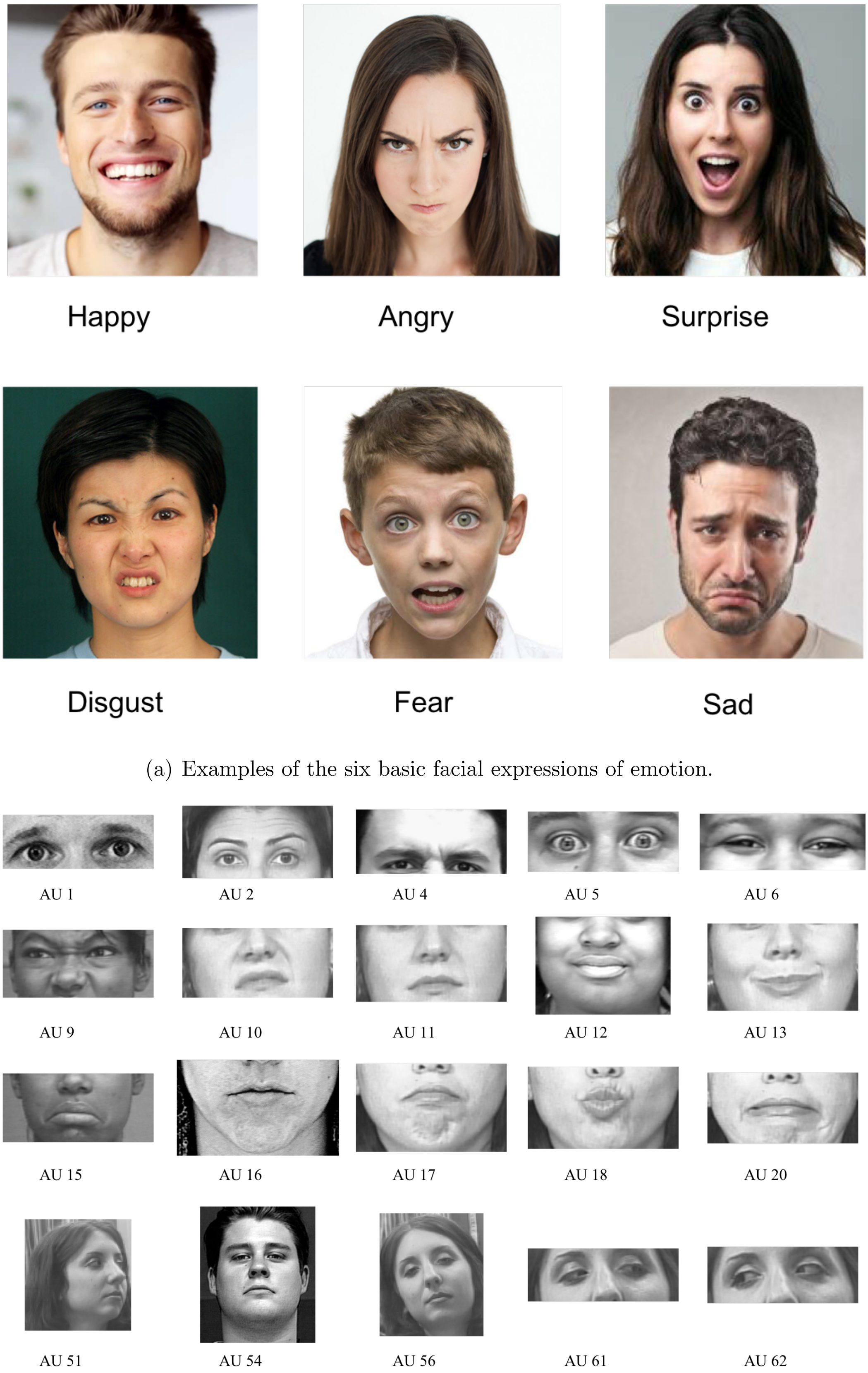
In summary, the aforementioned studies show that personality is associated with individual's health conditions, including mental health such as depression. Thus, it is interesting to explore a reliable way to apply the personality descriptors for automatic depression analysis, which is also a goal of this thesis (Please see Chapter.7 for details).

2.3 Facial expression analysis theory

Facial expressions (facial actions) can be defined as the facial changes in response to a person's internal states including emotion, intentions, etc. (Tian, Kanade, Cohn, Li, & Jain, 2005). They are represented as one or more motions or positions of the muscles beneath the skin of the face.

The research of facial expression analysis started with the work of Darwin in 1872 (D. et al., 1981; K. Scherer & Ekman, 1982; P. Ekman, 1989; Friesen & Ekman, 1978), and such study was sometimes confused with emotion analysis in the computer vision domain. However, besides emotion, facial expressions can also convey other information such as physical efforts, ideas and intentions (Carroll & Russell, 1996; Russell, 1991a, 1991b). Meanwhile, understanding emotions needs higher level knowledge than facial expressions. It can be considered that emotion is to facial expressions as the climate is to weather. Facial expression analysis can be categorized as two main theories. Darwin proposed an emotion-based facial expression theory (D. et al., 1981), which classified facial expressions into a set of prototypic emotional expressions, including Happy, Angry, Surprise, Fear, Disgust and Sad (Fig.2.1(a)). This theory has been further revised by Ekman and Friesen (Ekman & Paul, 1993; P. Ekman, 1976) and Izard et al. (Izard, Dougherty, Hembree, & Izard, 1983).

However, these prototypic expressions have imbalanced chances of occurrence in everyone's daily life, and they may not be able to represent facial status under some conditions. Motivated by this, another facial expression theory called Facial Action Coding System (FACS) was proposed by Ekman (Ekman & Paul, 1993). It is a human-observer-based system that aims to capture subtle changes in one or a few discrete facial features, e.g., eyebrows or eyelids which is the typical of paralinguistic displays. The FACS applies 44 action units (AUs) (Fig.2.1(b)) to represent the facial changes, where 30 of them are anatomically related to the contraction of a specific set of facial muscles (P. Ekman, 1989). The 14 reminder AUs are referred as miscellaneous facial actions. Specifically, in FACS theory, the occurrence of AUs can vary in intensity, i.e., the degree of muscle contraction can be represented by 3- or 5-point ordinal intensity scales.



(a) Examples of the six basic facial expressions of emotion.

(b) Examples of some facial action units (AUs)

Figure 2.1: Visualization of two facial expression theory.

While some existing facial expression analysis studies (Essa & Pentland, 2002; Lien, Kanade, Cohn, Li, & Zlochower, 1998; Kimura & Yachida, 1997; Jaiswal & Valstar, 2016) indicates the intensity variation within AUs can be automatically and accurately estimated under some specific conditions (Please refer to (Tian et al., 2005) for details), the individual differences in appearance and expressiveness bring challenges to such tasks. In particular, the face shape, skin color, gender, age and ethnic background (Farkas & Munro, 1987; Zlochower, Cohn, Lien, & Kanade, 1998) make individuals have completely different faces, resulting in similar facial expressions to have different appearances. Meanwhile, the difference in expressiveness would lead to different degrees of facial plasticity, morphology, and frequency of intense expression (MANSTEAD, 1991). For example, Zlochower et al.(Zlochower et al., 1998) found that the performance of optical flow-based algorithms designed for adolescents degraded significantly for infants, which may due to the lack of transient furrows, the reduced texture of infants' skin, and juvenile facial conformation. Thus, when developing and validating facial expression analysis algorithms, it usually requires a large number of examples across various hairstyles, gender, ethnic backgrounds, ages, etc.

2.4 Relationship between personality, emotions and facial expressions

There is convergent evidence showing that personality can be defined by specific tendencies to experience and express certain emotions. While John et al. (John, 1990) suggested that verbal language is frequently associated with personality traits and emotions, e.g., many words are related to specific emotion tendencies, previous studies also claimed that non-verbal expressive behaviours are informative to emotions and personality. Specifically, Keltner et al. (Keltner, 1996) found that both emotion and personality traits can be expressed in distinct, observable behaviours that evoke responses in others. In addition, the studies conducted by Ekman et al.(P. Ekman, 1992b) and Izard et al. (Izard, 1977) showed that human nature, including personality and emotions of individuals, can be interpreted by a certain number of universal non-verbal behaviours characteristics. Specifically, DePaulo (DePaulo, 1992) concluded that personality traits are likely to be strongly influenced by an individual's implicit predispositions and internal states to display certain emotions. Thus, a person who tends to express certain emotions is more likely to consistently evoke certain emotions and actions in others, contributing to the others' perception (apparent personality) on this individual.

According to (Calder, Ewbank, & Passamonti, 2011), some brain areas involved in facial displays recognition are also biologically involved in processing emotions. While emotions provide vital non-verbal cues to reflect personality traits, facial expressions are quick, reliable, and seemingly universal signals to represent emotions. In other words, emotions can be expressed by a set of consecutive facial expressions. As a result, facial expressions/displays are an important non-verbal source to infer personality traits (Knapp et al., 2013), and thus a large number of personality studies (Borkenau et al., 2009; DeBruine et al., 2006; Shevlin et al., 2003; Zebrowitz et al., 2002) build on facial behaviours.

Many of these studies (J. A. Hall et al., 2008; Kenny, 1994; Knapp et al., 2013) showed that when targets were photographed with spontaneous facial expressions (under spontaneous conditions), the personality traits values predicted from observers were very similar to the ground-truth for almost all the traits. Particularly, they claimed that consistently displayed facial expressions could define how the individual is perceived by others as well as the character of the individual's interactions and social relations. Consequently, facial expressions are always employed as key clues in observers' inferences about the individual's personality traits. For example, a study discussed in (Knutson, 1996) shows that happy expressions are associated with high dominance and affiliation; facial expressions of anger and disgust refer to high dominance and low affiliation and from fearful; sad facial expressions reflect low dominance. These findings suggest that facial expressions can not only convey the internal state of individuals, but also form interpersonal perceptions.

Meanwhile, since the big five personality traits model is prevalent, some studies have explicitly investigated the relationship between these traits and facial expressions. Edgar et al. (Edgar, McRorie, & Sneddon, 2012) found that the extraversion trait is consistently correlated with the sensitivity to positive emotional stimuli. This trait is reflected by facial expressions of positive emotions and negative emotions that can encourage positive social contacts (Keltner, 1996). In contrast, neuroticism is found to be negatively linked to positive emotions. This can be reflected by behaviours of: 1. increased facial expressions of negative emotions such as anger, contempt and fear; 2. embarrassing face with various emotions, including sympathy, amusement, embarrassment, and distress; 3. less likely to display Duchenne smiles. Agreeableness is claimed to be positively associated with facial expressions of emotion such as Duchenne laughter and displays of sympathy, which would benefit cooperative and friendly social interactions. Meanwhile, this dimension is negatively related to some negative emotions, including anger and disgust, reflected by the facial actions such as oblique eyebrows of sadness and forward head movements (Eisenberg et al., 1989). According to (Keltner, 1995), individuals who have high scores on conscientiousness are more likely to display facial expressions showing embarrassments, such as controlled smile, gaze aversion, head movements down and face touching, and sometimes they laugh for reasons other than pleasure. In addition, a fine-grained research (Watson & Clark, 1992) showed that this trait is strongly related to facial muscle actions that inhibited the smile, consistent with the role of impulse control in conscientiousness. Finally, individuals who score high in openness are more likely to keep their face and body open, and display curious facial expressions.

2.5 Depression and its severity measurement

Major Depression Disorder (MDD) is a psychiatric disorder defined as a low mood state with a problematic severity of duration for at least two weeks. This mental illness is caused by many aspects such as biological, psychological and genetic factors. It can negatively impact one's day to day life, causing people to become reluctant or unable to perform activities, which can negatively affect a person's personal, work, school life, as well as sleeping, eating habits, general health, etc., and affects thoughts, behaviour, feelings, and sense of well-being (Edition, Association, et al., 1994). In extreme cases it can lead to suicide, which is the leading cause of death for men under 50 in the UK. Depression is currently the most prevalent mental health disorder and the leading cause of disability in developed countries.

A correct diagnosis of depression can provide vital information about how to reduce inappropriate feelings of blame, shame, loneliness and low self-esteem for the corresponding patients and also facilitates the communication between (potential) patients and health professionals about the support and services they need (Craddock & Mynors-Wallis, 2014). It is the key to choose which interventions are suitable for treating a patient. Standard clinical depression assessments usually employ verbal psychological questionnaires completed by patients or caregivers to make judgements (Cohn et al., 2009). Among them, Patient Health Questionnaire (PHQ-9), Beck Depression Inventory II (BDI-II) and Hamilton Depression Scale (HAMD-17) are the most popular ones in recent years.

PHQ-9 is a self-administered instrument based on the nine DSM-V criteria for Major Depressive Disorder. It consists of 9 questions, where each of them has four ratings ranging from 0 to 3. Based on the total scores, four depression status can be predicted:

- 0-9: minimal depressive symptoms
- 10-14: mild depressive symptoms
- 15-19: moderate depressive symptoms
- 20-27: severe depressive symptoms

This inventory is usually employed to guide criteria-based diagnosis of depressive symptoms to assist in identifying treatment goals, determining severity of symptoms, as well as guide clinical intervention.

BDI-II is the revised version of the Beck Depression Inventory (BDI), and is also designed to assess depressive symptoms based on DSM-V criteria. This inventory is usually used for adults and adolescents over 13 years old, allowing them to self-report on symptoms from the last 2 weeks. It consists of 21 items. Similar to PHQ-9, each item of this inventory has four ratings with associated values ranging from 0 to 3. Based on the total scores, four depression statuses can be predicted:

- 0-13: minimal depressive symptoms
- 14-19: mild depressive symptoms
- 20-28: moderate depressive symptoms
- 29-63: severe depressive symptoms

The BDI-II has been widely used in clinical or hospital settings as well as community mental health settings. From the perspective of affective computing, the PHQ-9 and BDI-II questionnaire have been widely employed as the label to the automatic depression analysis dataset (Gratch et al., 2014) and AVEC challenges (M. Valstar et al., 2013, 2014, 2016; Ringeval et al., 2017, 2019) since 2013, which are used for evaluating the proposed approaches in this thesis.

HAMD-17 is the Hamilton Depression 17-item questionnaire which was developed in 1957 before the publication of the DSM-III and thus did not evaluate such criteria for depression. The HAMD-17 has been used extensively within the medical community. It is a standard clinical instrument that has been proven useful for determining a patient's level of depression before, during, and after the treatment. Unlike the PHQ-9 and BDI-II, this is an observer rating, and the questionnaire is not typically used by psychologists or used as a self-report instrument but should be administered by a clinician for psychiatric patients. It can be used when there are concerns about the patient's self-report accuracy as its scores correlate well with BDI-II scores. In particular, HAMD consists of 17 items with a rating of either 0 to 4 or 0 to 2, with total scores ranging from 0 to 54. Based on the total scores, four depression status can be predicted:

- 0-7: No depression
- 8-16: Mild depression
- 17-23: Moderate depression
- 24-54: Severe depression

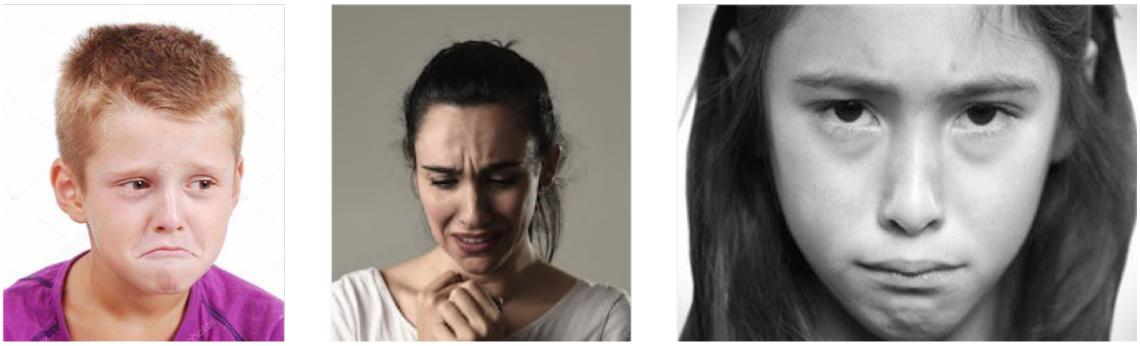


Figure 2.2: Examples of some prototypical facial displays that reflect depressive feelings

In the clinical usage, a decrease of 50% or more in the HAM-D score often indicates the positive treatment response, whereas a score of 7 or less is considered equivalent to a remission.

2.6 Relationship between depression, emotions and facial expressions

Besides the personality traits, this thesis also attempts to extend the proposed approaches to the automatic depression analysis, i.e., infer depression status from non-verbal facial behaviours. Thus, this section reviews the psychological evidence that depression can also be reflected by non-verbal facial behaviours.

As many researchers suggested, there is a strong relationship between non-verbal facial expressions of emotions and depression. An early study (Goldstein, 1964) found that some patients with depression frequently experienced ‘over arousal’. After that, Clark et al. (Clark & Watson, 1991) interpreted depression in terms of valence, which captured the positivity or negativity of an emotional state. The results showed that depressed people present deficient positivity and excessive negativity. In terms of the facial displays, a key finding that has been frequently validated is that depression is usually accompanied by reduced positive facial displays. This hypothesis has been frequently mentioned and validated across various studies (Chentsova-Dutton et al., 2010; Gehricke & Shapiro, 2000; Renneberg et al., 2005; Rottenberg, Kasch, Gross, & Gotlib, 2002; Sloan et al., 2001; Tsai, Pole, Levenson, & Muñoz, 2003; Tsai et al., 2003), and indeed has been replicated in this thesis. Additionally, depression can lead to a general reduction in facial expressiveness (Gaebel & Wölwer, 2004; Renneberg et al., 2005) and head movements (Fisch, Frey, & Hirsbrunner, 1983; Joshi, Goecke, Parker, & Breakspear, 2013). In particular, Ellgring et al. (Ellgring, 2007) summarized typical symptoms of depression in terms of facial expressions, indicating that depression is not only associated with sad facial displays but also with *"a total lack of facial expressions corresponding to the lack of affective experience"*. Regarding the negative facial expressions, researchers have conflicting conclusions. While most studies (Sloan, Strauss, Quirk, & Sajatovic, 1997; Reed, Sayette, & Cohn, 2007; Brozgold et al., 1998) argued that depression is marked by the increased negative expressions, some research reported that depressed individuals are more likely to experience reduced negative expressions (Gaebel & Wölwer, 2004; Renneberg et al., 2005), and that expressive behaviour is thus more bland and neutral. Some prototypical depression-related facial expressions are shown in Fig.2.2.

Since facial expressions of emotions can reflect the depression, many researchers attempted to directly apply non-verbal facial cues to infer individual's depression status. Cohn et al. (Cohn et al., 2009) firstly explored the feasibility of applying ML models to process audio and mid-level non-verbal facial cues for depression classification. This study extracted three sets of descriptors based on three models, i.e. manually annotated Facial Action Units (AUs), Active Appearance Model (AAM) and vocal prosody descriptors. Then, Support Vector Machines (SVM) is employed as the classifier to predict depression occurrence. The results showed that all of them were informative to depression, with facial AUs achieving the best accuracy of 88%, which indicates that facial behaviour primitives are informative to automatic depression analysis. Girard et al. (Girard et al., 2014) also specifically investigated the relationship between depression and non-verbal facial behaviours, e.g., AUs and head poses, using both manual and automatic systems. Both results showed that participants with high depression severity presented fewer affiliative facial expressions (AUs 12 and 15), more non-affiliative facial expressions (AU 14) and diminished head motion. Besides, a large number of machine learning-based automatic depression analysis systems (L. Yang et al., 2018; Song et al., 2018, 2020; Al-gawwam

& Benaissa, 2018; Meng et al., 2013; Pampouchidou et al., 2016) were building on the non-verbal facial cues, and most of them achieved relatively accurate occurrence detection or severity estimation performance (Please see Sec.3.2 for details).

2.7 Relationship between personality and depression

In the past decades, biologists researchers have found that personality is genetically associated with depression. For example, some studies (Vukasović & Bratko, 2015; Lo et al., 2017) claimed that some genes and environmental factors which determine individuals' personality are also highly correlated with mental health conditions, such as depression. Meanwhile, other studies (Kendler et al., 2006, 2007; Hettema et al., 2006) claimed that high neuroticism and depression are correlated in the genetic factors, indicating that neuroticism can reflect the genetic susceptibility to depression.

In addition to the biological evidence, psychological studies also frequently claimed a strong correlation between personality and depression. Dennis et al. (DENNIS et al., 1981) found that personality disorder is associated with depression. In particular, this study summarized several conclusions: 1. personality disorder is more common in unipolar non-melancholic depressed patients than in unipolar melancholic or bipolar depressed patients; 2. obsessive-related traits have been found to predominate the unipolar melancholic patients; 3. histrionic, hostile, and borderline traits were frequently experienced by the nonmelancholic patients. Matsudaira et al. (Matsudaira & Kitamura, 2006) investigated the impacts of personality variability on depression and anxiety. This study employed the Temperament and Character Inventory (TCI) as the personality measurement and the Hospital Anxiety and Depression (HAD) scale as the depression and anxiety measurement. The experimental results showed that people who have lower scores on some personality dimensions, e.g., Self-Directedness, Reward-Dependence, Cooperativeness, Persistence, and Self-Transcendence, are most likely to feel depressed while the immaturity of all character dimensions denoting the higher risk for specific depression. De Moor et al.(De Moor et al., 2006) examined the relationship between regular exercise and depression and personality, showing that people with lower neuroticism and higher extraversion are more likely to do regular exercise, and those people tend to have lower depression risk.

Since both personality and depression are long-term status, some longitudinal studies were specifically conducted, aiming at exploring the relationship between big five personality traits and depression. These works consistently found that both the onset and the chronicity of depression are associated with neuroticism personality (Ormel, Oldehinkel, & Vollebergh, 2004; Kendler, Kuhn, & Prescott, 2004; Krueger, Caspi, Moffitt, Silva, & McGee, 1996), while people who have a low level of extraversion are more likely to experience the onset of depression (Krueger et al., 1996; Hirschfeld et al., 1989). Klein et al.(Klein et al., 2011) surveyed the relationship between personality and depression. This study summarized that most personality traits, e.g. neuroticism, extraversion and conscientiousness, are related to depression. In particular, neuroticism is associated with the onset stage of depressive disorders. Similar conclusions were also made by (Chioqueta & Stiles, 2005), claiming that neuroticism trait and openness trait are positively correlated to feelings of depression while people who have high scores on extraversion are less likely to experience depression.

2.8 Summary

In summary, previous biological and psychological studies have been frequently concluded that personality traits and depression can be reflected by human facial behaviours, which provides the theoretical basis of the thesis, i.e., automatic analysis of facial behaviours for personality and depression recognition. Additionally, individuals with high scores in some personality traits (e.g., Neuroticism) are more likely to experience depression, indicating that personality can be a potential predictor for depression. In this thesis, the widely-used Five-Factor Model is employed as the measurement for personality traits and PHQ-9 inventory is employed as the severity measurement for depression.

Chapter 3

Related Work

Since previous psychological studies have been frequently claimed that personality traits and depression status are marked by non-verbal expressive facial cues (summarized in Chapter. 2), a large part of recent automatic personality and depression analysis systems were building upon them. This chapter systematically reviews studies that applied machine learning techniques to analyze human non-verbal expressive cues, especially facial behaviours, for automatic personality and depression recognition. In particular, the automatic personality and depression analysis literature are reviewed in Sec.3.1 and Sec.3.2, respectively.

In particular, most video-based approaches were built on the assumption that the provided video data contains human facial behaviours that are well associated with their personality or depression. While existing personality and depression datasets have utilised various stimulated contents/environments to trigger participants behaviours, including: 1. asking participants conducting winter survival task, i.e., ranking a list of 12 items in order to survive an airplane crash in winter (Sanchez-Cortes, Aran, Mast, & Gatica-Perez, 2011) ; 2. watching emotional short videos and movie segments with strongly affective multimedia contents (Correa, Abadi, Sebe, & Patras, 2018); 3. conducting a set of pre-defined tasks such as Lego building (Palmero et al., 2020), read paragraphs (M. Valstar et al., 2013, 2014) and describe figures (Palmero et al., 2020; M. Valstar et al., 2013); 4. interacting with either a human with pre-defined questions (Celiktutan, Skordos, & Gunes, 2017; Jaiswal, Song, & Valstar, 2019) or self-decided topics (Cafaro et al., 2017; Gratch et al., 2014); 5. interacting with robots that show different non-verbal behaviours (Celiktutan et al., 2017) or ask verbal questions (Jaiswal, Song, & Valstar, 2019). However, there is no clear agreement regarding the best stimuli to trigger facial behaviours for automatic personality and depression analysis. Thus, exploring a superior way to encode multi-scale facial dynamics as well as utilizing all available information are the main targets of this thesis. In this sense, Sec.3.3 explicitly summarized temporal modelling strategies of the reviewed approaches and their shortcomings, and Sec.3.4 discussed existing literature that encodes fixed-length feature representation of variable length videos.

3.1 Automatic personality analysis

As an essential part of machine learning, labels are important for models' training and validation. Although personality trait models have been criticized that they are purely descriptive and do not correspond to actual characteristics of individuals (Junior et al., 2018), this type of method is still the dominant way to model people's personality as they can provide explicit and quantifiable description across various aspects of human personality. Practically, the verbal questionnaires reported by individuals themselves or others who observed them have been frequently adopted as the self-reported or apparent personality traits labels (Corr & Matthews, 2009). Among those questionnaires, the Big Five Inventory (BFI) developed by John et al. (John, Donahue, & Kentle, 1991) has been widely used to measure the five traits of people's personalities, which contains 44 questions. Besides, Beatrice et al. (Rammstedt & John, 2007) and Gosling et al. (Gosling, Rentfrow, & Swann Jr, 2003) have proposed short and time-saving versions of BFI, containing only 10-items, which also achieved many attentions. Based on such labels, the following two subsections will systematically review existing automatic personality analysis approaches.

3.1.1 Frame/short segment-level personality modelling

The dimensions of face images used in automatic personality analysis studies are usually high, i.e., the resolutions are usually higher than 120 × 120 × 3, resulting in at least 40, 000 dimensions for each image. Then, it is challenging to directly feed such high-dimensional data to standard ML models for personality/depression analysis. Instead, a large part of existing methods attempted to firstly extract low-dimensional descriptors from each given image or short video segment (usually less than 20 frames), and then predict personality from low-dimensional descriptors. These approaches generally consist of two main steps: 1. image-level or short segment-level feature extraction using either hand-crafted methods (Aran & Gatica-Perez, 2013b; Pianesi, Mana, Cappelletti, Lepri, & Zancanaro, 2008; Lepri, Mana, Cappelletti, Pianesi, & Zancanaro, 2009; Kalimeri, Lepri, & Pianesi, 2010; Fang et al., 2016; Staiano, Lepri, Subramanian, Sebe, & Pianesi, 2011; Joshi, Gunes, & Goecke, 2014; Okada, Aran, & Gatica-Perez, 2015) or deep learning models (Batrinca, Lepri, & Pianesi, 2011; Güçlütürk et al., 2016; Subramaniam, Patel, Mishra, Balasubramanian, & Mittal, 2016; Wei, Zhang, Zhang, & Wu, 2018; Kampman et al., 2018); 2. feeding image/segment-level features into pre-trained classifiers or regressors to generate predictions. In particular, for video-based approach, the decision-level fusion has been frequently made from all frame/segment-level predictions of a video, to generate the final predictions.

Still image-based applications

While personality trait is defined to be stable over time for a particular person but differs from others (Kassin, 2003), some studies were attempted to predict personality from a single image such as the profile image in social media. These approaches usually start with facial landmarks detection, by which face regions can be obtained with a simple background. Then, based on the landmarks, various mid-level cues or semantic facial appearance and geometrical features can be extracted for the personality analysis.

For example, Joo et al. (Joo et al., 2015) conducted studies on face images of 650 American politicians of white ethnicity, including Senators, Congressmen and Governors.

This approach extracted Histogram of oriented gradients (HOG) features as the low-level per frame representation. Additionally, it calculated mid-level binary attributes: Glasses, Bald, Blonde, etc., as well as scalar attributes: Eye-height, Eye-width, Drooping-eye, etc. The RankSVM was then employed to combine them with high-level social dimensions to predict personality traits scores. Yan et al. (Yan et al., 2016) firstly divided each face image into several semantic regions, and extracted multiple low-level descriptors from each region, respectively. Then, SVM is employed to learn correlations between facial regions from these low-level descriptors, to explore the relationship between facial appearance and personality impressions. Vernon et al. (Vernon, Sutherland, Young, & Hartley, 2014) proposed to extracted 179 fiducial points to locate key features from each face and used ANN to calculated 65 numerical attributes from them, which contain both local or global facial geometric and appearance information. After that, the computed numerical attributes were reduced to three factors corresponding to three personality traits. A cascaded network was then adopted to synthesize a cartoon face-like image by decoding the 3 factors to 393 image properties, according to which the social traits can be assessed.

Images displayed in users' social media are also important sources to understand their personality. Celli et al. (Celli, Bruni, & Lepri, 2014) analyzed profile images from Facebook. They utilised the bag-of-visual-words model to represent each image. The extracted visual words were then fed to a series of regressors such as Support Vector Regression (SVR), Naive Bayes, decision tree and logistic regression, to predict users' personality traits. Dhall et al. (Dhall & Hoey, 2016) extracted both hand-crafted and deep-learned facial region features from Twitter profile images. It is worth mentioning that the background information was also taken into consideration in this work. The results showed that better predictions could be produced when using the background as additional information compared to use users' body or face regions only. In contrast, Qin et al. (Qin et al., 2018) recorded face images in a very controlled environment to avoid the influence of facial cosmetics, jewellery, and other decorations. In addition, 'irrelevant' information such as hair, clothes, background, etc. was also removed. Based on the pure face, this study extracted five hand-crafted features (HOG, Local Binary

Pattern (LBP), Gabor, Scale-invariant feature transform (SIFT) and Generalized Search Trees (Gist)) which were fed to several standard classifiers and regressors, e.g. Parzen Window, Decision Tree, etc. to estimate self-reported personality (16PF) and intelligence. The experiment results demonstrated that despite some personality traits can be reliably predicted from the facial features, others may largely depend on the social environment.

Besides the face image, some studies investigated other types of images posted on social media, and claimed that they can also partially reflect users' behaviours and tendencies. Guntuku et al. (Guntuku, Qiu, Roy, Lin, & Jakhetiya, 2015) researched 123 images collected from Sina Weibo (Chinese Twitter). This study extracted several typical hand-crafted features such as Colour Histograms, SIFT, LBP, from each image, and fed them to SVM with RBF kernel to detect mid-level descriptors, from which personality traits can be estimated. The results showed that these mid-level cues outperformed low-level hand-crafted features for most traits. Based on Instagram, Ferwerda et al. (Ferwerda, Schedl, & Tkalcic, 2016) explored the way users manipulate pictures by investigating the relationship between image features extracted by colour filters, and real personality traits.

Video-based applications

Videos can provide temporal information in addition to the static facial displays, which have been claimed to be crucial to reflect individuals' internal dispositions. As a result, video-based approaches can usually place personality analysis on a new level with a broader range of possible applications. Although videos were provided, some early studies still failed to utilise temporal information. For example, Biel et al. (Biel, Teijeiro-Mosquera, & Gatica-Perez, 2012) addressed the personality traits estimation task by extract frame-wise facial activity statistics and then utilised SVR to learn them independently. This work evaluated the most prominent facial expressions for personality traits modelling, claiming that Extraversion is the easiest trait to predict. This conclusion also has been validated by (Biel, Aran, & Gatica-Perez, 2011; Biel & Gatica-Perez, 2013). In other words, this approach still infers personality at the image-level.

As an extension of (Biel et al., 2012), Teijeiro et al. (Teijeiro-Mosquera, Biel, Alba-

Castro, & Gatica-Perez, 2015) studied the relationship between facial expressions and apparent personality traits in Vlog. Instead of modelling personality at the still image level, they introduced a short time-window to model temporal dynamics from the per frame statistical representation of facial expressions. It investigated the influence of the video slice duration and facial location on personality traits estimation, showing that the quality of extraversion predictions is negatively correlated with time-window length. Besides, this study also found that the video slice at the beginning of each video is enough to form viewers' impressions. Aran et al. (Aran & Gatica-Perez, 2013a) also focused on using temporal dynamics of the short segment to infer personality traits. In particular, they first encoded dynamics of each video segment into a weighted Motion Energy Images (MEI), from which the dynamic features were extracted. Then, Ridge Regression and SVM were employed to analyze extracted features for personality prediction.

Besides the hand-crafted approaches, various deep learning models also have been introduced to video-based personality analysis studies. Batrinca et al. (Batrinca et al., 2011) proposed a Descriptor Aggregation Network (DAN) to extract frame-wise features at multiple spatial resolutions. This method was later adopted to return per frame personality predictions in (Ventura, Masip, & Lapedriza, 2017). The result shows that facial information, especially eyes, nose and mouth, plays key roles in the personality trait estimation. Guccluturk et al. (Güçlütürk et al., 2016) proposed an audio-visual Residual Network for personality traits estimation. This structure consists of two streams: visual stream and audio stream, by which visual and audio features were deep learned. Both features were then combined at the fully connected layer (FC) to provide a personality prediction for each frame. The final video-level prediction is then made by averaging these frame-level predictions. Wei et al. (Wei et al., 2018) proposed a Deep Bi-modal Regression Network to extract per frame audio and visual features. For visual information, they also applied DAN to aggregate features from different convolution layers by using either global average pooling or global max pooling. Similar to above approaches, the final prediction is made by the fusion of frame-level predictions. In short, all the aforementioned deep learning frameworks predict personality from still images without utilizing temporal information.

To consider important dynamic cues, another CNN-based approach (Subramaniam et al., 2016) employed Long-short-term-memory Networks (LSTMs). This approach consists of three steps: 1. videos are firstly divided into several equal-length short segments, where a frame from each video segment is selected to generate a 3D aligned cropped face image; 2. The 3D face images are fed into either CNNs or LSTMs; 3. Personality traits were predicted at the frame-level, and then the video-level prediction is made by the decision-level fusion of frame-level predictions.

While most studies aimed at returning a single personality traits' label/score for each person, Celiktutan and Gunes attempted to estimate apparent personality traits continuously over time. In (Celiktutan & Gunes, 2014), human annotators were asked to provide a big-five rating for each frame, and then several low-level hand-crafted features are combined to predict personality. Celiktutan et al. (Celiktutan & Gunes, 2015) also explored the relationship between frame-level mid-level facial cues (gaze, attention and head) and self-reported personality traits in the context of human-robot interactions. Both results showed that there is a plausible relationship between the facial expressions, personality traits and social dimensions. These works have been further investigated as a part of the personality traits challenge (Celiktutan, Eyben, Sariyanidi, Gunes, & Schuller, 2014) and presented as a demo in (Celiktutan, Sariyanidi, & Gunes, 2015). Besides, Subramanian et al. (Subramanian, Yan, Staiano, Lanz, & Sebe, 2013) also investigated the relationship between frame-level facial cues and the social attention descriptor (automatically detected head poses). They found that social attention descriptors are excellent predictors of the Extraversion and Neuroticism traits.

In summary, all the methods reviewed above attempt to learn personality from a single frame or a super-short segment, i.e., assuming video-level personality labels as the frame/segment-level labels to train Machine Learning (ML) models. This is problematic because people with different personalities can of course display the exact same configuration of behaviour primitives in a single frame (e.g., facial expression, head pose, etc.). What indicates personality is how they change over time in response to certain stimuli.

Thus, a training strategy that learns relations between short segments and personality labels would lead to an ill-posed machine learning task, i.e., the same input pattern has multiple labels, making it practically impossible to learn a good hypothesis. Additionally, while a person's frame/segment-level facial status can change rapidly in a short period, video-level (event) labels will remain unchanged for each video, as personality traits do not change on that time scale. Therefore, we assume that a brief segment (a single frame or couple of frames) of expressive behaviours does not carry sufficient information to encode a personality trait-specific pattern.

3.1.2 Video-level personality modelling

Since it is not realistic to infer reliable personality traits from a frame or a super short segment, some studies constructed video-level descriptors and attempt to model reliable and stable personality traits from long-term human behaviours.

To generate video-level descriptors, computing global statistics of hand-crafted descriptors from all frames/segments (Bekhouche et al., 2017; Aran & Gatica-Perez, 2013b; Fang et al., 2016; Nguyen et al., 2013) is the most popular way. However, these reviewed descriptors usually fail to retain detailed temporal correlations between frames/segments, which are essential components of facial behaviours, and crucial to reflect personality. The automatic system proposed by (Joshi et al., 2014) is a typical example. This work calculated the mean and standard deviation of PHOG features (extracted per frame) for each video, as the video-level facial behaviour descriptor, and then fed them to linear SVR for traits prediction. Aran et al. (Aran & Gatica-Perez, 2013b) aimed to infer personality of participants from videos of small group meetings, from which multiple hand-crafted audio and visual features were computed for each frame. Then, statistics such as average speaking turn, prosodic features and visual activity were computed for each video, represented as video-level descriptors, which were fed to Ridge Regression to estimate personality traits. A similar method was proposed in (Pianesi et al., 2008), where hand-crafted acoustic and visual features were first extracted from each frame, and then video-level descriptors were obtained by computing their statistics. The final predictions were produced by feeding video-level descriptors to SVM. Bekhouche et al. (Bekhouche et al., 2017) firstly extracted Local Phase Quantization (LPQ) and Binarized Statistical Image Features (BSIF) from each face image of the video. Then, Pyramid Multi-Level (PML) was introduced to combine both features at the frame-level. The long-term information of a whole video was obtained by averaging the features over all frames. After that, SVR and a Gaussian Process Regression (GPR) were employed to process these features for personality traits estimation.

Besides the low-level features, mid-level cues were also found to be informative to personality. In (Staiano et al., 2011), mid-level visual attention cues (represented as the head pose and gaze) were extracted for personality analysis. Similarly, human postures and gestures cues during the speaking were considered in (Nguyen et al., 2013). The statistics of these frame-wise mid-level cues were computed as the video-level descriptor and fed to Ridge Regression and Random Forest for personality prediction. In (Fang et al., 2016), three types of video-level features, including intra-personal (related to only one participant), dyadic (related to a pair of participants) and one vs all features (related to one participant versus the other members of the group), were proposed. In particular, they were computed from statistics of frame-wise audio and visual cues, such as speech activity, prosodic and motion activity. Then, SVR and Ridge Regression were used to learn these features to predict personality traits.

Instead of using statistics, other types of video-level descriptors produced from the mid-level visual and acoustic cues also have been widely explored. For example, in (Lepri et al., 2012; Fleeson, 2001), authors employed medium-grained meeting behaviours such as automatically detected speaking time and social attention features to classify extraversion trait. It concluded that speaking time and social gaze are effective indicators of extraversion, and the distribution of peers' social attention is a crucial clue. Based on the frame-level hand-crafted mid-level visual features, e.g., body/head motion, prosodic features, etc., Okada et al. (Okada et al., 2015) proposed to construct a multi-channel time-series binary signal for each video. They then explored a co-occurrent pattern of them using a graph clustering algorithm, which is treated as the video-level non-verbal behaviour descriptor. Finally, the personality traits and social impressions are estimated based on the produced co-occurrent descriptors using Ridge Regression and linear SVM.

While the approaches described above modelled long-term information from extracted descriptors, some systems proposed to model it at the input level, e.g., summarizing all frames into a single matrix before feeding to ML models. Kampman et al. (Kampman et al., 2018) introduced CNNs for personality prediction, where a tri-modal architecture extracts feature from audio, video and text. Interestingly, this work only took a random frame from each video for CNN-based feature extraction. In other words, they only used a single image to represent the whole video. Meanwhile, for the audio part, it split raw waveforms as the raw waveform itself and the signal with squared amplitude, as input to CNNs. Zhang et al. (L. Zhang et al., 2019) extended the temporal segment networks (L. Wang et al., 2016) to the personality analysis task. The method operates on a pool of sparsely sampled faces from the entire video, i.e., a face image is selected from each short segment. Then, a consensus strategy is employed to process all the selected face images, to produce video-level predictions for personality traits. In (Lepri et al., 2009), a Motion History Image (MHI) was computed as the representation of each 1-minute slice, from which a set of energy-related acoustic and visual features, e.g., head fidgeting, hands fidgeting, pose fidgeting, etc, were extracted. The result indicates a similar conclusion made by (Pentland, 2006) and (Kalimeri et al., 2010), i.e., 1-minute slices are long enough to compute reliable features. Meanwhile, this length is also small enough to capture the transient nature of social behavior.

3.2 Automatic depression analysis

Non-verbal expressive behaviours are also informative to people's depression status. In the past decade, automatic depression analysis has attracted a lot of attentions, where a series of challenges have been organized (M. Valstar et al., 2013, 2014, 2016; Ringeval et al., 2017, 2019). Among these approaches, a large part of them (Wen, Li, Guo, & Zhu, 2015; Zhou, Jin, Shang, & Guo, 2018; Zhu, Shang, Shao, & Guo, 2017; Jan, Meng, Gaus, & Zhang, 2018; L. Yang et al., 2018) base their predictions on the non-verbal audio-visual behaviours of participants expressed during interviews.

3.2.1 Frame/segment-level depression modelling

Early automatic depression analysis works (Meng et al., 2013; Gupta et al., 2014; Senoussaoui, Sarria-Paja, Santos, & Falk, 2014; Pampouchidou et al., 2016) generally used traditional Machine Learning models, e.g. SVR (Gong & Poellabauer, 2017; M. Valstar et al., 2013), decision tree (Sun et al., 2017; L. Yang et al., 2018, 2016), Logistic regression (Dibeklioğlu, Hammal, & Cohn, 2018), etc., to predict depression from a single frame or short segment using traditional hand-crafted features such as LBP (Dhall & Goecke, 2015; L. He, Jiang, & Sahli, 2018), Low-Level Descriptor (LLD) (Sun et al., 2017; Stepanov et al., 2018; L. Yang et al., 2018), HOG (M. Valstar et al., 2014), etc). In addition, the DepressNet proposed by Zhou et al. (Zhou et al., 2018) is a typical deep learning approach that analyzes depression at the frame level. It learns depression representations with visual explanation, aiming to identify the salient facial region for depressed people. In this approach, the facial region that is most informative to depression was highlighted and used to predict depression at the frame level. The video-level depression score was computed by averaging the predictions from all frames.

Meanwhile, the short segment-based approaches can retain short-term temporal information in the input and force the back-end model to consider it. This type of approaches generally divide each video into several equal-length segments, and learn features from each segment independently. For example, Meng et al. (Meng et al., 2013) extracted LBP and EOH as visual features and LLD as audio features. Then, this approach applied Motion History Histogram (MHH) to summarize dynamics from features of each short segment of a video. These dynamics features were then individually fed to Partial Least Square (PLS) regression to return segment-level depression predictions, where the video-level prediction is made by the decision-level fusion of them using the linear opinion pool. Yang et al. (L. Yang et al., 2018) proposed to select several equal length segments in each video to balance the number of depressed and non-depressed training examples. They also proposed a Histogram of Displacement Range (HDR) method that records the dynamics of facial landmarks in a video segment. They used CNNs to learn deep features from hand-crafted audio and video descriptors (including HDR) and the final decision is made by the decision-level fusion of audio, video and text predictions using decision trees. Ma et al. (Ma, Yang, Chen, Huang, & Wang, 2016) proposed a DeepAudioNet to deep learn vocal features at the short segment-level, of which the outputs are fed to Long-Short-Term-Memory Networks (LSTMs), to return a depression prediction for each segment. Gupta et al. (Gupta et al., 2014) used LBP-TOP to summarize short-term temporal information from each segment, which is combined with motion features and facial landmarks as the segment-level representation. A feature selection step is then applied to select features to train an SVR model for segment-level depression prediction. Al Jazaery et al. (Al Jazaery & Guo, 2018) employed C3D network to extract short-term depression-related dynamic features from each short video segment. Then, these features were fed to a Recurrent Neural Network (RNN) to model long-term dynamics and make segment-level predictions. The final score was obtained by averaging all segment-level predictions. A similar approach was proposed by Melo et al. (de Melo, Granger, & Hadid, 2019), which also used C3D CNNs to model short-term dynamics and predict depression severity at the short segment-level and then the final decision was also made by averaging all segment-level predictions. Recently, Haque et al. (Haque, Guo, Miner, & Fei-Fei, 2018) employed Causal Convolutional Networks (C-CNN) (Bai, Kolter, & Koltun, 2018) to predict depression severity from audio, text and 3D facial landmarks at the long sentence-level, which has been shown to outperform recurrent neural networks (RNNs) on long sequences for inferring depression.

3.2.2 Video-level depression modelling

While the reviewed segment-level based approaches can efficiently encode short-term dynamics, most of them ignored long-term temporal behaviour patterns of participants except RNNs and LSTMs were used in (Al Jazaery & Guo, 2018) and (Ma et al., 2016). These approaches assumed the video-level depression labels as the frame/segment-level labels when training models. As discussed before, behaviour descriptors extracted from a single frame or a short segment can be ambiguous and explained by various causes, e.g., a smile may be caused by feeling happy or feeling helpless. Additionally, the same short-term behaviours can be expressed by subjects with different levels of depression. Thus, this strategy would make trained models practically impossible to learn a good hypothesis. In other words, depression levels can be more reliably described using the entire video rather than short segments of the video.

For these reasons, some studies devoted to building a video-level descriptor for each clip, where the Gaussian Mixture Model (GMM) and its extensions have been frequently employed. For example, Williamson et al. (Williamson et al., 2013; Williamson, Quatieri, Helfer, Ciccarelli, & Mehta, 2014) who were the winners of the AVEC 2013 (M. Valstar et al., 2013) and AVEC 2014 depression challenges (M. Valstar et al., 2014), based their approaches on audio data. They utilised formant frequencies and delta-mel-cepstra to represent underlying changes in vocal tract shape and dynamics. After that, by exploring the correlations between these features and using PCA, an 11-dimensional feature vector (five principal components for the formant domain and six principal components for the delta-mel-cepstral domain) is obtained. Finally, a Gaussian Staircase Model, an extension of the GMM, was introduced as the regression model to summarize descriptors of each clip. The approach proposed by Cummins et al. (Cummins et al., 2013) is also based on GMM where a GMM-UBM model was applied to learn features representing each entire clip, which contain both audio and visual information. Jain et al. (Jain, Crowley, Dey, & Lux, 2014) used GMM (Fisher Vector) to encode the segment-level extracted LBP-TOP, HOG, HOF and MBH features into a video-level visual descriptor. Nasir et al. (Nasir, Jati, Shivakumar, Nallan Chakravarthula, & Georgiou, 2016) proposed to use i-vector, which is an extension model of GMM. This model summarizes several segment-level extracted audio features such as TECC and MFCC, and encoded them into a video-level descriptor. He et al. (L. He et al., 2018) is another typical approach using the combination of hand-crafted features and traditional ML models. This work extended the LBP-TOP feature to MRLBP-TOP for extracting short-term dynamics and then Fisher Vector is used to aggregate them as the long-term representation.

The topics of the interview tasks decide the context of the interview, and thus would trigger participants' expressive behaviours in different ways. As a result, a few studies constructed video-level descriptors based on interview topics. Gong et al. (Gong & Poellabauer, 2017) investigated the relationship between the interview topics and depression level. They found that there are 83 different topics in video contents in DAIC-WOZ database (Gratch et al., 2014) and each clip only contains a few of them. To predict depression severity, they constructed a long vector containing each topic's feature in a video, including the occurrence of the topic, statistics of hand-crafted audio features, and facial attributes. Sun et al (Sun et al., 2017) also investigated the interview topics, where they found that only 6 topics are potentially highly correlated to the depression levels.

3.3 Temporal modelling of facial behaviours

Temporal dynamics are important sources of information for video-based face analysis. The basis of this thesis is that personality traits and depression status can be reflected by expressive facial behaviours which are made up of both appearance and dynamics. Traditional hand-crafted features and recently proposed deep learning architectures since AlexNet (Krizhevsky et al., 2012) have been proved to be powerful to encode spatial information from images. However, the optimum ways to encode the temporal dynamics of facial behaviours for personality and depression analysis have not been fully explored. Thus, this section explicitly reviews and summarizes the typical temporal models used for related face tasks. In particular, this section separately reviews the short-term temporal modelling methods and long-term temporal modelling methods.

3.3.1 Short-term temporal models

The short-term temporal modeling of the face usually accomplished either by generating a single set of features from multiple consecutive frames at a time (early modeling of temporal dynamics (Bilen et al., 2017)) or by using memory-based models, such as Recurrent Neural Networks (RNNs) and Markov Models (late modeling of temporal dynamics (Al Jazaery & Guo, 2018; de Melo et al., 2019)), or by a combination of both (Jaiswal & Valstar, 2016; Song, Sánchez-Lozano, Kumar Tellamekala, et al., 2019).

Among these solutions, the late modeling of short-term facial dynamics usually achieved by computing the temporal correlations between frame-wisely extracted descriptors. To do so, some early studies extended hand-crafted spatial features to the temporal domain. For example, some features, e.g., LBP, HOG, etc., have been extended the spatial domain to the temporal dimension, referred as the Three Orthogonal Planes (TOP), and were widely used by AVEC baselines (Schuller, Valster, Eyben, Cowie, & Pantic, 2012; M. Valstar et al., 2013, 2014). Particularly, the LBP features are extended to the temporal domain as the LBP-TOP in (Kaltwang, Todorovic, & Pantic, 2016). When combining this LBP-TOP with a sparse regression method, the excellent AU recognition and affect estimation performances were yielded on the SEMAINE database (McKeown, Valstar, Cowie, Pantic, & Schroder, 2012). In (Kaya, Gürpınar, & Salah, 2017), histogram-based features, such as LPQ, LBP and LGBP, were extended to the temporal dimension, which were further combined with deep-learned features to summarize spatio-temporal patterns. Gupta et al. (Gupta et al., 2014) also extend LBP to LBP-TOP to encode short-term temporal information for personality analysis.

However, the TOP extension of features grows drastically in complexity as the number of frames increases, and thus learning temporal models is a better choice. While graphical models such as Hidden Markov Models (HMMs) or Conditional Random Fields (CRFs) provide powerful temporal representations, they are prone to failure when modeling long-term dynamics. These drawbacks can be tackled with Recurrent Neural Networks (RNNs), which are feed forward networks of latent states that can be learned through back-propagation. RNN models exploit the temporal information by applying latent variables that are supposed to model the intrinsic correlation that exists between the input and the output at a given frame, conditioned to the latent states at previous frames. This type of model, including Long-short-term-memory networks (LSTM), Gated Recurrent Units (GRUs), etc., can be used with either hand-crafted features or in combination with CNNs.

While RNNs are relatively sensitive to the hyper-parameters and training data, which leads them to be overfitting, some extensions handling such back-propagation problems have been proposed too. As a result, RNN-based models have been widely adopted to model temporal dynamics of emotions and facial actions from the face (Jaiswal & Valstar, 2016; K. Zhang et al., 2017; Kollias et al., 2019b; Jaiswal & Valstar, 2016) recently. In (Nicolaou, Gunes, & Pantic, 2011), a Bidirectional-LSTM (BLSTM) architecture is employed to learn hand-crafted features, which showed superior performance than using SVR as the back-end model. This is because the BLSTM can learn facial dynamics from videos in addition to the spatial information. Hasani et al. (Hasani & Mahoor, 2017) extracted features using Inception module (Szegedy et al., 2015). When feeding the deep-learned features to an LSTM, the prediction performance was enhanced compared to using the Inception module only in a per-frame basis. Similar approaches were adopted in (Khorrami, Le Paine, Brady, Dagli, & Huang, 2016) and (Kollias et al., 2019b), where relatively shallow CNNs were used in combination with an RNN or GRU, to learn short-term spatio-temporal facial patterns. In (Han, Zhang, Cummins, Ringeval, & Schuller, 2016), the output of RNNs is combined with an SVR, thus preventing the former to incur in overfitting, and the latter not to consider the temporal domain. The proposed approach, coined Strength Modeling algorithm, applies the two models in a hierarchical manner.

Besides the RNNs and TOP-based models, other temporal models that aims to learn dynamics from features, such as co-occurrent pattern (Okada et al., 2015), statistic-based dynamic features (Teijeiro-Mosquera et al., 2015), Histogram of Displacement Range (HDR) method (L. Yang et al., 2018) etc., also have been investigated in recent years, to infer short-term dynamics from video segments. In (Nicolle et al., 2012), global and local features were extended to the temporal domain through the magnitude of the Fourier transform of each of them.

However, the late modelling methods described above mainly focused on learning temporal information from the extracted frame/segment-level features rather than face images themselves. Since the feature extraction significantly reduces the input face image's dimensionality, this process may result in the loss of some critical spatio-temporal patterns. Instead, some studies attempted to model temporal evolution directly from the face image sequence, and consequently, the dynamics can be encoded in the context of the face. For example, in (de Melo et al., 2019), the C3D network was utilised to learn short-term temporal patterns from a set of stacked input frames. However, this model would become high-complex when modelling relative long sequences, as the number of input channels equals the number of frames. Thus, the number of input frames are usually limited to less than 20. Aran et al. (Aran & Gatica-Perez, 2013a) directly encoded facial dynamics of an image sequence by producing a weighted Motion Energy Images (MEI), which can be easily forwarded to back-end models for feature extraction. Lepri et al. (Lepri et al., 2009) proposed to produce a Motion History Images (MHI) as the dynamic representation of a 1-minute slice, from which a set of energy-related acoustic and visual features, e.g. head fidgeting, hands fidgeting, pose fidgeting, etc, were extracted. Similarly, Bilen et al. (Bilen et al., 2017) proposed a dynamic image algorithm that also encodes dynamics of an image sequence into an image-size matrix. The dynamic image is an on-the-fly sequence descriptor, i.e., it computes a representation for a given set of frames. The form of this descriptor is that of a kernel that, when projected into the preceding frames, assigns to each a score that allows them to be sorted in time. This approach was further extended to the face domain in (Song, Sánchez-Lozano, Kumar Tellamekala, et al., 2019), and achieved good performance.

Some researchers also proposed to combine both early and late modeling of temporal dynamics, and clearly shown the superior performance in comparison to use either of them only. Jaiswal et al. (Jaiswal & Valstar, 2016) proposed a spatio-temporal CNN on a set of stacked input frames to encode facial dynamics at the input level for facial AU recognition, with its center being the target frame. The output of the CNN is forwarded to a BLSTM to additionally model the temporal information at the feature-level. In (Al Jazaery & Guo, 2018), C3D network is employed to model dynamics directly from a set of frames and then RNNs is introduced to model dynamics from frame-wisely extracted features again, for depression analysis.

In summary, most late temporal modelling approaches attempted to learn **temporal patterns of facial features** rather than the **temporal patterns of facial images**. However, during the feature extraction process which significantly reduces the dimensionality of the input data, some crucial dynamic cues may lost. Although some works have proposed to encode dynamics in the context of the original data (face), by constructing a dynamic representation from a set of frames (Jaiswal & Valstar, 2016; Bilen et al., 2017; Aran & Gatica-Perez, 2013a), these methods have some practical drawbacks, as the learning of such approaches can become quite complex for long-term sequences. To this end, this thesis attempted to encode temporal information directly from face images.

3.3.2 Long-term temporal models

Both personality and depression status are stable in a relatively long-term that would be much longer than the duration of an interview video, and thus relatively long-term information may provide more stable predictions. The majority of related approaches reviewed in this chapter focused on using recurrent or latent-based models to model long-term temporal information of the face video. For example, in (Subramaniam et al., 2016; Al Jazaery & Guo, 2018; de Melo et al., 2019; Ma et al., 2016), authors applied RNN-based models, e.g., LSTM, BLSTM, to model temporal dynamics from of an entire clip, for either personality traits or depression level prediction. In particular, the employed models learn dynamics from all frame/segment-level extracted features, i.e., from the beginning of the clip to the end of the clip, which are extracted by CNNs. Thus the long-term temporal dynamics can be obtained through this process. Instead of using RNNs, Haque et al. (Haque et al., 2018) employed Causal Convolutional Networks (C-CNN) (Bai et al., 2018) to model dynamics for long sentences, which has been shown to outperform recurrent neural networks (RNNs) for such cases. Zhang et al. (L. Zhang et al., 2019) extended the temporal segment networks (L. Wang et al., 2016) to personality analysis. Then, they proposed a consensus strategy to process all the temporal correlations of all selected faces.

In summary, memory-based neural networks such as RNNs, are easy to be overfitting, and they usually need frame/segment-level annotations during the training. As a result, recent approaches have frequently assumed video-level labels as the frame/segment-level labels. However, as discussed in Sec.1, this would result in the trained models having poor generalization capability because a frame/short-segment is not sufficient to represent either personality traits or depression status. Other approaches, such as Causal Convolutional Networks, haven't shown the outstanding capability to model very long-term (more than 500 frames) temporal information for personality and depression analysis. Meanwhile, the use of temporal segment networks (L. Wang et al., 2016) leads to a large part of frames to be removed (L. Zhang et al., 2019), resulting in a significant loss of local temporal and spatial facial information. To address such issues, this thesis proposed two long-term facial dynamics modelling approaches, namely **spectral representation** and **person-specific facial dynamic descriptor**. Both of them can not only encode the video-level facial dynamics of variable-length videos into length-independent dynamic representations, but also avoid using video labels as the frame/segment-level label during models' training.

3.4 Video-level representations encoding of variable-length video data

In real-world applications, collected videos that contains human behaviours usually have variable duration (e.g., videos in SEMAINE dataset (McKeown et al., 2012) range from 25 minutes to 50 minutes). While standard ML models require fixed-size inputs, how to represent each video of an arbitrary number of frames using a length-independent representation while retaining important clues is an important research question. A straightforward solution is to select a fixed number of frames/segments from each video to construct a fixed-length sub-video. For example, Bishay et al. (Bishay, Priebe, & Patras, 2019) divided the shortest videos into N clips without overlap, and longer videos into N equally-spaced clips. As a result, each video is represented by N sub-clips of equal length. Similarly, a fixed number of equal-length video segments are extracted from each depression video in (L. Yang et al., 2018). Li et al. (Li et al., 2020) selects 32 frames from each video for personality recognition. In addition, a large number of action recognition approaches also utilised this strategy, i.e., selecting a fixed number of frames (Feichtenhofer, Fan, Malik, & He, 2019; C. Yang, Xu, Shi, Dai, & Zhou, 2020; Bilen, Fernando, Gavves, Vedaldi, & Gould, 2016) or segments (Munro & Damen, 2020) to represent a whole video with an arbitrary duration.

The main drawback of the approaches described above is that they discarded some frames, which may contain crucial human behaviour clues. Thus some researchers proposed to re-sample per-frame representations of a video (which can be a multi-channel time-series data) to a fixed length by using interpolation, Dynamic Time Warping (DTW), etc. However, these approaches will distort the original signals. To avoid distortions, other studies employed fixed-size histograms or statistics to summarize the distribution of per-frame representations. Specifically, they generate video-level descriptors by computing statistics of features (Pampouchidou et al., 2016; Gupta et al., 2014; Gong & Poellabauer, 2017; Stepanov et al., 2018), using Gaussian Mixture Model (GMM) (Williamson et al., 2014; Cummins et al., 2013; Jain et al., 2014; Dhall & Goecke, 2015; Anis, Zakia, Mohamed, & Jeffrey, 2018; Hao et al., 2019) or fisher vector (Dhall & Goecke, 2015; L. He et al., 2018; Bishay, Palasek, Priebe, & Patras, 2019), etc. Although these methods summarized undistorted information, important temporal dynamics between segments/frames, such as the order of events, may be lost after creating the statistics. Besides, Vector of Local Aggregated Descriptors (VLAD) has been widely employed for video-based action recognition (Tu et al., 2019; Duta, Ionescu, Aizawa, & Sebe, 2017). The main strength of VLAD is that it constructs a fixed-size super vector to represent a video-level feature, regardless of the video's duration. While it can avoid: 1. disporting original data; 2. ignoring frames; 3. losing short-term temporal correlations between frames, this representation still lacks the capability to encode multi-scale long-term dynamics.

3.5 Summary

In summary, temporal dynamics are important assets for automatic personality and depression analysis. As described above, the information carried by short segments/single frames is not reliable for personality and depression analysis. In other words, pairing video-level personality/depression labels with short segments/single frames to train personality/depression analysis models is problematic. While some studies are devoted to extracting video-level visual descriptors for automatic personality and depression analysis, those approaches usually failed to retain multi-scale dynamics. Meanwhile, the systems based on task topics require human annotated verbal information to decide the time location of each topic, which makes them not being the fully automatic approaches. To this end, this thesis proposed two fully automatic long-term descriptor extraction approaches, namely spectral representation (Song et al., 2018, 2020) (Chapter. 5) and person-specific adaptation descriptor (Song, Jaiswal, et al., 2021) (Chapter. 4). Both of them can encode multi-scale facial behaviours dynamics of an entire video into a length-independent descriptor, for automatic depression and personality analysis, regardless of the length of the video.

Chapter 4

Self-supervised learning of person-specific facial dynamics

As discussed in Chapter 2 and Chapter 3, facial dynamics are important sources to reflect individuals' emotion, personality and depression status. Meanwhile, the personality trait is defined as the aspects of personality that are relatively stable over time but differ across individuals (Kassin, 2003). Motivated by this, this Chapter proposes to learn video-level (long-term) person-specific facial behaviour dynamics for personality traits estimation. However, while it is not possible to directly feed an entire video to networks, frame-level manual annotation of face videos is time-consuming, expensive and subjective. Moreover, the optimum frame-level label for inferring personality is still unclear. Meanwhile, although assuming the video-level personality label as the frame-level label is a popular solution that has been frequently used by existing approaches, it may not train good models. To deal with such issues, this thesis proposes a novel rank loss (Song, Sanchez, Shen, & Valstar, 2021) that allows networks to learn facial dynamics in a self-supervised manner without using any human annotations.

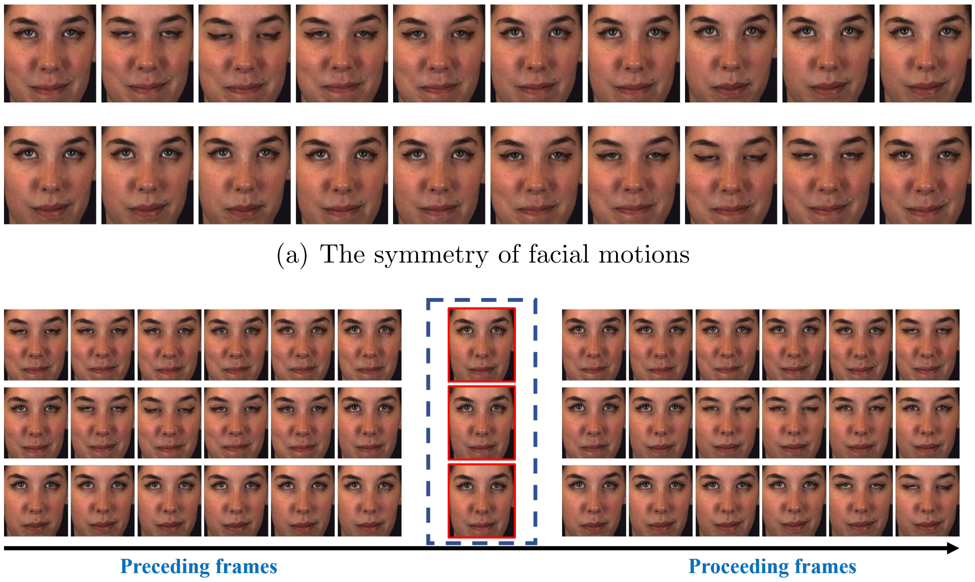
Specifically, this Chapter first explains the theoretical basis of the proposed self-supervised learning approach: dynamic image algorithm (Bilen et al., 2017), and shows how to extend this algorithm to model short-term facial dynamics from image sequences. Then, the self-supervised short-term facial dynamic modelling approach is proposed to learn generic short-term facial dynamics from unlabelled face videos, which utilises natural facial temporal evolution (the temporal orders of frames) to supervise models' training process, without requiring any human-annotated label. This approach can not only infer facial dynamics from any previous unseen still face image but also provide the basis for the video-level person-specific facial representation construction. Finally, this Chapter shows how to learn such video-level person-specific facial representation by training a set of intermediate filters using the same self-supervised learning algorithm, which are incorporated into the previous learned generic dynamic network. This way, the learned filters' weights are person-specific, making them a valuable source for modeling person-specific facial dynamics. Their weights are then concatenated as a person-specific representation, which can be directly used to predict the personality traits without needing other parts of the network.

Compared to the techniques such as Recurrent Neural Networks (RNNs) and Hidden Markov Model (HMM) that learn dynamics at the feature level, which may result in the loss of critical facial details, the proposed approaches model dynamics directly from the face. As reviewed in Sec. 2.7, there is convergent evidence indicating that personality and depression are reflected by specific tendencies to experience and express certain emotions. In this sense, the final section not only evaluates the proposed person-specific dynamic representation on the task of personality traits estimation, but also evaluates the short-term dynamics encoding capability of the proposed approach on the task of emotion (dimensional affect) estimation.

4.1 Dynamic facial models

To learn a dynamic descriptor from a short face video segment, this section describes a method that extends the dynamic image algorithm (Bilen et al., 2017) to the face domain. The dynamic image algorithm was originally presented as a sequence descriptor with a fixed size regardless of sequence length, which can be used in a plug-and-play format in CNN-based networks for action recognition. The form of this descriptor is that of a kernel that, when projected onto the preceding frames, assigns to each frame a score that allows them to be sorted in order of time. In other words, the goal is to have a sequence descriptor whose per-frame scores are increasingly similar to the final frame as the sorted sequence progresses. Such a representation is conjectured to be a good dynamic descriptor of a sequence. This assumption was empirically validated in (Bilen et al., 2017; Song, Sánchez-Lozano, Kumar Tellamekala, et al., 2019; Song, Sanchez, et al., 2021). This idea itself is very similar to temporal templates introduced by Bobick & Davis in 2001 (Bobick & Davis, 2001) but has proved to be a superior descriptor for action recognition (Bilen et al., 2017). The use of a summarized dynamic image allows CNN-based architectures that are designed to take still images to be easy to process an image sequence of variable lengths.

Although the dynamic image algorithm has been successfully applied to human action recognition, its extension to model the dynamics of facial actions is not straightforward because: 1. Bilen et al.(Bilen et al., 2016) made use of whole images to generate dynamic appearance, without segmentation of specific, semantically meaningful regions of objects (the human body, or the face, etc.), which are highly valuable for human action analysis; 2. Most human actions have asymmetric temporal patterns, displaying a clear arrow of time, which makes their temporal evolution predictable and distinguishable. In contrast, it can be observed that facial actions display a **symmetric** temporal pattern of changes in shape and appearance. As shown in Fig. 4.1(a), facial actions are in many cases indistinguishable from their temporally reversed counterparts. This means the facial actions are still plausible "motions" when reversed in time; 3. The human action recognition task returns a single prediction per video, and thus the objects (instruments, tools, etc.) in videos are good features for classification because the objects used in different human actions are normally different. However, the difference of people's faces (identities) can not be utilised for our task because the goal is to encode facial dynamics such as emotions to benefit personality/depression analysis, which means a person's face can display variable expressions and different people can display the same expressions.



(a) The symmetry of facial motions

(b) The ambiguity of facial motions

Figure 4.1: Temporal ambiguity and symmetry of facial motions. Fig. 4.1(a) illustrates the symmetric temporal pattern of facial actions. The lower facial action is the **temporally reversed counterpart** of the upper one, where both sequence can be a plausible facial action. Fig. 4.1(b) displays the temporal ambiguity of facial actions, as the same facial display (in the center) can occur in different facial actions.

4.1.1 Novelty and contributions

The main novelty of this approach resides in the encoding of short-term facial spatio-temporal information into a length-independent representation, in the context of the face. This work differs from (K. Zhang et al., 2017; Kollias et al., 2019b) as the proposed approach encodes short-term dynamics directly from the face images rather than extracted features. It also differs from that of Nicolle et al. (Nicolle et al., 2012) in that it does not rely on the frequency domain, as their approach contains nuisance factors that are hard to capture with a CNN. Finally, it differs from (Jaiswal & Valstar, 2016; de Melo et al., 2019; Al Jazaery & Guo, 2018) in that the dynamics of a sequence are encoded into a length-independent matrix that has the same size of the input images, allowing the use of a flexible number of frames, rather than as a concatenation of frames, which in practice limits the time extent of the short-term encoding.

Although the proposed approach is based on the dynamic image algorithm (Bilen et al., 2017), it differs from it in that: 1. while the original dynamic image is proposed as the sequence descriptor, the proposed approach constructs per frame dynamic representation;

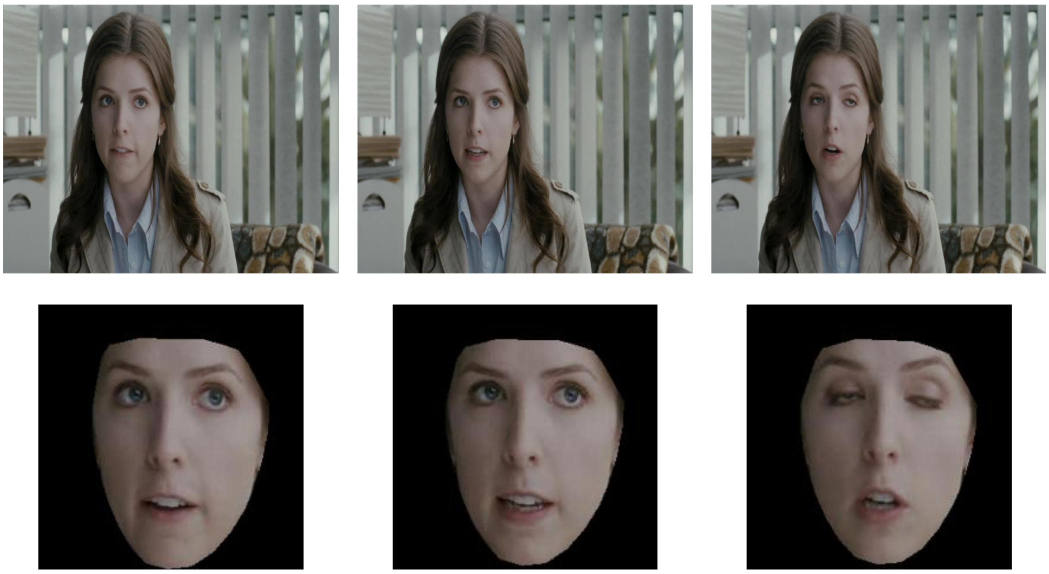


Figure 4.2: Examples of static face images (SFIs). The upper row displays original frames in the video and the lower row displays the corresponding SFIs.

2. the proposed Dynamic Facial Representation (DFR) model can jointly encode both preceding and succeeding dynamics rather than one of them; 3. the proposed approach takes the shape information into account, allowing only the dynamics of the target (face) rather than both target and background to be encoded.

4.1.2 Static facial image

The sequence-based approach starts with detecting a set of 68 facial landmarks for each video frame using OpenFace 2.0 toolkit (Baltrusaitis et al., 2018). Based on these landmarks, a binary mask is applied to the original face image, whereby only the pixels lying within the convex hull defined by the landmarks are set to one. This mask is applied to the input image to generate the static facial image (SFI), which basically accounts for the facial appearance. Then, only the face region is cropped and aligned. This way, the background noise is removed before the feature extraction process. Some examples are shown in Fig.4.2.

4.1.3 Dynamic facial image

Based on the obtained faces, this section extends the dynamic image algorithm to the face domain, summarizing the proceeding and succeeding temporal evolution of each given face image. The dynamic image is a matrix that has the same size as the input images. Its parameters are learnt to rank the position of the given frames from their features by implementing dot product between the per-frame features and the dynamic image. That is to say, a dynamic image is an operator that encodes the temporal evolution of the given sequence. In particular, it is worth noting that the score, defined as a dot product, represents the *similarity* between the generated representation d_t and the feature descriptor V_a. By enforcing this score to be higher for frames that are closer to the input image I_t, this approach are targeting a feature representation that is "decreasingly similar" to the features computed at adjacent frames. This way, the learned representation encodes dynamics as a summary of weighted variations. In this thesis, the term 'facial dynamics' represents the temporal evolution of facial behaviours, which can be learned from the adjacent sequence of the target face image. Importantly, the generated dynamic images are spatially corresponded to the given face images, allowing back-end CNNs to deep learn the spatio-temporal patterns in the context of the face. Consequently, by extending this algorithm to include shape and adapting it to the face domain by leveraging facial landmarks, the dynamic facial image (DFI) without background noise can be obtained.

Let $I_t \in \mathbb{R}^{m\times n}$ be the t-th face image of a sequence composed of T consecutive face-aligned images, all of size $m \times n$, and let $V_t = \frac{1}{\tau}\sum_{\tau=1}^{t} I_\tau$ be the average value image up to frame t. The mapping chosen in this paper is the same as that which attained highest performance in the original paper by Bilen et al. (Bilen et al., 2017), which defined ϕ to be the identity function. Let $\mathbf{d} \in \mathbb{R}^d$ be the DFI of the image sequence. The ranking score for frame t is defined as the dot product between $\mathbf{d}$ and V_t:

$$S(\mathbf{d}, V_t) = \langle \mathbf{d}, V_t \rangle \tag{4.1}$$

Thus, for the frame I_c, the goal is to learn a DFI $\mathbf{d_c}$ so that if $c > b > a > e$, then

$S(\mathbf{d_c}, V_b) > S(\mathbf{d_c}, V_a)$, where a, b, c, e are time stamp, i.e., I_e is the first frame of the subsequence and I_c is the last frame (target frame) of the subsequence. Specifically, the frame I_b that is temporally closer to the target frame I_c would have higher score than the frame I_a. In other words, $\mathbf{d_c}$ is learned so that when projected into the aggregated kernel of the input image size, it returns a score that sorts frames by time. The learned kernel $\mathbf{d_c}$ ranks these input SFIs based on their temporal orders, and hence encodes facial temporal evolution that provided by the sub-sequence (from I_e to I_c), making it a good facial behaviour descriptor. In order to learn $\mathbf{d_c}$, hinge loss is minimized between pairs of scores:

$$\mathbf{d_c}^* = \underset{\mathbf{d_c}}{\operatorname{argmin}}\, E(\mathbf{d_c}) \tag{4.2}$$

$$E(\mathbf{d_c}) = \frac{\lambda}{2}\|\mathbf{d_c}\|^2 + \gamma \sum_{b>a} \max\{0, 1 - S(\mathbf{d_c}, V_b) \ + S(\mathbf{d_c}, V_a)\} \tag{4.3}$$

where $\gamma = \frac{2}{c(c-1)}$, is the L2-norm regularised error. The second term in Eq. 4.3 defines the accumulated errors on the subset that are produced by the incorrectly ranked pairs. A pair $b > a$ is said to be correctly ranked if $S(\mathbf{d_c}, V_b) \geq S(\mathbf{d_c}, V_a) + 1$. The minimisation of Eq. 4.3 is accomplished with RankSVM (Smola & Schölkopf, 2004). The parameters in the final learned kernel $\mathbf{d_c}$ are in the real space. As this kernel has the same size of the input images, it can be interpreted as a three-channel raster image. It is worth highlighting that during both training and inference stages, each DFI is generated from a set of its proceeding frames using RankSVM, i.e., each DFI is a RankSVM kernel learned from the proceeding subsequence of the target facial image. In other words, the proposed approach trains a RankSVM kernel (DFI) for each training and testing frame individually.

4.1.4 Encoding bidirectional temporal evolution

The dynamic image was originally used as the short video descriptor for human action recognition. While it is possible to segment a long personality/depression video into a set of short segments and encoding each of them into a dynamic facial image, deciding the optimal start and end time stamp for each segment requires extra effort. On the contrary, this thesis aims to produce a per-frame dynamic facial representation (DFR) that can summarize both preceding and succeeding temporal evolution of the given frame. In other words, we propose to summarize bidirectional short-term facial dynamics for each frame.

While the DFI can be produced by the aforementioned dynamic algorithm by treating the given face image as the end of the sequence, summarizing the past temporal evolution of the given face, i.e., how was the current facial status evolved from the past, the succeeding sequence were not considered by this way. Thus, this section also proposes to generate the dynamic image of the succeeding sequence called Backward dynamic facial image (BDFI), which encodes the future temporal evolution of the given face, i.e., how would the current facial status change in the future. In this setting, the latest frame of the succeeding sequence is treated as the beginning of the sequence and the given frame is denoted as the end of the sequence. Specifically, T face-aligned frames are taken after the given face I_c, all of size $m \times n$. Then, the average value image down to frame t can be denoted as

$$V_t = \frac{1}{T - t + 1} \sum_{t}^{T} I_t \tag{4.4}$$

where $c \leq t \leq T$. V_t is defined as the average of a given feature mapping of the image, $\phi(I_t)$. Follow the similar process of the DFI algorithm, RankSVM with the loss function defined in Equation 4.3 and Equation 4.2 is employed to learn a kernel that if $q > c$, then $S(\mathbf{d_c}, V_q) < S(\mathbf{d_c}, V_t)$. As a result, this kernel can be treated as the representation of given frames as it summarized reversed dynamics of the image sequence after I_c. eq:argmin

4.1.5 Dynamic facial representation

Based on the methods described above, three facial representations, e.g., SFI, DFI and BDFI, can be produced for each frame of a video (except the first $T - 1$ and last $T - 1$ frames, where T denotes the length of the time-window). Each of them is a 3-channel matrix that has the same size as the original aligned face images. In this sense, this section proposes to combine three obtained representations as a 9-channel matrix for each frame, unifying spatial and bidirectional facial temporal evolution in a single dynamic facial representation (DFR). As a result, spatial and dynamic channels in the DFR are

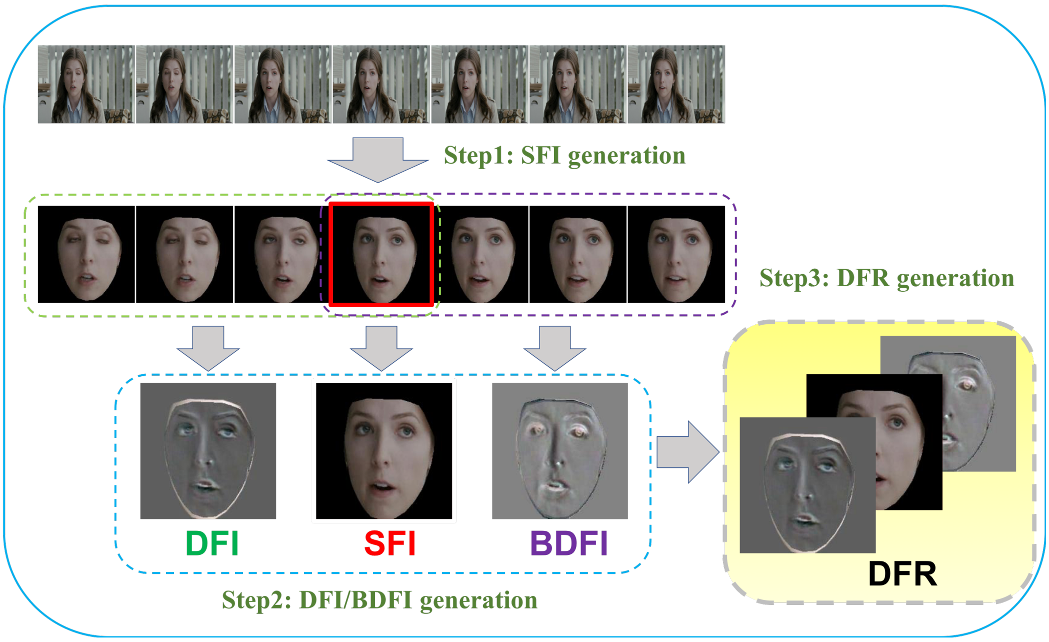


Figure 4.3: The diagram of the DFR generation. This approach starts with obtaining aligned face region from each original frame (Step 1). Then, DFI and BDFI are computed from preceding and succeeding frames of the given frame, based on the dynamic image algorithm (Bilen et al., 2017) (Step 2). By concatenating DFI, SFI and BDFI of the given frame, DFR is produced (Step 3).

spatially corresponded at the pixel level, which means they can guide the back-end model focusing on most **task-discrmintative** part of the face. Importantly, in doing so it keeps the size of the representation fixed regardless of the sequence length, making it easy to be combined with most developed front-end and back-end models for downstream tasks. The entire pipeline of this approach is illustrated in Fig.4.3.

4.2 Learning person-specific video-level facial dynamic representation

A common challenge in automatic personality traits analysis concerns how to bridge the gap between generic and person-specific models. While there are underlying similarities in how people express their facial behaviours, there are also subtle differences at the individual level, in how these facial actions are expressed. Since there is lack of person-specific information annotation, how to effectively incorporate person-specific information to the models at test time is a generic and open problem. Inspired by the dynamic facial model described above, this section explores how to encode person-specific video-level facial dynamics from an unlabelled face video of the individual.

The DFR described above achieved excellent performance on dimensional affect recognition (Please see Sec.6.3.2 for details), indicating that if an image-size representation can rank images of a sequence based on their temporal positions, it would be a good descriptor to summarize the temporal evolution of the sequence. Motivated by this, this person-specific dynamic encoding approach is also building upon the same assumption. In particular, this approach starts with pre-training an encoder using emotion data (explained in Sec.4.2.5), as there is evidence (Subramanian et al., 2016; Keltner, 1996) that personality and depression are associated with emotion (Subramanian et al., 2016; Keltner, 1996). Then, self-supervised learning of a dynamic representation of faces that built on DFR is described in Sec. 4.2.4, allowing the network's training to be supervised by the temporal evolution of natural facial actions, i.e., the temporal order of frames, where no manual annotation is required, and consequently the network can learn short-term facial dynamics for each frame, respectively. Based on this self-supervised learning strategy, Sec.4.2.5 proposes to train a Dynamic Facial Neural Network (DFNN) that learns generic facial dynamics, and Sec. 4.2.6 introduces a novel domain adaptation approach to train a set of person-specific adaptation layers (PALs) for each individual. Finally, Sec. 4.2.7 shows how the learned weights of PALs can indeed be used as a basis for personality traits prediction. This pipeline of the person-specific adaptation approach is illustrated in Fig. 4.4.

4.2.1 Novelty and contributions

The proposed self-supervised learning approach builds upon the assumption described in Bilen et al. (Bilen et al., 2017) and the sequence-based DFR model presented in Sec.4.1. In contrast to the sequence-based approach that encodes frame-level DFR from adjacent image sequences, this approach creates a dynamic representation (DR) only from a single face image (the DR mentioned here is different from the DFR described in previous sections). Specifically, based on the rank loss, a UNet-style network: Dynamic Facial

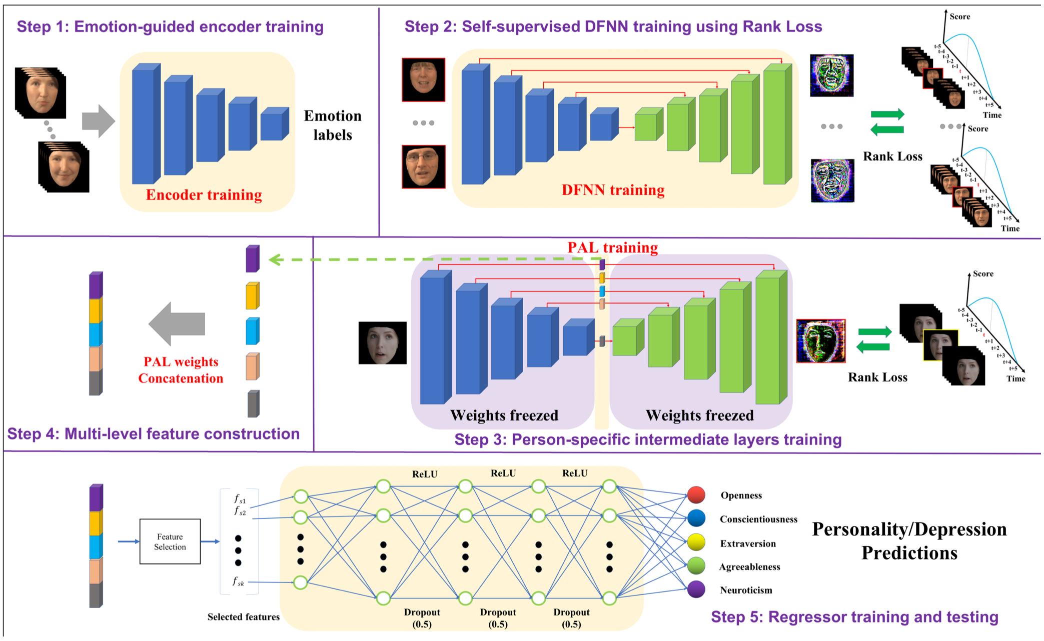


Figure 4.4: The pipeline of person-specific descriptor extraction and personality traits/depression prediction. The approach starts with pre-training an emotion guided encoder (**Step 1**), and then applying the proposed rank loss to train a Dynamic Facial Neural Network (DFNN) (**Step 2**) in a self-supervised manner, which can infer generic facial dynamics of any given face. After that, the DFNN is frozen, and a set of intermediate filters (person-specific adaptation layers (PALS)) are added to skip layers, which are trained using person-specific videos for each individual, respectively (**Step 3**). Then, the learned PALs weights are concatenated as a person-specific facial behaviour representation for the corresponding individual (**Step 4**). Finally, this representation can be fed to a pre-trained regressor (ANN) for personality traits/depression prediction (**Step 5**).

Neural Network (DFNN) can be trained using a large pool of unlabelled face videos. The well-trained DFNN is expected to be able to infer temporal dynamics from face images of any previous unseen individual, i.e., generating a DR that has the same size as the input face image, which can sort adjacent frames of the input face image based on their relative temporal distances to the input face image. Here, the generic facial dynamics is defined as the facial behaviour dynamics that can be observed in most people's face.

Based on this self-supervised learning approach, the proposed domain adaptation approach then incorporates person-specific information to the pre-trained network through adaptation layers. It can be observed that while the proposed Rank Loss allows a network to learn rich generic facial behaviour dynamics in a self-supervised setting, it lacks person-specific information, crucial for the task of personality traits estimation. To incorporate such information, this approach adds a set of **person-specific adaptation layers (PALs)** to the pre-trained DFNN which remains frozen, and resume the self-supervised training with a single video only. Under this setting, the adaptation layers would only carry **person-specific facial dynamics**. After that, contrary to existing approaches that build on top of the pre-trained network, the proposed approach **uses the weights of the PALs as the source of information** for the downstream tasks, in this thesis predicting personality traits/depression severity. PALs have been trained to adapt to the facial behaviours on a person-specific basis, so these weights constitute a person-specific facial dynamic descriptor for a given subject, and their sizes are independent to the length of the video and time-window, reducing the variability of the produced representations. It can be observed that a small fully connected network is capable of delivering promising results in the task of personality prediction from videos using our person-specific weights.

To the best of the author's knowledge, the proposed approach is the first work that brings the advantages of the dynamic image algorithm at summarizing sequences to scenarios where only still images are given. Even though some early works attempted to summarize temporal dynamics from image sequences (Bilen et al., 2017), predicting or anticipating motion from still images (Pintea, van Gemert, & Smeulders, 2014; Walker, Doersch, Gupta, & Hebert, 2016), or generating an output image or sequence according to a target attribute or style (Choi et al., 2018; Isola, Zhu, Zhou, & Efros, 2017; Long, Shelhamer, & Darrell, 2015; T.-C. Wang et al., 2018; Yi, Zhang, Tan, & Gong, 2017; R. Zhang, Isola, & Efros, 2016; Pumarola, Agudo, Martinez, Sanfeliu, & Moreno-Noguer, 2018), none of the existing studies have attempted to infer a DR from a single still face image. Meanwhile, this approach is not only the first work that extends a self-supervised learning approach to automatic personality analysis, but also the first work that uses CNN's weights as a feature descriptor for automatic personality/depression analysis. In comparison to other related studies, this approach are: 1) in contrast to (Güçlütürk et al., 2016; Wei et al., 2018; Subramaniam et al., 2016), the proposed approach is learned in a self-supervised manner without requiring human annotated labels, overcoming the lack of per-frame label problem; 2). the proposed representation is learned from all available frames of a video without losing any short-term details. This differs from (L. Zhang et al., 2019; Subramaniam et al., 2016; Kampman et al., 2018) that use only a small part of each video.

4.2.2 The inputs and targets

Since the goal is to train a network that learns facial dynamics, this section aims to learn a similar dynamic representation (DR). However, instead of learning a separate DR from each given image sequence (as proposed in (Bilen et al., 2017)), this section aims to learn a generic network which can predict the DR for any facial image sequence given a single (central) image of that sequence. This way the network is forced to learn generic temporal evolution of faces within any short sequence. While the DFR could be directly used as target representations for the proposed self-supervised learning task, it can be observed that facial actions display a **symmetric** and **ambiguous** temporal pattern, that could lead to weak representations, as the "predictive" task, understood as ranking preceding frames, would become harder. As shown in Fig. 4.1(a), the temporally reversed counterparts of facial actions are also plausible facial actions. Let's denote an image sequence as $\{I_1, I_2, I_3, I_4, I_5\}$, then its 'temporally reversed counterpart' can be represented as $\{I_5, I_4, I_3, I_2, I_1\}$. Important sources of such ambiguous information are the activation (onset) and deactivation (offset) phases of facial displays. As shown in Fig. 4.1(b), three similar frames can carry completely different dynamics, and thus if one is to define the corresponding facial representation from the preceding frames only, one would end up with completely different descriptors. Having very different descriptors for similar inputs is known to make the learning process harder. In order to partially overcome this limitation, the output DR ranks both preceding and successive frames of the input frame, based on their temporal distance relative to it. The use of temporal distance (regardless of direction) relative to the given image allows the modeling of symmetric patterns while ranking the surrounding frames rather than learning to predict these frames addresses the ambiguity issue in a more efficient way.

Formally, let $I_t \in \mathbb{R}^{m\times n}$ be a given face image, and let $I_{t-T}, I_{t-T+1}, \cdots, I_{t-1}$ and $I_{t+1}, I_{t+2}, \cdots, I_{t+T}$ be the frames corresponding to a window of $2T+1$ frames, centered at I_t. Let $V_a, a \in [t-T, t+T]$ be the static representation for the frame a. This section choose V_a to be the image itself, i.e. $V_a = I_a$. The self-supervised learning task is defined as learning a network $f(\cdot,\, \theta)$, with parameters θ, that produces a dynamic representation (DR) $\mathbf{d_c}$ from just a single input image I_c. The DR is tasked with encoding temporal evolution from the adjacent frames to that of I_c. In particular, $\mathbf{d_c}$ is defined as a representation with the same size as that of V_a, that can rank preceding and succeeding frames based on their relative temporal distance to I_c. The ranking of frames is performed by assigning a score to each, which is defined as the (Frobenius) inner product between the DR $\mathbf{d_c}$ and the representation V_a, with $a \in [c-T, c+T]$. Mathematically speaking, the score for frame $a \in [c-T, c+T]$, assigned by the DR $\mathbf{d_c}$, is defined as $S(\mathbf{d_c}, V_a) = \langle \mathbf{d_c}, V_a \rangle$. The scores are then used in an ordinal manner to sort the frames within the given window. In particular, this approach is interested in assigning ascending scores for frames $a, b < c$, and descending scores for $a, b > c$. This way, the difference between the scores computed at time a and b can be defined as:

$$\delta_{ab}(t) \doteq S(\mathbf{d_c}, V_a) - S(\mathbf{d_c}, V_b) \tag{4.5}$$

where a is chosen to be closer to c than b, with both a and b being either preceding or succeeding frames w.r.t. c. Then, the goal is to learn a DR d_c that makes $\delta_{ab}(c) > 0$ for pairs a, b corresponding to $a, b > c$ or $a, b < c$, with $|a - c| < |b - c|$. As shown in Sec. 4.2.4, only the actual value of $\delta_{ab}(c)$ will be considered when the pair a, b has been incorrectly ranked, i.e. when $\delta_{ab}(c) < 0$. Thus, the target representation can be defined as a d_c that meets the following criteria:

$$\delta_{ab}(c) > 0 \quad \text{for} \quad \begin{cases} |a - c| < |b - c| \\ (a - c)(b - c) > 0 \end{cases} \tag{4.6}$$

It is important to remark that the proposed approach computes the scores for the cases in which both frames are either before the current frame c, or after it. In other words, this approach only interested in computing ascending scores when $a, b < c$, and descending scores when $a, b > c$, as the cases where e.g. $a > c$ and $b < c$ would raise ambiguous definitions.

Discussion: Before defining the target loss function, let us delve into the interpretation of what the DR is enforced to describe. In particular, it is worth noting that the score, defined as a dot product, represents the *similarity* between the generated representation $\mathbf{d_c}$ and the frame descriptor V_a. By enforcing this score to be higher for frames that are temporally closer to the input image I_c, this approach is targeting at a feature representation that is "decreasingly similar" to the features computed at adjacent frames. This way, the learned representation encodes dynamics as a summary of weighted variations. In practical, the length of time-windows has a significant impact on the ranking capablity of learned DRs. This is explicitly evaluated and discussed in Sec. 6.3.2.

4.2.3 Network architecture

In this section, the choice of the network $f(\cdot, \theta)$ is the U-Net network (Ronneberger, Fischer, & Brox, 2015) (The thesis coins this network **Dynamic Facial Neural Network** (**DFNN**)), which is an 5-layer encoder-decoder network with multiple skip layers at different spatial resolutions (see Fig. 4.5). The encoder contains five blocks, where

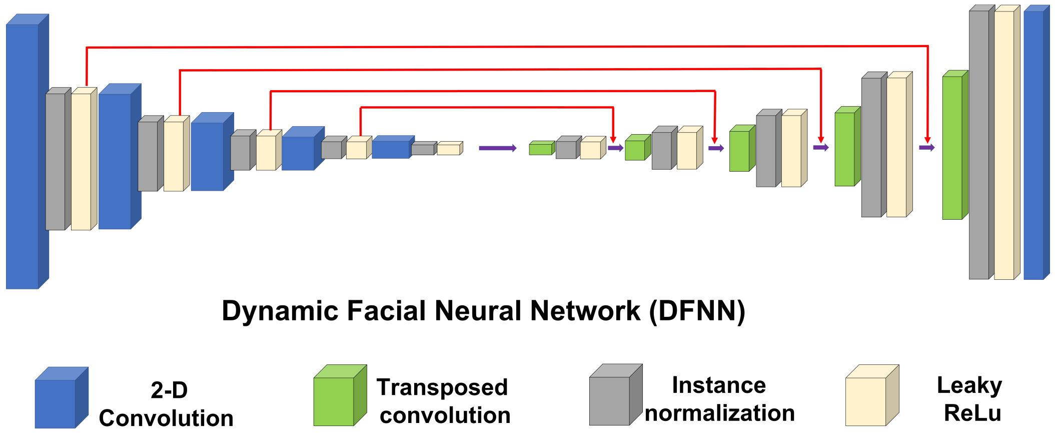


Figure 4.5: The DFNN architecture.

each block is made up of a 2-D convolution layer, an instance normalization layer and a Leaky ReLU activation function. The decoder also contains five blocks. Each block in the decoder contains a transposed convolution layer which first doubles the size of input feature maps. This is then fed to a instance normalization layer and Leaky ReLU. All five pairs of encoder-decoder layers are connected by skip layers.

To infer person-specific facial dynamics, this section also extends DFNN architecture to a DFNN-PALs structure, where a set of intermediate filters are incorporated into the DFNN. In particular, the DFNN-PALs inserts a convolution block consisting of a convolution layer with kernel size 1, a instance normalization layer and a Leaky ReLU to each skip layer of the DFNN. During the person-specific training, the weights of DFNN are frozen and only the weights of the inserted PALs will be adjusted. The number of filters (kernel size of 1) for PALs are set as $32, 64, 128, 256, 512$ in this thesis, respectively. During the person-specific training, the weights of DFNN are frozen and only the weights of the inserted PALs are adjusted. The network structure is shown in Fig.4.6.

4.2.4 Training strategy

The above definition loosely defines the task of the proposed self-supervised learning approach $f(\cdot, \theta)$. The goal is to learn a projection d_t from the input face image I_t as $d_t = f(I_t, \theta)$. A priory, this could be accomplished by defining a target d_t for each input I_t using the original dynamic image algorithm, according to the neighboring frames and

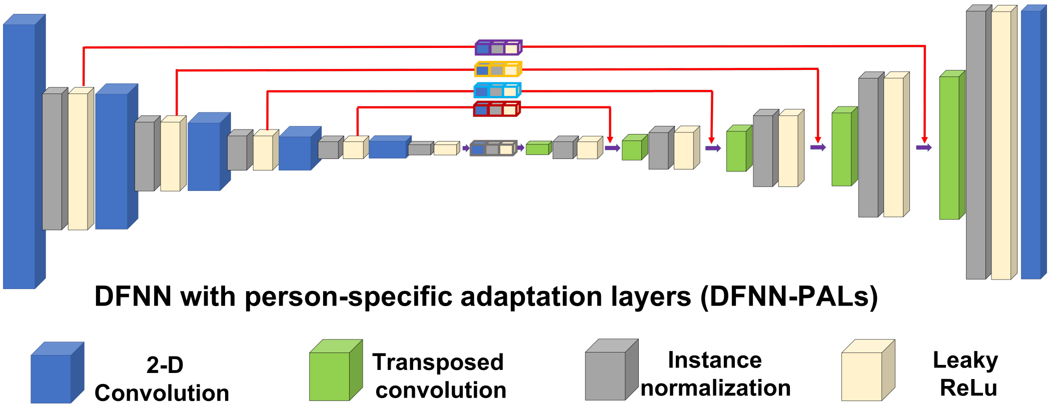


Figure 4.6: The structure of DFNN-PAL.

the aforementioned criteria. In that case, one could use a reconstruction loss between the generated output and the corresponding target, so that the network learns to reproduce such a representation. However, such a loss function would not account for the ranking capabilities of the generated representation, i.e. *subtle errors in the reconstruction loss do not necessarily correlate with errors in the ranking of frames*. This hypothesis has been empirically validated in (Song, Sánchez-Lozano, Shen, et al., 2019) that, when using a pre-defined representation and a reconstruction loss, the task of sorting adjacent frames degrades substantially.

Instead, this section proposes to learn the DR by enforcing the network to produce outputs that directly meet Eqn. 4.6, i.e. this approach allows the network to also help design the DR. More specifically, when the network generates an output from a given face image, the proposed approach projects it onto its preceding and succeeding frames within a window of $N = 2T + 1$ frames with centre I_c, and compute the pair-wise scores using Eqn. 4.5 and Eqn. 4.6. Then the proposed Rank Loss only penalizes the negative scores, i.e. those frames that would be incorrectly sorted based on their scores. To avoid small fluctuations over zero to influence the loss, the rank loss function includes a rank success factor σ. Mathematically speaking, let I_c be the given frame, corresponding to the central image of a window of $N = 2T+1$ frames. Let $\mathbf{d_c} = f(I_c, \theta)$ be the output of the network for the given frame. the DR is generated to minimize the following rank loss function:

$$
\begin{aligned}
L_f(d_c) = \gamma \times \|\mathrm{d_c}\|^2 \\
&+ \sum_{b=c-T}^{c-1} \sum_{a=b+1}^{c} \min(\max(0, \eta - \delta_{ab}(c)), \varepsilon_1) \\
&+ \sum_{a=c}^{c+T-1} \sum_{b=a+1}^{T} \min(\max(0, \eta - \delta_{ab}(c)), \varepsilon_2)
\end{aligned}
\tag{4.7}
$$

where $\delta_{ab}(c)$ is defined in Eq. (4.5). In Eq. (4.7), γ is a regularization factor (set to 1) and η denotes the minimum value that separates the scores between two correctly ranked images. Meanwhile, ε_1 and ε_2 were used to avoid extremely large loss value caused by outliers, e.g., incorrectly detected face regions, which can cause the network to be biased. Specifically, they were used as a relaxation factor to set an upper bound to the rank loss of each single pair. In the experiments of this thesis, the setting is $\varepsilon_1 = \varepsilon_2$, which were ranged from 0.35 to 0.66 based on the size of time-window. The loss $L_f(d_c)$ can be differentiated w.r.t. the parameters of the network f, and therefore the network can be learned through typical backpropagation methods. The training process is also illustrated in Fig. 4.7.

4.2.5 Generic facial dynamics modelling

Since most people express their facial actions in a similar manner (P. Ekman, 1992b), the first goal is to learn a DFNN network that encodes general facial dynamics that can be observed from most individuals. This section proposes to pre-train the encoder, by adding a regression head consisting of some fully connected layers, plugged on top of the encoder, to perform the task of valence and arousal estimation. By pre-training the encoder in such a way, the generated DR is meant to capture the temporal variations on the emotion feature space of related facial behaviours. In other words, as personality traits are found to be reflected by certain emotions (John, 1990; Keltner, 1996; DePaulo, 1992), this pre-training would guide the network to capture personality-related information, rather than other kinds of information such as identity.

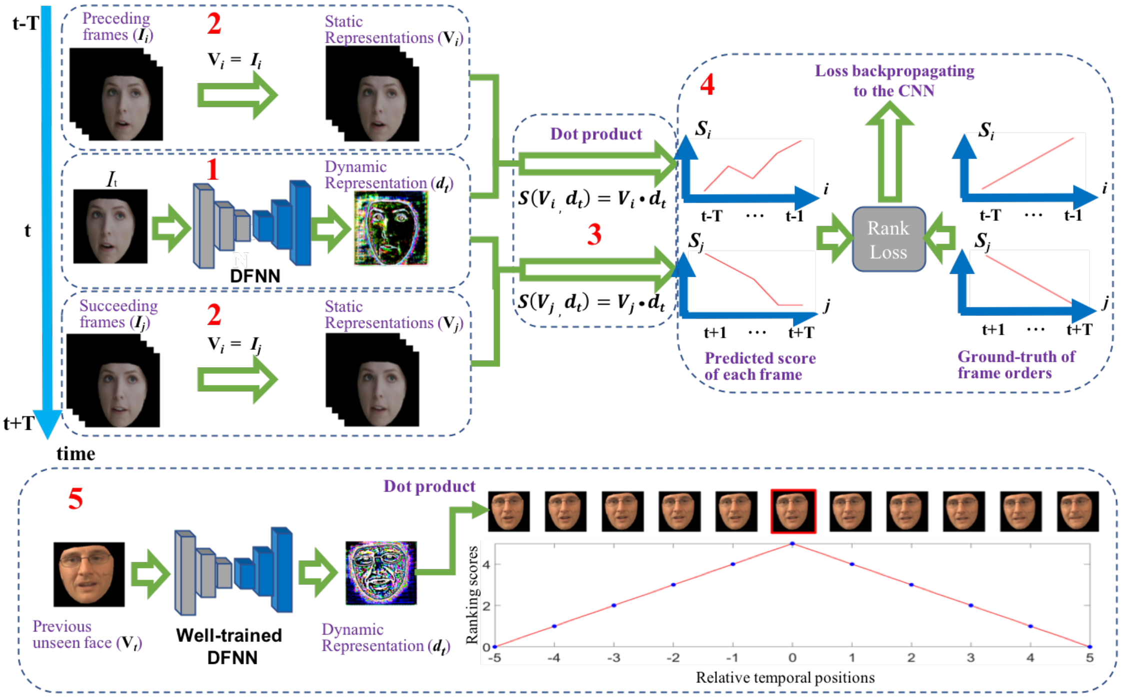

t-T
t
t+T
time
Preceding frames (I_i)
2
$V_i = I_i$
Static Representations (V_i)
I_t
1
DFNN
Dynamic Representation (d_t)
Succeeding frames (I_j)
2
$V_i = I_j$
Static Representations (V_j)
Dot product
$S(V_i, d_t) = V_i \cdot d_t$
3
$S(V_j, d_t) = V_j \cdot d_t$
4
Loss backpropagating to the CNN
S_i
i
t-T ⋯ t-1
S_j
j
t+1 ⋯ t+T
Rank Loss
Predicted score of each frame
Ground-truth of frame orders
5
Previous unseen face (V_t)
Well-trained DFNN
Dynamic Representation (d_t)
Dot product
Ranking scores
0
2
4
-5
-4
-3
-2
-1
0
1
2
3
4
5
Relative temporal positions

Figure 4.7: During training, a set of sequences is given, which are the adjacent frames of a given image I_t. In (1), the given image I_t is forwarded to the network that needs to learn, that produces a DR d_t. The ranking capabilities of d_t can be measured by projecting it onto the preceding and succeeding frames (2). To rank the frames, the difference of each pair-wise scores is computed as a dot product between the generated DR and the corresponding preceding or succeeding frame (3). These scores are used to compute a Rank Loss, which allow us to measure the extent of which the current d_t is correctly ranking the frames within the sequence. The Rank Loss can be backpropagated w.r.t. the parameters of the network that has produced the DR d_t (4). This way, the network not only learns to produce a target dynamic representation d_t, but also contributes to define it. (5) During the inference stage, when an unseen face is used as input, the generated d_t can produce a score for each of its preceding and succeeding frames based on their relative temporal positions to the input frame.

After the pre-training of the encoder, the DFNN is trained with the self-supervised learning approach proposed in Sec.4.2, as a pretext task. The proposed rank loss allows the DFNN learning general facial dynamic patterns from a large pool of unlabelled face videos which can be easily obtained. In particular, the DFNN learns to rank adjacent frames of various facial displays showed by people of different age, gender, ethnic, identity, etc. Thus, it should encode facial action patterns that can be observed in most people regardless of their identities.

4.2.6 Person-specific facial dynamics modelling

While there are underlying similarities in how people express their emotional and cognitive states through facial behaviours, there are also subtle individual differences (discussed in Sec.2.3). In this thesis, it can be observed that, while the ranking accuracy (defined by Equation 6.4 in Sec. 6.2.5) at training time is around 96% (The average results obtained by time-windows equaling 7 and 11 frames, with the stride of 2), the ranking capabilities of the DR at test time is of 87% accuracy on pairwise comparisons. In order to boost this performance for each individual at test time, a set of Person-specific Adaptation Layers (PALs) is attached as the skip layers into the generically trained DFNN network, aiming at encoding person-specific spatio-temporal patterns. This setting allows the PALs to encode dynamics from feature maps at multiple spatial resolutions.

These PALs are trained for each person independently. Departing from the trained generic network, which is kept frozen, only the parameters of the PALs layers are adjusted based on person-specific videos, by optimizing Eq. (4.7). This makes the PALs parameters person-specific, i.e. the parameters θ split into a generic and a person-specific subset, corresponding to the UNet and the PALs, respectively. In other words, let $\theta = \{\theta_g, \theta_{ps}\}$, with θ_g being the weights of the DFNN, resulting after optimizing Eq. (4.7) in a large pool of generic face videos, and θ_{ps} the weights of the extra PALs. Now, the optimization problem consists of optimizing Eq. (4.7) only w.r.t. θ_{ps}, assuming that only a set of person-specific videos are given. Under this setting, it can be observed that the average ranking capabilities of the DR on unseen videos of each specific person goes up to 93%

(the dataset and experimental settings are presented in Sec. 6.1) after adaptation, as shown in Fig. 4.8.

It is clear that DRs generated by the DFNN-PALs network led to the improved performance (from 87% to 93%) in ranking frames of previous unseen videos for each specific individual, but reduced the performance for other individuals. This section hypothesizes the reason for the enhanced performance in inferring dynamics for the specific person is that the adjusted PALs can better encode person-specific facial dynamics of the given person, in addition to the general facial dynamics provided by DFNN. Additionally, while the facial displays varied a lot for each video, the PALs still helped the DFNN to accurately infer facial dynamics from most frames (93% ranking accuracy). These indicate that the dynamics encoded in PALs are stable over various facial displays of the given person. Meanwhile, based on the decreased ranking performance on other individuals, this section also hypothesises that the facial dynamics encoded in PALs rarely occur in others due to the subtle differences in facial dynamics at individual level. As a result, the PALs is expected to encode person-specific facial dynamics that is relatively stable over time for a given person but differ compared to others.

4.2.7 Person-specific representation

As we can see, this approach offers a person-specific representation that is independent of video length: that of the PAL layers. In particular, **this section proposes to use the learned weights and bias θ_{ps} as the representation** for the downstream tasks. Because the weights and bias θ_{ps} have been specifically trained by the videos of the given person, they are person-specific and constitute a fixed-size representation regardless of the length of the video. In this thesis, each produced person-specific representation used in this thesis is made up of the filter weights of 5 PALs, whose number are $32, 64, 128, 256$, and 512, respectively, making the produced representation 1984 dimensional (992 weights and 992 biases).

Since the temporal scale of the person-specific representation is decided by the size of time-window but the optimum time-scale for personality analysis is still unclear, this

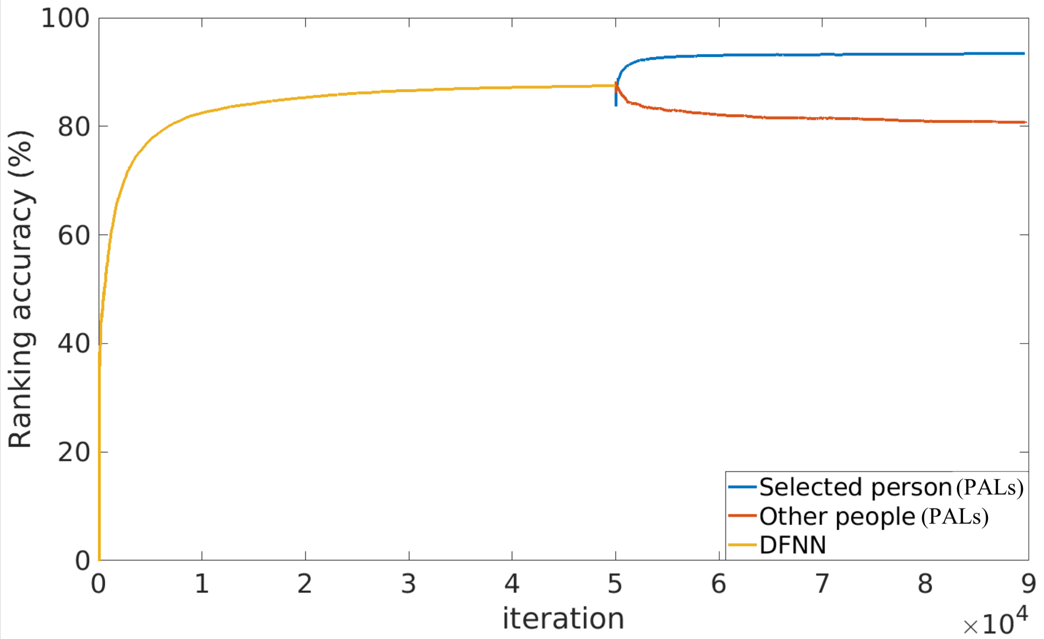


Figure 4.8: The average ranking accuracy curve of DFNN and DFNN-PALs. The DFNN is trained by 50000 iterations, where the ranking accuracy of DRs generated by DFNN in validation set is finally stable at about 87%. Then, the weights of DFNN is frozen, and train PALs for another 50000 iterations. It is clear that the PALs lead the enhanced ability of inferring facial dynamics from the given person (93% ranking accuracy) but the reduced ability of inferring other people's facial dynamics (80% ranking accuracy).

section proposes to construct multi-scale person-specific descriptors, each capturing facial dynamics at a unique temporal scale. This can be achieved by using a set of time-windows with different lengths, to train multiple sets of PALs. As a result, the combination of these person-specific descriptors represent facial dynamics obtained at multiple temporal and spatial resolutions, for a given person.

4.3 Applications: personality traits estimation and dimensional affect estimation

4.3.1 Facial affect analysis using multi-scale DR

Given that the generated DRs are 3-channel raster images displaying a "summary" of possible facial motions, they can be feed to most state-of-the-art Machine Learning frameworks for any still image-based facial analysis tasks. To evaluate the usefulness of the proposed still image-based approach in encoding multi-scale facial dynamics, this section further proposes to generate multiple sets of DRs, each capturing a different temporal scale by using a different window length to train CNN models. It should be noted that each generated DR d_c is a 3-channel tensor with its size equaling to the given face image I_c, regardless of the choice of the time-window size. Thus, they can be easily concatenated before applying them to further related tasks. Herein, this section will explore the use of a **Single Dynamic Representation** (**SDR**), using just the generated DR, and the use of a **Multi-level Dynamic Representation (MDR)**, which combines the output of networks trained using different time-windows, as a multi-channel matrix representing multi-scale temporal dynamics. In particular, a multi-level representation is used for the dimensional affect estimation experiments presented in Sec.6.2.5. The first level corresponds to $T = 0$, i.e., a window length of one frame. In practice, this level does not require the training of a DR, as it basically consists of the input image. Then, the number of rest channels and the temporal scale of rest channels are decided by the validation process to capture the most valuable temporal dynamics along with the static appearance. Fig. 4.9 shows a description of the MDR for this three-level approach.

4.3.2 Personality traits prediction

This main goal of this chapter is to propose and apply the person-specific facial dynamic encoding approach to video-based automatic personality analysis tasks. To this end, an Artificial Neural Network (ANN) structure used in (Jaiswal, Song, & Valstar, 2019) is extended here as the regressor. Fig. 4.10 shows the topology of the ANN used in this section. It consists of 4 fully-connected hidden layers, where each of them is equipped with a dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) with probability 0.5, to prevent model from overfitting. The last layer has 5 output units, each corresponding to one of the Big-Five personality traits. This architecture enables the model to jointly learn all the five personality traits.

Before feeding the produced person-specific representations to the personality recognition model (ANN is employed in this thesis), their dimensions were reduced using

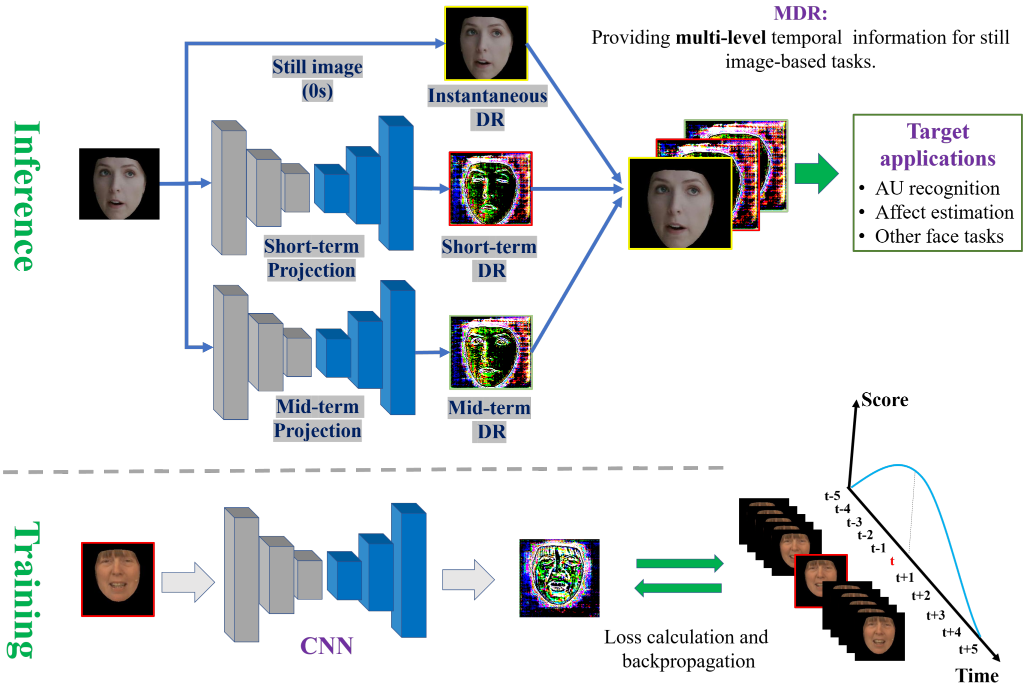


Figure 4.9: The proposed MDR approach (top) that can be used to infer several time-length dynamics, and further combined to enhance face-related tasks such as dimensional affect estimation. During training (bottom), a set of videos is used to learn the DRs, without explicitly generating a fixed set of target representations. In this thesis, the short-term refer to the time duration that is less than 1.5 seconds and mid-term refers the time duration ranges from 1.5 seconds to 5 seconds.

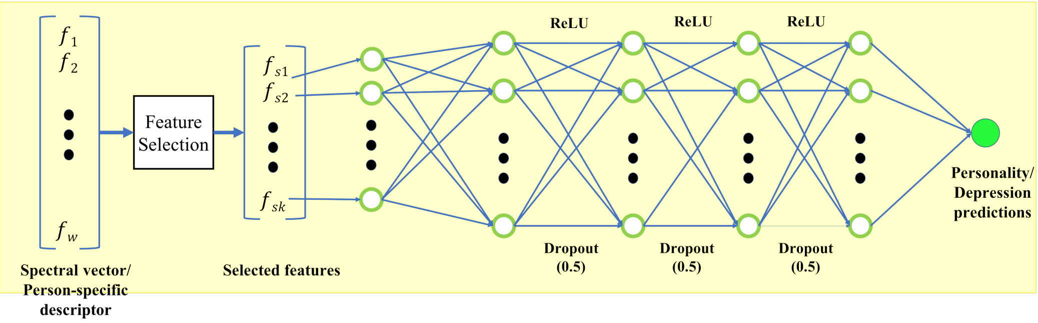


Figure 4.10: The ANN model used to learn spectral vectors and person-specific representations.

Correlation-based Feature Selection (CFS) (M. A. Hall, 1999). The CFS algorithm hypothesizes that a good feature set contains features that are highly correlated with the target, yet uncorrelated with each other. In this thesis, CFS is employed to select a subset of the most important person-specific features from produced person-specific representations. This process was done on training sets, where Pearson's linear correlation coefficient was employed to measure the correlations between each feature and the target.

4.4 Summary

The person-specific approach proposed in this chapter is the first work that utilises CNN parameters as the feature representation for personality analysis. Its novelties and advantages have been summarized in Sec. 4.1.1, while the main drawback of this approach is time-consuming, i.e., it has to re-train a set of intermediate layers for each test video. Consequently, this approach is not efficient for real-world applications considering the limitation of current computation resources. In addition, as shown in Fig. 6.3 and Fig. 6.5, this approach is only good at encoding short-term dynamics. Thus, the next chapter will propose a spectral approach to encode multi-scale long-term facial dynamics, which also encodes facial dynamics more efficiently. The evaluation of the person-specific approach and DFR models are presented in Chapter 1.

Chapter 5

Modelling multi-scale facial dynamics

While the person-specific approach encodes person-specific short-term facial dynamics, it lacks the ability to encode long-term multi-scale facial dynamics, which are also crucial for understanding one's personality traits and depression levels. Let us recall the main problems of existing long-term encoding approaches that used in related studies: 1. statistical features and GMM-based approaches (Bekhouche et al., 2017; Fang et al., 2016) usually failed to retain facial dynamics between adjacent frames; 2. memory-based models such as Recurrent Neural Networks (RNNs) are easy to be overfitting. These approaches also require frame/segment-level labels for model training, which regard video-level labels as the frame/segment-level labels (Güçlütürk et al., 2016; Wei et al., 2018; Subramaniam et al., 2016). As discussed before, these training strategies would end up utilising the same input pattern with multiple labels, making it practically impossible to train models that have good generalization capability; 3. other approaches (L. Zhang et al., 2019; Subramaniam et al., 2016; Kampman et al., 2018) either select a set of key frames to represent an entire video, or attempted to only used a short segment of a video to represent the global information. They ignore the information contained in the discarded frames.

To addresses problems summarized above, this chapter proposes a spectral analysis approach that provides multi-scale video-level facial dynamics for automatic personality

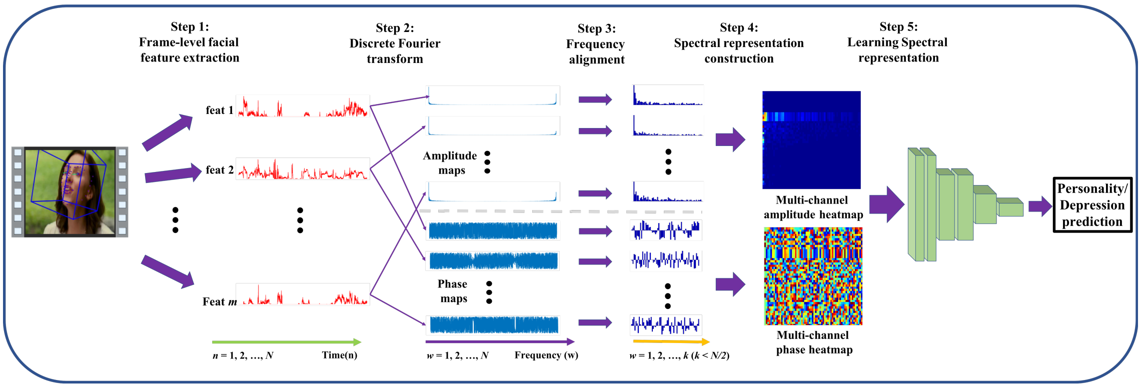

Step 1:
Frame-level facial feature extraction
Step 2:
Discrete Fourier transform
Step 3:
Frequency alignment
Step 4:
Spectral representation construction
Step 5:
Learning Spectral representation
feat 1
feat 2
Feat m
Amplitude maps
Phase maps
Multi-channel amplitude heatmap
Multi-channel phase heatmap
Personality/ Depression prediction
n = 1, 2, …, N
Time(n)
w = 1, 2, …, N
Frequency (w)
w = 1, 2, …, k (k < N/2)

Figure 5.1: The pipeline of the spectral approach. This approach starts with using low dimensional multi-channel human behaviour time-series data (e.g. latent descriptor, behaviour primitives, etc.) to represent videos (**Step 1**), and then converts them to spectral signal consisting of multiple frequency information of all frames (**Step 2**). Since spectral signals are symmetric, only the first half of them are kept.Then, the frequency alignment is implemented by removing high frequency components and obtaining common frequencies for human behaviours of all videos (**Step 3**). Finally, the aligned spectral signals of all dimensions (**Step 4**) are concatenated and are feed to ML models for personality/depression analysis (**Step 5**).

traits and depression analysis. It first concatenates per-frame descriptors to produce a multi-channel time-series signal describing the visual facial expressive behaviours across a video. Then, this approach employs two frequency alignment methods to create spectral representations of equal size and frequency coverage, regardless of variation in the length of input videos. The final generated long-term spectral representations contain multiple video-level facial behaviour frequencies, representing multi-scale video-level facial temporal information in the frequency domain, where each frequency component stands for a unique scale of dynamics. Meanwhile, the lengths of interview videos are usually variable, with the duration of the longest video sometimes several times longer than the shortest one. Yet most Machine Learning models require fixed-size input. To this end, the proposed spectral approach also aims to encode information of variable length videos into fixed-size video representations. In this Chapter, $f_c^m(n)$ is defined as the c_{th} time-series signal in m_{th} video. The pipeline of this approach is shown in Fig. 5.1.

5.1 Novelty and contributions

Compared to other recent methods, the main advantages and novelties of the spectral approach are: 1). it can convert long and variable length time-series data to a small and fixed-size representation, allowing the long-term information of the entire video to be used for analysis, which differs from (Güçlütürk et al., 2016; L. Yang et al., 2018; Biel et al., 2012; Kalimeri et al., 2010; Wei et al., 2018) in the sense that they can only process fixed-length segments from videos for analysis; 2). the spectral representation contains multi-scale video-level temporal information, in contrast to (Joshi et al., 2014; Aran & Gatica-Perez, 2013b; Nguyen et al., 2013; Teijeiro-Mosquera et al., 2015) that did not retain temporal information, or (Jaiswal & Valstar, 2016; Stepanov et al., 2018; L. He et al., 2018; Anis et al., 2018) where only a fixed-length time window is used to encode a single-scale short-term dynamic, thereby losing temporal information at other scales. In summary, this approach converts long and variable length time-series data, (the multi-channel time-series signal that is made up of the per-frame facial descriptors), to short and length-independent spectral representations that contain multi-scale temporal dynamics, which can be easily used with standard Machine Learning techniques.

5.2 Encoding multi-scale video-level dynamics

Given that the difference in personality/depression status causes people to have different expressive behaviours in videos when facing a set of similar stimuli, which can be represented by time-series signals, both temporal patterns and spatial patterns are significant. Additionally, both personality and depression status are expected to be relatively stable in a interview time scale. Thus, this approach aims to extract representations that reflect video-level information including multi-scale temporal facial dynamics.

The proposed approach firstly uses the Fourier Transform (FT) to convert time-series signals of a video to the frequency domain. The resulting spectral representation is a decomposition of the original time-series signal into its constituent frequencies. Let $f(x)$ be a time-series signal, then the Fourier Transform can convert it to a spectral representation $F(w)$

$$F(w) = \int_{-\infty}^{\infty} f(x) e^{-(2\pi ixw)/N} dx \tag{5.1}$$

where w can be any real number and $F(w)$ is a complex function that can be re-written as

$$\begin{aligned} F(w) &= \int_{-\infty}^{\infty} f(x)(\cos((2\pi ixw)/N) - i\sin((2\pi ixw)/N))dx \\ &= \int_{-\infty}^{\infty} (\mathrm{Re}(fc(x)) + i\mathrm{Im}(fs(x))) \\ &= \mathrm{Re}(F(w)) + i\mathrm{Im}(F(w)) \end{aligned} \tag{5.2}$$

where $fc(x)$ and $fs(x)$ denote $f(x)\cos((2\pi ixw)/N)$ and $-f(x)\sin((2\pi ixw)/N)$, respectively. $R(F(w))$ is the real part of $F(w)$ and $Im(F(w))$ is the corresponding imaginary part of $F(w)$. Here, w determines the frequency $(2\pi w)/N$ that $F(w)$ represents. Consequently, the spectral representation $F(w), w \in [-\infty, \infty]$ contains information from all frequencies presented in $f(x)$.

In the applications of personality and depression analysis, each video is made up of a set of frames, resulting in one discrete time-series signal for each frame-level feature. Therefore Discrete Fourier Transform (DFT) is applied to the feature signal $f_c(n)$, where $c = 1, 2, \cdots, C$ denotes the feature index and $n = 1, 2, \cdots, N$ denotes the frame index, as given below:

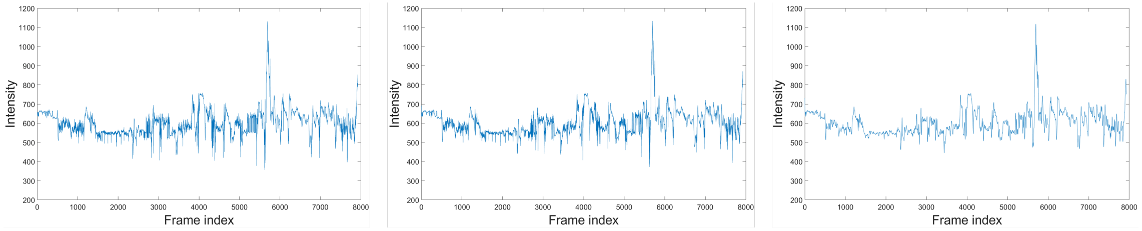
$$\begin{aligned} F_c(w) &= \sum_{n=0}^{N-1} f_c(n) e^{-\frac{2\pi i}{N} wn} \\ &= \sum_{n=0}^{N-1} f_c(n)[\cos(2\pi wn/N) - i\sin(2\pi wn/N)] \\ &= \sum_{n=0}^{N-1} (\mathrm{Re}(fc_c(n)) + i\mathrm{Im}(fs_c(n))) \\ &= \mathrm{Re}(F_c(w)) + i\mathrm{Im}(F_c(w)) \end{aligned} \tag{5.3}$$

where $f_c(n)$ is the time-series signal of c_{th} feature, which consists of N frames and $F_c(w)$ is the DFT of the signal $f_c(n)$ at frequency w, where $w = 0, 1, 2 \cdots, W-1$.

As can be seen from Eqn. 5.3, each frequency component is computed from all frames of the $f_c(n)$. This is to say, each component in the spectral signal summarizes a unique frequency information presented in the entire video. Therefore, the spectral signal contains information corresponding to W frequencies given by $2\pi w/N, w = 0, 1, 2, \cdots W-1$. These components represent different types of facial dynamics, i.e. high frequency components represent sharp facial behavioural changes and low frequency components represent more gradual changes in facial behaviour. As the relationship between the duration of a $T_c(w)$ and its frequency $F_c(w)$ is $T_c(w) = 1/F_c(w)$, the produced spectral signal can be said to summarize multi-scale temporal information of the entire video. Here, the number of discrete frequency components W in $F_c(w)$ is set to be the same as N in order to completely summarize the information contained in the discrete time-series data $f_c(n)$ (It is well known that the $f_c(n)$ can be fully reconstructed from $F_c(w)$ if $W = N$).

5.3 Frequency alignment

Motivation: As mentioned above, a time-series signal consisting of N frames can be converted to a spectral signal that has $W = N$ frequency components without losing any information. Thus, the spectral signals of variable-length videos will have different number of frequency components, which would again lead to feature representations of varying dimensionality. To make them equal, it should be noted that the spectral signals of time-series data are always symmetric around their central frequency $W/2$, i.e. if $F(w) = Re(w) + iIm(w)$ and $F(W-w) = Re(W-w) + iIm(W-w)$, then $Re(w) = Re(W-w), Im(w) = -Im(W-w)$. This means that the first $W/2$ components of the spectral signal can fully represent the information contained in $f(n)$. Additionally, as facial actions are continuous and smooth processes, high-frequency information usually represents noise or outliers caused by e.g. incorrectly detected faces, errors in facial points localization, etc. In practice, after removing the high-frequency information, the reduced spectral signal can still represent the original time-series data well, as applying the inverse DFT to the modified spectral signals can recover most of the information present in the original time-series data. This is illustrated in Fig. 5.2 and Fig. 5.3. In both figures, all unused frequency components were replaced by zeros.



Original signal

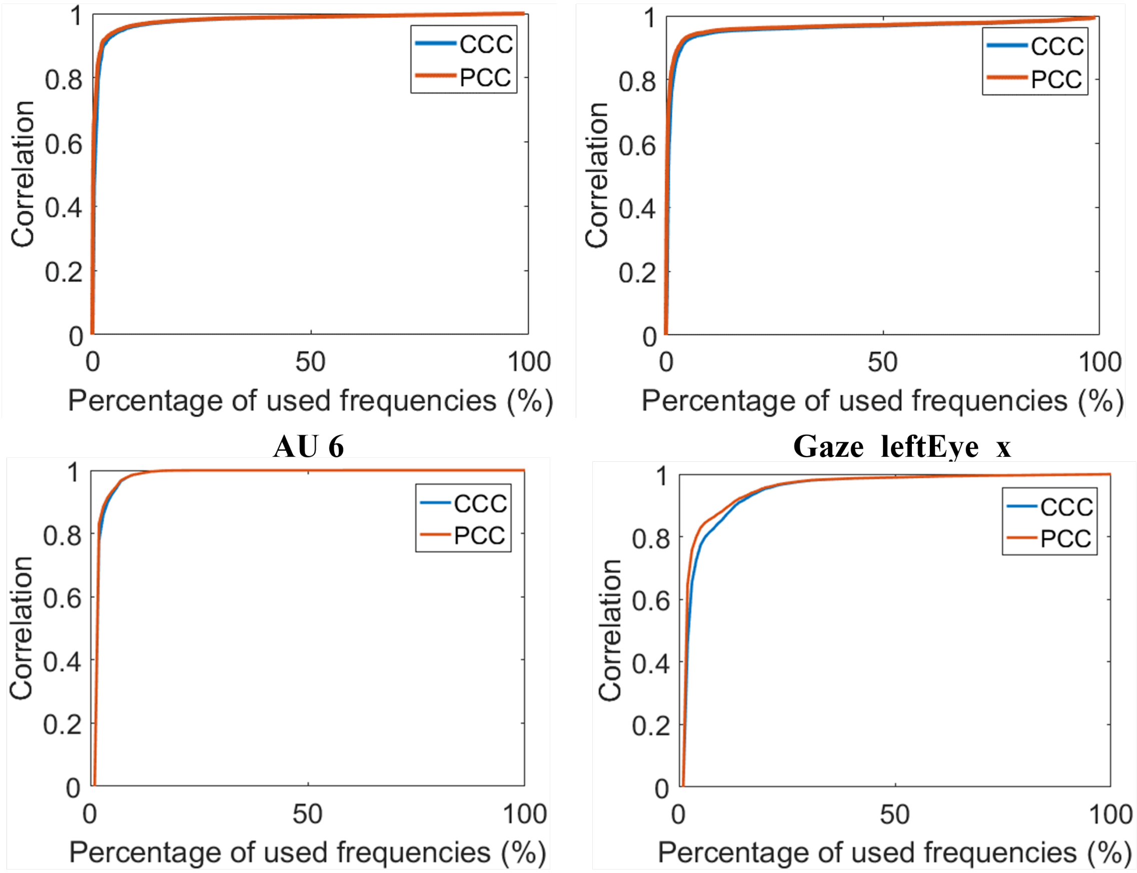
Remove 3000 highest frequencies

Remove 7000 highest frequencies

Figure 5.2: Examples of the reconstructed time-series signals after removing high frequency components. The original signal has 7923 frames and its spectral signal also has 7923 frequencies.

Solutions: Motivated by this, the proposed approach only keeps the first $W/2$ components of spectral signals. Then, components corresponding to high frequencies are also removed. Since the goal is to generate video-level spectral representation of the equal size from the variable length time-series data, one may consider to keep the first K lowest frequencies of spectral signals for all videos, with $K < W/2$. However, the w_{th} component in videos of different lengths will represent different frequencies. Consider two time-series signals $f^1(n)$ and $f^2(n)$ of length N_1 and N_2 respectively. Additionally, consider their corresponding spectral representations as $F^1(w)$ and $F^2(w)$ respectively. If $N_1 \neq N_2$, the w_{th} component $(0 < w < N_1/2, N_2/2)$ of the spectral signal $F^1(w)$ denotes the DFT value at frequency $2\pi w/N_1$ while the w_{th} component of $F^2(w)$ denotes the DFT value at frequency $2\pi w/N_2$. Clearly, $2\pi w/N_1 \neq 2\pi w/N_2$, and thus the w_{th} component of spectral signals $F^1(w)$ and $F^2(w)$ do not represent the same frequency. In order to resolve the problem of misaligned frequencies, this section proposes the following two solutions:

Solution 1: Zero-padding is a common method often used to increase the frequency resolution after Fourier Transformation of a discreet time series. In this method, zeros are appended to the time-series data to increase its length, allowing the DFT of this time-series data to have more frequency components. In particular, the frequency resolution W of the spectral signal is equal to the number of frames N in the original time-series data. By padding with zeros, N_{add} zeros are added at the end of the original time-series signal to create a new time-series of length $N + N_{add}$. Consequently, the spectral signal of the new time-series signal will have $W + N_{add}$ frequency components. Please see (Lindsten,



AU 6

Gaze leftEye x

Pose_rotation_z

Gaze_leftEye_y

Figure 5.3: Examples of the average correlation between reconstructed behaviour primitive time-series signals and original behaviour signals as a function of the percentage of used frequencies. It can be observed that even after removing more than 90% of high frequency components, the reconstructed signals still have high correlation with the original signals. (The Pose_rotation_z is the least useful behaviour primitive for depression analysis; the Gaze_leftEye_y achieved lowest correlation (CCC) with the original signals when removing 90% high frequency components)

2010) for detailed theoretical explanation of this method. For the task conducted in this thesis, the proposed approach first set a number that is not smaller than the longest video in both training and testing sets (or potential test videos). Then, zeros are added to the time-series signals extracted from the rest of the videos, making time-series signals of all videos to have the same length as the longest video. Consequently, the spectral signals of all zero-padded time-series signals will have the same resolution. By further selecting only the first K components of each spectral signal, the dimensionality can be significantly reduced.

Solution 2: Although zero-padding can increase the frequency resolution of spectral signals, the values of the increased frequency components are estimated. Moreover, the multi-channel facial behaviour time-series signals added by zero-padding are zero-signal. This strategy assumes that the facial status in the added frames is neutral and remains unchanged, which is not accurate. Therefore, the extended multi-channel time-series signal may not accurately represent the facial behaviour patterns of the corresponding person. To avoid this, the second solution extracts fixed-size spectral signals from variable length time-series data by choosing k common frequencies from the spectral signals obtained from each video. In this case, the values of k chosen frequencies are obtained from the original signal rather than an zero-padded signal. Hence, each component in the produced representation represents the accurate value rather than the estimated value of the corresponding frequency. It should be noted that the advantage of this method is at the cost that the spectral signals gets downsampled thereby losing some information. Assuming that there are M time-series signals $f^1, f^2, \cdots f^M$ corresponding to M variable length videos, this solution can be achieved by the following steps:

1. Choosing a fixed frequency resolution R, i.e. the number of frequency components used to represent each time-series data, and then shorten the time-series, reducing the total number of frames in the original time-series signal $f^m(n)$ from N_m to $N_m - (N_m \bmod R)$ frames, which is a multiple of R, resulting in a slightly shorter time-series signal $S(f^m(n))$. In practice, the first $(N_m \bmod R)/2$ frames and the last $(N_m \bmod R)/2$ frames are removed from each video. For example, if R is chosen as

100 or 500, which means the maximum length of removed video contents were less than 99 frames or 499 frames, respectively.

2. Each time-series $S(f^m(n))$ is converted to its spectral signal $S(F^m(w))$ using Eq 5.3. Since the number of frequency components is equal to the number of frames, the number of frequency components in $S(F^m(w))$ will also be multiple of R, which can be defined as $W_m = (t_m \times R), m = 1, 2, \cdots, M$. As a result, the frequencies represented in each spectral signal can be denoted as $2\pi w_m/(t_m \times R), w_m = 0, 1, 2, \cdots, t_m \times (R-1)$.

3. As the number of frequencies in each spectral signal is a multiple of R, all of them would contain the same R components whose frequencies are given by:

$$\begin{aligned} n_f(m) &= 2\pi w_m(r)/W_m \\ &= 2\pi r \times t_m/(R \times t_m) \\ &= 2\pi r/R \end{aligned} \tag{5.4}$$

   where $r = 0, 1, 2, \cdots (R-1)$. It is clear that the R selected frequencies are independent of t_m, and these R frequencies, i.e. $2\pi \times 0/R, 2\pi \times 1/R, 2\pi \times 2/R, \cdots, 2\pi \times (R-1)/R$, are encoded in all spectral signals. This process is also illustrated in Fig.5.4. Finally, the proposed approach removes those high frequency components and only keeps the first K components.

As a result, the solution 2 could not only align the frequencies of variable-length time-series signals, but also prevents the distortion of the aligned spectral signals.

5.4 Spectrum representations

After obtaining the aligned spectral signal for each time-series signal, this section also describes two methods to construct a fixed-size joint representation so that spectral signals of all features (all behaviour primitives) can easily be used as the single input for standard ML techniques.

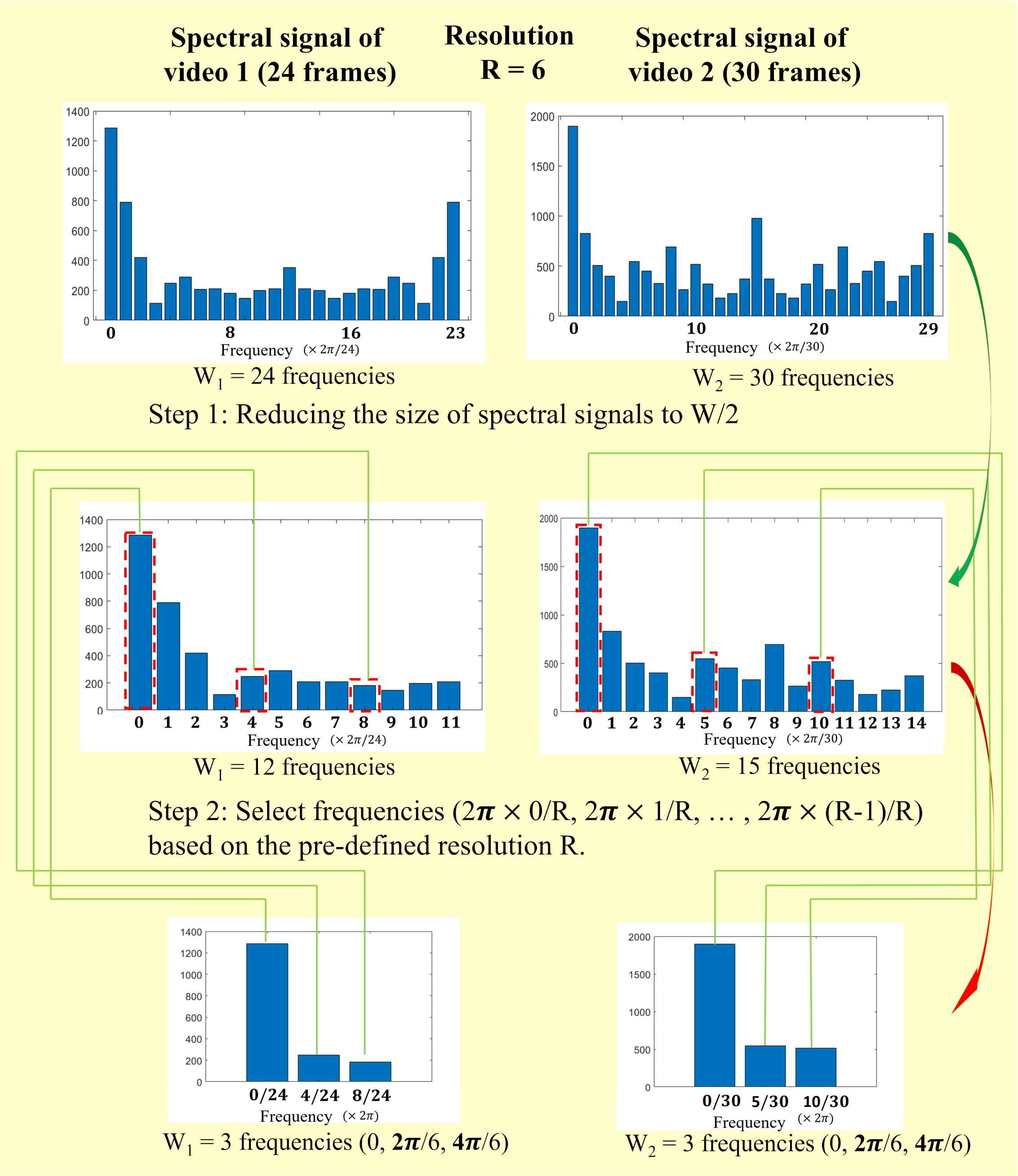


Figure 5.4: Illustration of the frequency selection (Step 3 of the solution 2): After the DFT, the second half of the spectral signals are removed as they are symmetric.

Assuming that the extracted per frame behaviour descriptor has C dimensions, the proposed approach produces C aligned spectral signals consisting of K frequencies for each video. Since the values in spectral signals are complex numbers, each of them is converted to two spectrum maps in the real domain: an amplitude map and a phase map, where the amplitude map can be computed by

$$|F_c^m(w)|/N = \sqrt{\mathrm{Re}_c^m(w)^2 + \mathrm{Im}_c^m(w)^2}/N \tag{5.5}$$

and the phase map can be computed by

$$\arg(F_c^m(w)) = \arctan\frac{\mathrm{Im}_c^m(w)}{\mathrm{Re}_c^m(w)} \tag{5.6}$$

where $\mathrm{Re}_c^m(w)$ and $\mathrm{Im}_c^m(w)$ are the real and imaginary part of $F_c^m(w)$ respectively, as defined in Equation(5.3). Hence, C amplitude maps and C phase maps are extracted from a video, all of which have K frequencies. Then, the following two methods are further proposed to combine them:

1. **Spectral heatmap**. A $C \times K$ multi-channel amplitude spectrum map and a $C \times K$ multi-channel phase spectrum map. In both maps, each row represents an amplitude map or a phase map of a single feature signal while each column represents a unique frequency. In this thesis, two spectrum maps are concatenated as a two-channel spectral heatmap.

2. **Spectral vector**. A 1-D vector that concatenates $C \times K$ amplitude features and $C \times K$ phase features from all time-series signals. As a consequence, the concatenated vector contains $C \times K \times 2$ components.

It is clear that both representations encode information from descriptors of all frames for all dimensions. Additionally, the length-independent property of the produced representations makes them suitable for use with standard ML techniques.

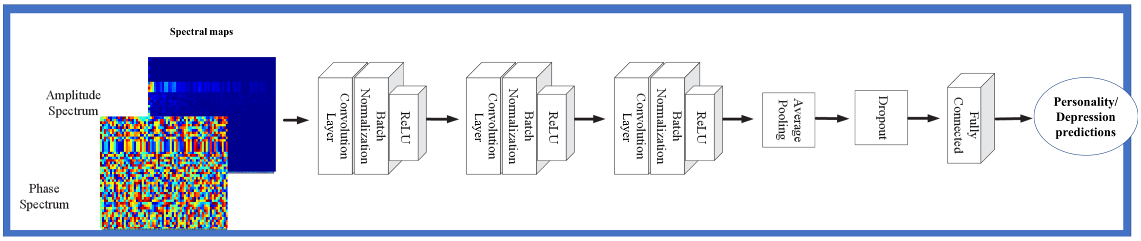


Figure 5.5: The 1-D CNN model used to learn spectral heatmaps.

5.5 Personality traits recognition models

Inspired by the success of applying deep learning work (Z. Wang, Yan, & Oates, 2017; Liu, Reimer, Song, Mehler, & Solovey, 2021) for multi-channel signal processing, a 1-D CNN structure that has been frequently used in the multi-channel time-series data analysis, is employed to extract features from **spectral heatmaps**. The reason behind this choice is that the rows in the heatmaps, which represent multiple facial features, have no natural ordering, spatial or otherwise. Therefore, standard 2-D CNNs may not be suitable. Hence, the proposed spectral heatmaps are treated as multi-channel 1-D data and a 1-D CNN is employed as the personality traits estimation back-end model. As shown in Fig.5.5, this CNN architecture is made up of three Conv-Batch-ReLU blocks, where each block contains a 1-D convolution layer followed by a batch normalization layer and a ReLU layer. In particular, each convolution layers consists of 128 filters of kernel size 7×1, 128 filters of kernel size 5×1, and 64 filters of kernel size 3×1, respectively. After that, a channel-level global average pooling layer is employed to obtain a feature of a single value from each feature map, producing a 64-D deep feature. Finally, a fully connected layer with 64 input neurons, a dropout layer (Srivastava et al., 2014) (probability factor $p = 0.5$) and an output layer of five neurons are used at the top of the average pooling layer to jointly predict five personality traits. Meanwhile, the ANN architecture introduced in Sec.4.3 is again employed to process spectral vector, allowing five personality traits can be jointly learned.

5.6 Summary

The spectral approach proposed in this section first employs Fourier Transform to convert time-series behavioural signals to the frequency domain as spectral signals, where each component in a spectral signal encodes different frequency information of the whole video. As a result, the produced spectral signals contain multi-scale video-level temporal information. However, due to the variation in the length of original videos, the length of their corresponding time-series behaviour signals and spectral signals are also variable. To allow spectral signals to be easily processed by standard ML models, two frequency alignment methods are also proposed. Additionally, two spectral representations, i.e., spectral heatmap and spectral vector, are employed to encode aligned spectral signals, allowing them to be learned by CNNs and ANNs, respectively.

In comparison to the person-specific representation encoding approach, the spectral approach theoretically has clear advantages in encoding multi-scale and long-term dynamics. In addition, it does not require training a set of layers for each test video, making it more suitable for real-world applications. However, it can only learn a representation from per-frame 1D features instead of the original video data, which means its performance is largely depending on the quality of the extracted features.

Chapter 6

Evaluation of the person-specific representation and spectral representation

This chapter evaluates the proposed person-specific dynamic encoding approach (presented in Chapter 4) and multi-scale facial dynamics encoding approach (presented in Chapter 5) on automatic true/apparent personality traits analysis tasks. Sec. 6.1 introduces the datasets used for not only personality studies but also for frame ranking and dimensional affect estimation tasks. Then, implementation details of two approaches and the evaluation metrics are described in Sec. 6.2 and Sec. 6.2.5, respectively. The performance of both spectral and person-specific approaches on personality traits estimation are compared with the existing state-of-the-art approaches in Sec.6.3.1. After that, a series of ablation studies are conducted to evaluate various aspects of the proposed two approaches. In particular, experiments on frame ranking (Sec. 6.3.2) and dimensional affect estimation (Sec. 6.3.2) tasks are firstly introduced to evaluate the performance of the proposed rank loss-based self-supervised learning strategy (used by the person-specific dynamic encoding approach) in modelling short-term facial dynamics. Then, the influence of different settings on two approaches are also evaluated in Sec. 6.3.2 and Sec. 6.3.3, respectively.

6.1 Datasets

6.1.1 Frame ranking datasets

Three datasets were used for the frame ranking experiments. The RECOLA dataset (Ringeval, Sonderegger, Sauer, & Lalanne, 2013) is solely used to train the DFNN based on the proposed self-supervised learning algorithm. Then, SEMAINE and BP4D datasets were used to evaluate the capabilities of the trained DFNN, by testing the ranking accuracy of the generated DRs.

The used subset of the RECOLA dataset contains 27 videos corresponding to the AVEC 2016 challenge (M. Valstar et al., 2016), where each video is approximately 5 minutes of 25 fps. During each recording, a pair of participants are communicating over a video conference to conduct a collaborative task. These videos were recorded from 16 female participants and 11 male participants. Among them, there are 20 Frenches, 5 Italians and 2 Germans.

The SEMAINE dataset recorded uncontrolled facial expressions of participants who have a conversation with an operator. All videos in this dataset were recorded with the frame rate of ~ 50 fps. For experiments conducted in this thesis, a subset of SEMAINE dataset that has been used by AVEC 2012 challenge (Schuller et al., 2012) is employed. This subset contains 31 videos for training ($501,277$ frames), 32 videos for development ($449,074$ frames) and 32 videos for the test ($407,772$ frames). The scenario used in the recordings is the Sensitive Artificial Listener (SAL) technique (Douglas-Cowie, Cowie, Cox, Amir, & Heylen, 2008), where each participant was asked to interact with emotionally stereotyped “characters” whose responses are stock phrases keyed to the user’s emotional state. In this subset, all frames have been annotated with valence and arousal intensities, each lying in the range $[-1, 1]$.

This thesis uses a subset of BP4D dataset, which is defined by the FERA 2015 challenge (M. F. Valstar et al., 2015). This subset contains $75,586$ frames for training, $71,260$ frames for development and $75,726$ frames in the test set. During the video recording, young adults were asked to respond to emotion elicitation tasks (happiness/amusement, sadness, surprise/startle, embarrassment, fear/nervous, physical pain, anger/upset, and disgust), where each task was governed by a professional actor/director of performing arts. The frame rate of all videos in this dataset is 25 fps.

To make all videos used in this thesis to have the same frame rate, wherever this thesis refers to the stride S when setting up the span of frames to be considered, this will be automatically scaled to $2S$ for the SEMAINE dataset. During the DFNN training, both the input and the output are tensors of size $224 \times 224 \times 3$. It should be noted that all videos in training, validation and test sets of SEMAINE and BP4D were used as the test set for cross dataset evaluation of frame ranking in this section.

6.1.2 Dimensional affect estimation datasets

The dimensional affect estimation experiments were conducted on two video-based datasets, including a dataset collected under lab environment: SEMAINE (described in Sec. 6.1.1) and a dataset collected in the wild: Affwild 2 (Kollias & Zafeiriou, 2018). The Affwild 2 dataset is the extended version of the Affwild dataset (Kollias et al., 2019a). This dataset is collected from Youtube, which contains 558 face videos of 458 subjects (279 males and 179 females) with the frame rate of 30 frames per second. Again, all frames have been annotated with valence and arousal intensities within $[-1, 1]$.

6.1.3 Personality traits analysis datasets

As discussed in Chapter. 1, automatic personality traits recognition can be categorized into two types (Vinciarelli & Mohammadi, 2014): self-reported personality recognition and apparent personality recognition. While multiple video-based databases (McKeown et al., 2012; Ponce-López et al., 2016; Biel & Gatica-Perez, 2010; Sanchez-Cortes et al., 2011) are publicity available for apparent personality research, to the best of author's knowledge, only a few public databases have been recorded for studying self-reported personality. Even among the available self-reported personality databases, most of them have very few subjects. For example, the database introduced in (Celiktutan et al., 2017) only has 18 subjects (Please see (Junior et al., 2018) for the survey of personality databases).

Therefore, this thesis conducts self-reported personality experiments on the VHQ dataset (Jaiswal, Valstar, et al., 2019; Jaiswal, Song, & Valstar, 2019) (Please also see Sec.7.2), which consists of 165 videos collected from 55 participants. In this dataset , each participant completed 3 questionnaire interview sessions (BFI-10, PHQ-9, and GAD-7). During each session, participants were asked to answer a set of questions verbally under three interaction modes: face-to-face interaction, video conference, and human-to-robot interaction. The labels of Big-Five personality traits were obtained by asking participants to fill the BFI-44 questionnaire online.

The apparent personality estimation experiments were conducted on the ChaLearn (Ponce-López et al., 2016) dataset, where the Big-Five apparent personality traits are also employed as the labels. The ChaLearn dataset contains $10,000$ talking-to-the-camera videos of $2,764$ YouTube users, which have been assigned to three subsets: training set ($6,000$ videos), validation set ($2,000$ videos) and test set ($2,000$ videos). All videos are around 15 seconds long and are recorded at 30 frames per second. In the experiments of this section, the videos of the ChaLearn dataset have been re-sampled to 25 frames per second. The personality trait labels in this database were obtained by multiple human annotators using Amazon Mechanical Turk.

6.2 Implementation details

6.2.1 Frame ranking and dimensional affect estimation

Configuration: All experiments were carried out using the PyTorch library (Paszke et al., 2017) for deep learning. Both the input and the output of the trained models are tensors of size $224 \times 224 \times 3$. To generate SFI, DFI, DFR and DR (MDR) per frame, multiple time-windows with different settings were employed (explained in Sec.6.3.2). For the sequence-based approach, the RankSVM algorithm is also applied to generate the DFI and DFR from a sequence at test time, i.e., it is learned on the go for each subsequence of images. In contrast, for still image-based approach, it is important to note that, at test time, each still face image is used as the input (treated as the center frame) to the trained DFNN to generate the corresponding DR/MDR individually. Each produced representation is then used for the downstream tasks (frame ranking/dimensional affect estimation).

Pre-processing: The publicly available OpenFace 2.0 toolkit of (Baltrusaitis et al., 2018) was employed to detect a set of 68 facial landmarks for each frame. Using these landmarks, face regions are cropped, resized and aligned to meet the network size. Then, all pixels corresponding to the outer part of the convex hull defined by the landmarks are set to zero to remove all non-facial appearance information.

Models training details: In this thesis, multiple DFNNs were used for generating DRs, which have been trained and validated on the entire RECOLA dataset, using an Adam optimizer (Diederik P. Kingma, 2015) with a learning rate of 10^{-3}, and $\beta = (0.5, 0.9)$. The parameter θ in Eqn. 4.7 is set beforehand to ensure the chance level ranking accuracy to be less than 0.1%. In terms of models for dimensional affect estimation, the VGG-16 networks pre-trained by the VGG face database were combined with the Bidirectional Gated Recurrent Units (BGRU) (Cho et al., 2014) networks (BGRU was set to have one hidden layer with 200 neurons). During the training, all settings were set based on (Kollias et al., 2019b), with the input layer modified to adapt to the size of the input DRs/MDRs. In summary, for the sequence-based approach (presented in Sec. 4.1), the DFI, BDFI and DFR were computed per frame by corresponding adjacent sub-sequences at both training and test time, and then fed to VGG-16 for valence/arousal estimation. For the still image-based approach (presented in Sec. 4.2), at test time the DRs/MDRs are generated from each frame only using the well-trained DFNN models.

6.2.2 Person-specific representation

Person-specific representation generation

DFNN training: This section first employed Aff-Wild dataset to pre-train the encoder, for the task of valence and arousal intensities estimation, which follows the setting described in (Kollias et al., 2019a). After that, the RECOLA dataset (Ringeval et al.,

2013) is solely used to train and validate the DFNN, whereas all videos in training, validation and test sets of SEMAINE and BP4D datasets were used as the test set for the cross-dataset frame ranking evaluation of the trained DFNNs.

Training of PALs: To produce PALs for each person, the proposed approach keeps the DFNN's weights frozen, and keep the initial weights of PALs the same for all individuals. During the person-specific training, frames are fed to the network based on their time stamps, i.e. from the beginning of the video to the video's end. This not only ensures that the weights of PALs always converge (during training) to the same set of values for a particular video, but also ensures that the difference between individually trained PALs is only influenced by person-specific facial dynamics rather than the initialization of weights or the order in which the frames are used for training. To adjust the hyper-parameters of the PALs' training, this section equally divided each video in training and validation sets of Chalearn into two parts, i.e., one for training and the other for ranking accuracy validation. This section finally chose a set of hyper-parameters that generated the highest ranking accuracy, and then used each entire video to train a set of new PALs for each person as the person-specific representation. Meanwhile, the same hyper-parameters obtained from Chalearn experiments were used to train PALs for videos in the VHQ dataset. Since this process does not need human annotations, PALs are produced for each video in Chalearn and VHQ. All experiments of DFNN and PALs were implemented in PyTorch, where ADAM (Kingma & Ba, 2014) was used as the optimization method.

6.2.3 Spectral representation

Frame-level descriptors:

In the experiments, two types of frame-level descriptors were utilised as the basis to construct spectrum representations. The first descriptor (called latent descriptor in this thesis) is the per-frame latent 1-D feature obtained from the last encoder layer of the well-trained DFNN network (introduced in Sec.4.2). The second descriptor is called facial behaviour primitives, which consists of multiple objective, visual and non-verbal human behaviour attributes that are easily interpreted by both people and machines, to wit Facial

Action Units (AUs), head pose and gaze directions. In this thesis, 17 AU intensities, 6 gaze directions and 6 head poses are estimated for each video frame using the OpenFace 2.0 facial behaviour toolkit (Baltrušaitis, Robinson, & Morency, 2016):

- **Head pose**: The poses of the head can convey vital information about the mental state of a person especially related to attention and interest levels. The head poses are encoded in terms of the angle of rotation of the head about the X (horizontal), Y (vertical) and Z (depth) direction (pitch, yaw and roll), with the origin of the axes fixed at the camera. In order to take into account the variation of these angles due to other factors (e.g. person's height).

- **Eye gaze**: The behaviour encoded by the direction of eye gaze can also represent important psychological information and has been previously shown to be useful for predicting depression (Song et al., 2018). The eye gaze directions are encoded in terms of the angles which the direction makes with the X (horizontal) axis and Y (vertical) axis.

- **Facial Action Units (AUs)**: Facial AUs, which can be described as the atomic units of different facial expressions, represent the movement of one or more localized groups of facial muscles. First proposed by Ekman and Friesen (P. Ekman & Friesen, 1977), the system of Facial Action Coding System (FACS), is widely used for objective analysis of human facial expressions. The facial expressions represented by these AUs can provide useful cues to the emotional state of a person and hence is potentially valuable for depression analysis. The intensities of a total of 17 AUs have been used in the experiments are: AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26 and AU45.

Pre-processing and post-processing:

As described above, the publicly available OpenFace 2.0 toolkit is employed to output an aligned face image for each frame, as well as the facial behaviour primitives. For frames in which no face was detected or if the confidence value of the detected face was small, they were removed from analysis. To minimize the effects of participants' identity, the value of each human behaviour primitive was normalized by subtracting its corresponding median value computed over the whole video, respectively. Due to the high dimensionality of spectral vector, correlation-based Feature Selection (CFS) (M. A. Hall, 1999) was employed to select only the most relevant features before feeding them to ANNs. CFS is a popular feature selection technique which only selects those features which are highly correlated with the output variable but uncorrelated with each other, thereby giving a very compact set of useful features. The feature selections were done on training sets, and validated on the corresponding validation sets. In this feature selection algorithm, the Pearson's linear correlation coefficient is used to measure the correlations.

6.2.4 Personality models and training details:

Due to the limited amount of data (165 videos from 55 participants) in the VHQ dataset, the leave-one-subject-out cross validation is conducted to evaluate the ANN and CNN models of self-reported personality traits. In each fold, videos of 54 participants were used for training and the remaining one was used for testing. The reported results were obtained by averaging the results over all 55 folds. For the ChaLearn dataset, standard training, validation and test procedure is implemeted based on the corresponding subsets defined in the dataset. The dimensions of selected features in different experiments ranged from 16 to 42. During the validation and test, the same features were selected from the validation and test data. The training and evaluation of personality models (ANN) were implemented in MATLAB 2019, where RMSProp with learning rate of 0.0005 was used as the optimization method and training was done for 150 epochs.

6.2.5 Evaluation Metrics

Three standard measures were used to assess the performance of the proposed approaches on the dimensional affect estimation; firstly the Pearson Correlation Coefficient (PCC, Eq. 6.1),

$$\rho_{x,y} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{6.1}$$

where the cov is the covariance and σ_X, σ_Y are the standard deviations, secondly the Concordance Correlation Coefficient (CCC, Eq. 6.2):

$$\rho_{ccc} = \frac{2\rho_{x,y}\sigma_x\sigma_x}{\sigma_x^2 + \sigma_y^2 + \left(\mu_x - \mu_y\right)^2}, \tag{6.2}$$

where μ_x and μ_y are mean values of time-series predictions and labels while σ_x and σ_y are standard deviations, and thirdly the Mean Squared Error (MSE).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right)^2 \tag{6.3}$$

where x_i and y_i are the i_{th} prediction and ground-truth, respectively. In addition, the ranking accuracy is employed to evaluate the ranking ability of the generated representations of both proposed approaches, and is defined as

$$\text{RA} = \frac{\sum_{i=1}^{I}\sum_{n_i=1}^{N_i}\text{CRIP}(i)(n_i)}{K(K-1)\sum_{i=1}^{I}n_i} \tag{6.4}$$

where i is the video index, n_i is the n_{th} input image of the i_{th} video and $\text{CRIP}(i)(n_i)$ is the number of correctly ranked pairs for the corresponding input. Here, for each input image, there are k preceding and k succeeding frames, resulting in $k(k-1)/2$ preceding image pairs and $k(k-1)/2$ succeeding image pairs need to be ranked. Consequently, $K(K-1)\sum_{i=1}^{I}n_i$ represents the number of all ranked image pairs.

Meanwhile, four metrics are utilised to evaluate the personality traits estimation performance, which is Pearson Correlation Coefficient (PCC) defined above; the Root Mean Square Error (RMSE) defined in Eqn 6.5; and the **mean accuracy** measurement used in the ChaLearn challenge (Ponce-López et al., 2016) (defined in Eqn 6.6).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_i - y_i\right)^2} \tag{6.5}$$

In the above equations, f_i is the i^{th} prediction of the prediction vector f, y_i is the corresponding ground-truth in the ground-truth vector y, cov is the covariance and σ_f, σ_y are standard deviations of f and y respectively, μ_f and μ_y are the mean values of predictions and labels respectively. Meanwhile, the ACC is defined as

$$\mathrm{ACC} = 1 - \frac{1}{N_t} \sum_{i=1}^{N_t} |t_i - p_i| \tag{6.6}$$

where t_i and p_i are the labels and predictions, respectively and N_t is the number of videos.

6.3 Personality traits recognition results

6.3.1 Comparison to the state-of-the-art

This section compares the best systems achieved by the proposed two video-level facial dynamic encoding approaches as well as the combination of them against other published automatic personality analysis approaches on the VHQ and ChaLearn datsets. To do such comparison, three previous published baselines, including histogram feature (Jaiswal, Song, & Valstar, 2019), deep regression network (video features only) (Wei et al., 2018) and deep residual network (Güçlütürk et al., 2016; K. He, Zhang, Ren, & Sun, 2016), were implemented based on the codes provided online. For VHQ dataset, all baselines were trained and evaluated using the combined videos (the combination of three types of videos), which is termed as the video-level fusion in this section. Meanwhile, Two frame-level descriptors, e.g. facial behaviour primitives (AUs, Gazes, Head Poses) and the latent features produced from the last layer of the well-trained DFNN encoder, were utilised to construct spectral representations (spectral vector and spectral heatmap).

Table. 6.1 reports the self-reported personality prediction results on the VHQ dataset. It should be noted that the reported 'Single-task' results are the average results over a set of three single task videos, i.e., PHQ-9, GAD-7 and BFI-10 while the reported 'Single-scale' results are the average results over three single temporal scale settings. The 'multi-task' denotes the decision-level fusion results of three types of video and 'multi-scale' denotes the decision-level fusion results of systems using three different temporal scales. It can be observed that in terms of the average performance, the spectral vectors of behaviour primitives provided better predictions than the spectral vectors of latent features. This may due to that the high dimensions of the latent features ($1894 - D$), which makes it hard to be properly learned by a simple back-end models. More importantly, the top-3 DFNN-PALs systems achieved the better results than the spectrum representations with either behaviour primitives or latent features, and also outperformed the four baselines, for all three measures. Among them, the multi-scale person-specific representations performed best in terms of RMSE and ACC measures while the multi-task models gave the best performance in terms of PCC measure (0.39). These results suggest that the proposed dynamic encoding approach can extract important clues from facial behaviour, for self-reported personality prediction. In particular, this method achieved promising results on predicting Agreeableness, Neuroticism and Openness traits. In addition, it is clear that the combination of spectral representation and person-specific representation generated the best result among all systems, indicating that both multi-scale facial dynamics and person-specific facial dynamics are informative to self-reported personality traits. While person-specific information provided better performance, each of them contains some unique clues for this task. Fig.6.1 shows the predictions generated by our multi-scale person-specific representation on VHQ dataset. It can be observed that for Neuroticism trait most predictions of people whose ground-truth scores range from 12 to 27 are very close to corresponding labels. This could caused by the fact that there is more training data in this range and hence the model is able to perform better in this range. It can also be observed that, for all traits, most predictions have lower values than the corresponding ground-truths. One of the reasons for this observation could be the uneven distribution of scores and presence of outliers towards the lower end of scores.

Table. 6.2 compares the apparent personality prediction results on ChaLearn dataset. Again, the spectrum vectors of behaviour primitives clearly beaten the spectral vectors of latent features. It is also clear that the multi-scale DFNN-PALs model produced the best average PCC and RMSE results except the combined system, with the best PCC performance on three traits and best RMSE results on all five traits. In addition, the predictions from DRN method (Wei et al., 2018) achieved highest PCC for Agreeableness trait. In terms of ACC, both of DFNN-PALs systems outperformed most existing methods and achieved comparable average results (0.9168) to the state-of-the-art-method (CR-Net) (0.9187). In particular, for Agreeableness and Openness traits, the proposed DFNN-PALs approach achieved significant improvement over most existing approaches. Meanwhile, the combined system achieved the best predictions on three traits with the second best performance on Extraversion trait. This result again can indicates that both multi-scale facial dynamics and person-specific facial dynamics are informative to apparent personality traits. According to Fig.6.2, unlike predictions on VHQ dataset, the distributions of predictions on Chalearn dataset, are more balanced rather than having lower values compared to corresponding labels. This could be due to the fact that the training data is more evenly distributed with fewer outliers. It can also be observed that the average prediction performance on Chalearn dataset is higher than that on the VHQ dataset. This may be contributed to the large number of training examples in Chalearn dataset, leading the trained model to have better generalization capability.

To investigate whether there is statistical difference between the performance of the proposed best system (i.e., person-specific representation) and other reproduced systems (e.g. Spectrum (BP), NJU-LAMDA and DCC) is statistically significant, Table 6.3 reports the p value computed between the L2 results of our multi-scale person-specific representation and each of the others, where the L2 results were obtained from all test examples (2000 results on Chalearn dataset and 55 results on VHQ dataset). In addition, 5-fold cross-validation is conducted on the entire Chalearn dataset to evaluate the statistical difference in terms of PCC and ACC measurements. At each fold, the PALs of 8000 videos were used for training and the rest 2000 videos were used for testing. The p-value results reported in Table 6.4 are computed from the 5-fold results between by our multi-scale approach and each of the others. Both tables show that our multi-scale approach has clear advantages over others in predicting Agreeableness and Neuroticism traits. Meanwhile, Table 6.4 also indicates that there is statistical significant difference in the PCC and ACC results of Openness predictions.

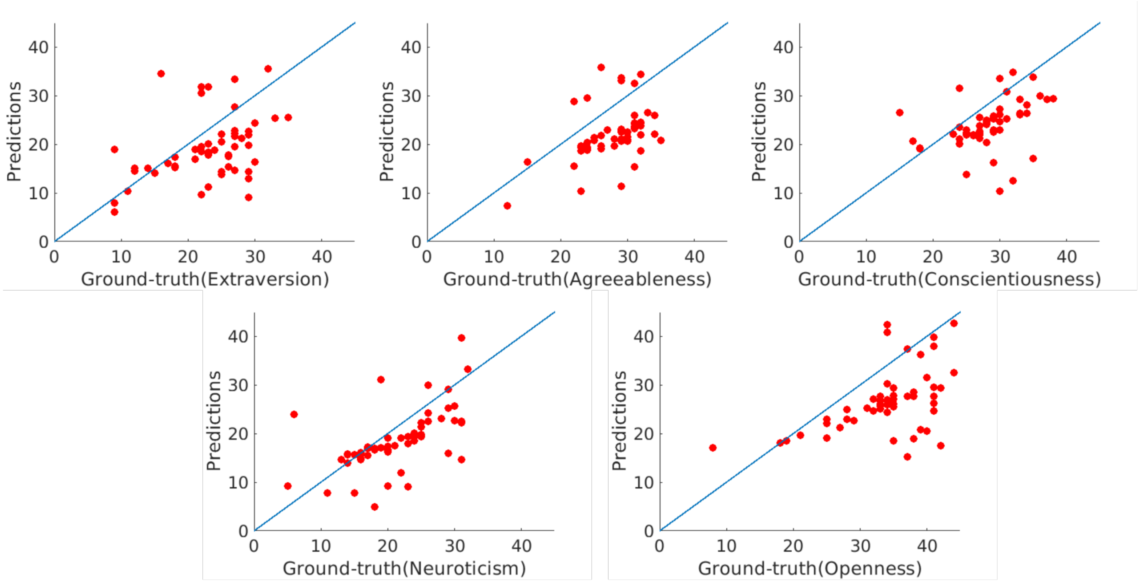


Figure 6.1: Predictions of our best system on the VHQ dataset.

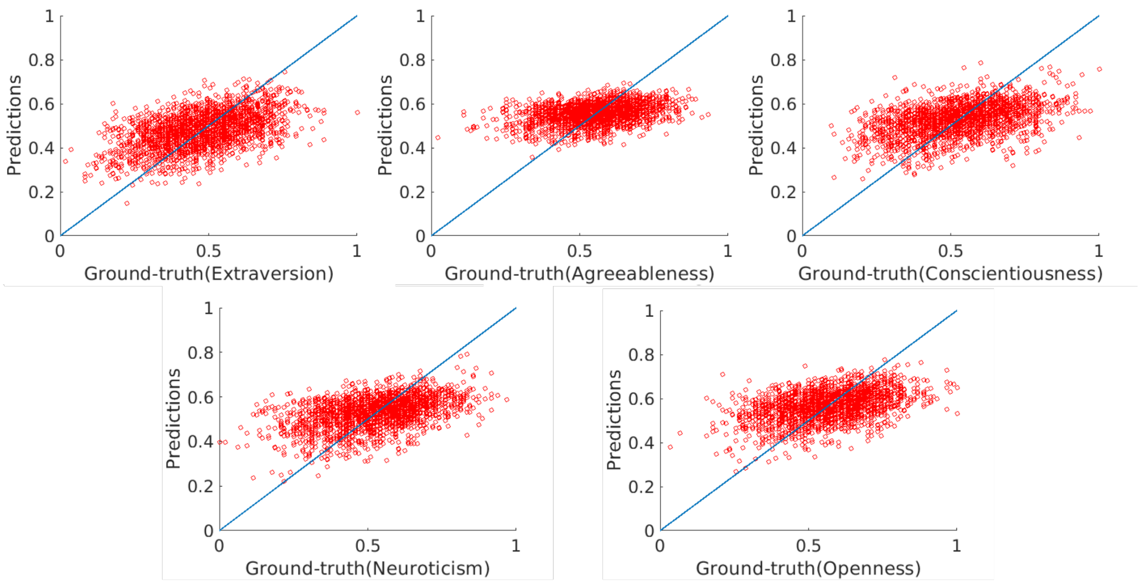


Figure 6.2: Predictions of our best system on the ChaLearn dataset.

	Traits	Extrav	Agree	Consc	Neuro	Open	Avg.
PCC	Histogram (Jaiswal, Song, & Valstar, 2019)	-0.15	-0.15	-0.28	0.04	-0.14	-0.14
	NJU-LAMDA* (Wei et al., 2018)	0.25	0.33	0.15	0.31	0.26	0.26
	DCC* (Güçlütürk et al., 2016)	0.11	0.08	0.02	0.12	0.14	0.09
	Spectrum (BP)	0.17	0.02	0.12	0.26	0.24	0.16
	Spectrum (LF)	0.11	-0.08	0.02	0.12	0.18	0.07
	DFNN-PALs (Single-task)	0.29	0.42	0.30	0.32	0.16	0.30
	DFNN-PALs (Multi-task:VF)	**0.42**	0.27	**0.34**	0.31	**0.36**	0.34
	DFNN-PALs (Multi-task:DF)	0.36	0.44	0.28	0.52	**0.36**	0.39
	DFNN-PALs (Single-scale)	0.19	0.31	0.18	0.32	0.27	0.25
	DFNN-PALs (Multi-scale)	0.34	**0.46**	0.29	0.51	0.27	0.37
	DFNN-PALs (Multi-scale) + Spectrum (BP)	0.37	0.43	0.33	**0.55**	**0.36**	**0.41**
RMSE	Histogram (Jaiswal, Song, & Valstar, 2019)	7.06	4.71	5.46	6.28	7.91	6.44
	NJU-LAMDA* (Wei et al., 2018)	6.19	4.33	4.88	6.09	6.97	5.69
	DCC* (Güçlütürk et al., 2016)	6.33	4.42	4.72	6.58	7.03	5.82
	Spectrum (BP)	6.50	4.50	4.82	6.23	7.19	5.91
	Spectrum (LF)	6.78	4.81	5.05	6.52	7.40	6.11
	DFNN-PALs (single-task)	6.02	4.13	4.55	6.01	7.00	5.54
	DFNN-PALs (Multi-task:VF)	**5.74**	4.37	**4.47**	6.08	**6.65**	5.46
	DFNN-PALs (Multi-task:DF)	6.03	4.07	4.60	5.49	6.85	5.41
	DFNN-PALs (single-scale)	6.24	4.18	4.64	6.00	6.82	5.58
	DFNN-PALs (Multi-scale)	6.02	4.03	4.56	5.45	6.79	5.37
	DFNN-PALs (Multi-scale) + Spectrum (BP)	5.86	**3.99**	4.50	**5.25**	6.71	**5.26**
ACC	Histogram (Jaiswal, Song, & Valstar, 2019)	0.8332	0.8410	0.8221	0.8361	0.8377	0.8340
	NJU-LAMDA* (Wei et al., 2018)	0.8412	0.8451	0.8385	0.8449	0.8438	0.8427
	DCC* (Güçlütürk et al., 2016)	0.8380	0.8433	0.8354	0.8370	0.8392	0.8386
	Spectrum (BP)	0.8408	0.8442	0.8397	0.8406	0.8417	0.8414
	Spectrum (LF)	0.8366	0.8423	0.8337	0.8314	0.8390	0.8360
	DFNN-PALs (single-task)	0.8426	0.8459	**0.8422**	0.8470	0.8387	0.8433
	DFNN-PALs (Multi-task:VF)	**0.8461**	0.8499	0.8395	0.8458	0.8450	0.8452
	DFNN-PALs (Multi-task:DF)	0.8425	0.8491	0.8380	0.8521	0.8416	0.8447
	DFNN-PALs (single-scale)	0.8407	0.8466	0.8398	0.8483	0.8425	0.8436
	DFNN-PALs (Multi-scale)	0.8430	**0.8517**	0.8415	0.8540	0.8433	0.8467
	DFNN-PALs (Multi-scale) + Spectrum (BP)	0.8455	0.8506	0.8418	**0.8587**	**0.8462**	**0.8486**

Table 6.1: Video-based self-reported personality prediction results on VHQ dataset. Multi-task:VF denotes the video-level fusion of multiple videos and Multi-task:DF denotes the decision-level fusion of multiple videos. Spectrum (BP) denotes the spectrum representation produced from behaviour primitives (Song et al., 2020). Spectrum (LF) denotes the spectrum representation produced by the frame-level features extracted from the last encoder layer of the well-trained DFNN. $*$ denotes the results obtained by our own implementation using the code provided online.

6.3.2 Ablation studies of the person-specific representation

This section firstly evaluate the short-term facial dynamic encoding ability of the proposed self-supervised learning strategy that is used by person-specific dynamic encoding, based

	Traits	Extrav	Agree	Consc	Neuro	Open	Avg.
PCC	Histogram (Jaiswal, Song, & Valstar, 2019)	0.30	0.05	0.22	0.22	0.20	0.20
	NJU-LAMDA* (Wei et al., 2018)	0.43	**0.37**	**0.45**	0.34	0.36	0.39
	DCC* (Güçlütürk et al., 2016)	0.36	0.12	0.20	0.25	0.25	0.24
	Spectrum (BP)	0.37	0.30	0.34	0.36	0.32	0.34
	Spectrum (LF)	0.23	0.19	0.25	0.33	0.23	0.25
	DFNN-PALs (Single-scale)	0.50	0.30	0.43	0.44	0.43	0.42
	DFNN-PALs (Multi-scale)	0.52	0.31	**0.45**	0.45	0.44	0.45
	DFNN-PALs (Multi-scale) + Spectrum (BP)	**0.56**	0.35	**0.45**	**0.47**	**0.47**	**0.46**
RMSE	Histogram (Jaiswal, Song, & Valstar, 2019)	0.17	0.15	0.17	0.17	0.16	0.16
	NJU-LAMDA* (Wei et al., 2018)	0.14	0.12	**0.13**	0.14	0.13	0.13
	DCC* (Güçlütürk et al., 2016)	0.15	0.14	0.15	0.15	0.14	0.15
	Spectrum (BP)	0.15	0.13	0.14	0.14	0.14	0.14
	Spectrum (LF)	0.20	0.15	0.16	0.14	0.15	0.16
	DFNN-PALs (single-scale)	0.12	0.13	0.14	0.14	0.12	0.13
	DFNN-PALs (Multi-scale)	0.12	**0.10**	**0.13**	0.12	**0.11**	0.12
	DFNN-PALs (Multi-scale) + Spectrum (BP)	**0.10**	**0.10**	**0.13**	**0.11**	**0.11**	**0.11**
ACC	Baseline (Escalante et al., 2018)	0.9112	0.9112	0.9152	0.9103	0.9111	0.9118
	PML (Bekhouche et al., 2017)	0.9155	0.9103	0.9137	0.9082	0.9100	0.9115
	BU-NKU (Kaya, Gurpinar, & Ali Salah, 2017)	**0.9212**	0.9137	0.9197	0.9146	0.9170	0.9172
	NJU-LAMDA (Wei et al., 2018)	0.9133	0.9126	0.9166	0.9100	0.9123	0.9130
	Evolgen (Subramaniam et al., 2016)	0.9150	0.9119	0.9119	0.9099	0.9117	0.9121
	DCC (Güçlütürk et al., 2016)	0.9107	0.9102	0.9138	0.9089	0.9111	0.9109
	ucas (Ponce-López et al., 2016)	0.9129	0.9091	0.9107	0.9064	0.9099	0.9098
	CR-Net (Li et al., 2020)	0.9200	0.9176	**0.9218**	0.9150	0.9191	**0.9187**
	PerEmoN (L. Zhang et al., 2019)	0.920	0.914	0.921	0.914	0.915	0.917
	Histogram (Jaiswal, Song, & Valstar, 2019)	0.8949	0.8970	0.9001	0.8913	0.8975	0.8962
	Spectrum (BP)	0.9165	0.9099	0.9178	0.9109	0.9117	0.9134
	Spectrum (LF)	0.8860	0.8997	0.9061	0.9082	0.9035	0.9007
	DFNN-PALs (single-scale)	0.9178	0.9255	0.9063	0.9132	0.9177	0.9161
	DFNN-PALs (Multi-scale)	0.9183	0.9262	0.9082	0.9133	0.9180	0.9168
	DFNN-PALs (Multi-scale) + Spectrum (BP)	0.9203	**0.9285**	0.9108	**0.9161**	**0.9197**	**0.9187**

Table 6.2: Apparent personality prediction results on ChaLearn dataset.

	Traits	Extrav	Agree	Consc	Neuro	Open
VHQ	Spectrum (BP)(Song et al., 2020)	+(2.94E-02)	+(1.58E-02)	+(3.11E-02)	-(6.68E-02)	-(5.27E-01)
	NJU-LAMDA* (Wei et al., 2018)	-(1.91E-01)	+(3.39E-03)	+(3.78E-02)	-(4.68E-01)	-(2.20E-01)
	DCC* (Güçlütürk et al., 2016)	+(3.01E-03)	+(5.90E-04)	+(2.14E-02)	+(2.97E-02)	+(6.34E-03)
ChaLearn	Spectrum (BP)(Song et al., 2020)	-(1.93E-01)	+(3.68E-03)	+(9.84E-04)	+(1.17E-02)	+(3.86E-03)
	NJU-LAMDA* (Wei et al., 2018)	-(1.54E-01)	+(8.16E-03)	-(5.81E-02)	+(4.70E-02)	-(5.37E-02)
	DCC* (Güçlütürk et al., 2016)	+(7.05E-03)	+(1.52E-04)	+(2.90E-02)	+(2.57E-04)	+(4.15E-05)

Table 6.3: Statistical significant differences of the L2 error between the proposed multi-scale approach's predictions and three reproduced systems' predictions on VHQ and Chalearn dataset, where + / - denotes that there is / there is no statistical significant differences between our approach and the other approach while the corresponding P-value is presented in the square.

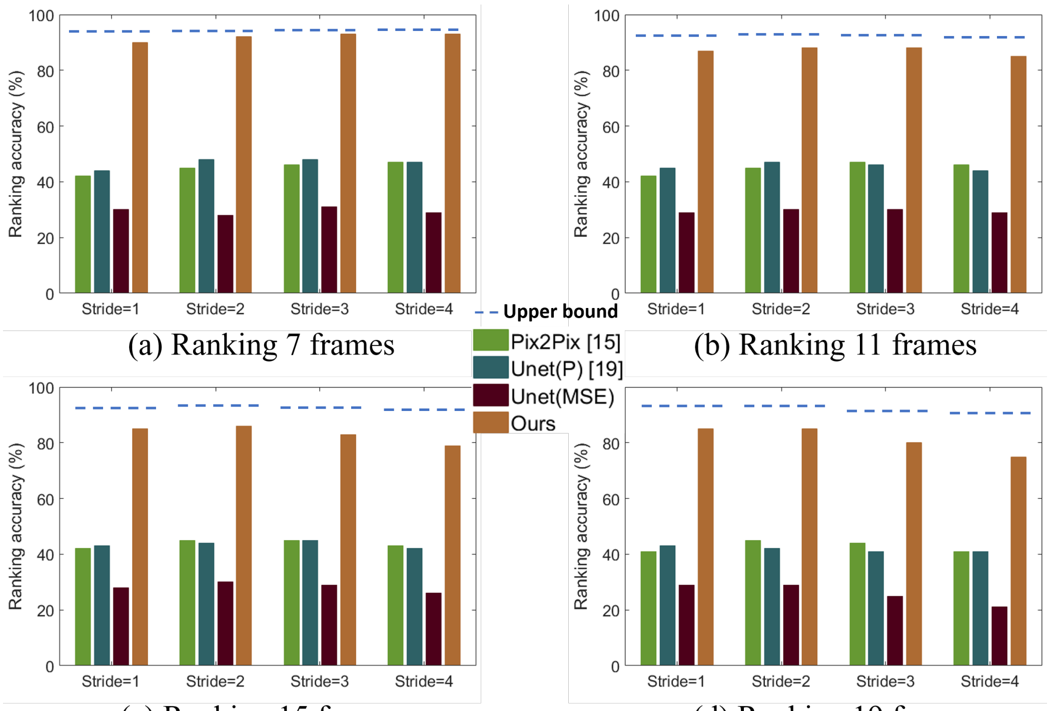
	Traits	Extrav	Agree	Consc	Neuro	Open
PCC	Spectrum (BP)(Song et al., 2020)	+(8.12E-04)	-(4.68E-01)	+(4.29E-05)	+(3.29E-04)	+(1.18E-05)
	NJU-LAMDA* (Wei et al., 2018)	+(1.42E-02)	-(1.04E-01)	-(9.94E-01)	+(1.12E-03)	+(3.55E-03)
	DCC* (Güçlütürk et al., 2016)	+(1.06E-04)	+(2.83E-06)	+(4.01E-08)	+(3.16E-06)	+(5.20E-06)
ACC	Spectrum (BP)(Song et al., 2020)	-(5.99E-01)	+(1.05E-05)	+(1.21E-03)	+(2.03E-03)	+(4.76E-04)
	NJU-LAMDA* (Wei et al., 2018)	+(6.15E-03)	+(2.74E-06)	+(9.21E-03)	+(2.61E-02)	+(1.85E-03)
	DCC* (Güçlütürk et al., 2016)	+(3.42E-03)	+(3.76E-05)	+(1.41E-02)	+(4.98E-03)	+(8.06E-04)

Table 6.4: Statistical significant differences of the PCC and ACC results between the proposed multi-scale approach and three reproduced systems in 5-fold cross validation on VHQ dataset.

on two applications: frame ranking (Sec. 6.3.2) and dimensional affect estimation (Sec. 6.3.2). Then, the influence of several factors, including encoder pre-training settings, the size of time-window, the employed fusion strategies, and task contents and video lengths, on personality recognition performance are evaluated.

Frame ranking results

Comparison with other generative methods: Fig.6.3 shows the capability of the DRs generated by DFNN in ranking the corresponding adjacent frames in SEMAINE and BP4D. The evaluation was conducted under different scenarios by choosing a set of different window lengths and strides to train models and generate the corresponding representations. The number of frames used per training image is $N = 2T+1$ (T preceding frames, T succeeding frames and the given frame). The experiments sample N frames using four different strides S. The image sequence range is then of $N \times S$ frames. The ranking capabilities of the proposed approaches are reported for $T = \{3, 5, 7, 9\}$ (i.e. $N = \{7, 11, 15, 19\}$). In the most extreme case, i.e. when $T = 9$ and $S = 4$, the ranking is measured on a window size of $N = 2T + 1 = 19$ frames, evenly sampled from a sequence of $N \times S = 76$ frames. At test time, frames are chosen following the same sampling procedure as that of the corresponding model. To compute the ranking accuracy, the DRs are generated for each of the images available in the corresponding datasets, which is then projected onto the frames lying within the corresponding window of N frames, sampled with stride S. The ranking accuracy is the percentage of pairs that are correctly ranked (Eq.6.4), i.e. the percentage of pairs for which $\delta_{ab}(t) > 0$. Meanwhile, the results



(a) Ranking 7 frames

(b) Ranking 11 frames

(c) Ranking 15 frames

(d) Ranking 19 frames

Figure 6.3: Average ranking accuracy (%) on two datasets. Four generative models are trained using RECOLA dataset and tested on SEMAINE and BP4D datasets. RankSVM classifiers were trained on SEMAINE and BP4D datasets, and each classifier only rank its training frames. The results obtained by RankSVM are treated as the upper bound.

shown with blue dash lines correspond to the results achieved by the sequence-based approach, i.e., applying a RankSVM at test time, trained using the frames that were later ranked by it. Thus, the results given by the RankSVM are treated as an *upper bound* for the ranking accuracy. In particular, Fig. 6.3 compares the DFNN trained by the proposed self-supervised learning algorithm against the following approaches:

- UNet (MSE). Using the dynamic images, the model is trained using as objective the Mean Squared Error.

- UNet (P). In this method, the dynamic images is used as target representation, and the objective function used to train the model is the Perceptual loss proposed in (Johnson, Alahi, & Fei-Fei, 2016).

- Pix2Pix (Isola et al., 2017) refers to using a conditional GAN, again using the dynamic images as the corresponding targets.

According to Fig. 6.3, the ranking accuracies of the proposed DFI (the ranking accuracy of BDFI is almost the same to the DFI) is much better than other methods, indicating that the extended dynamic image algorithm has strong capability to rank frames that contains facial behaviours. Then, the results show how the proposed still image-based self-supervised learning algorithm (DFNN) achieved similar results to those given by the sequence-based approach (DFI) that uses at test time the adjacent frames to compute the kernel. Remarkably, the still image-based approach (DFNN) yields around 76% ranking accuracy even for the longest cases (i.e., when ranking 19 frames with different strides). It is also clear that, when pairing the input images with a DR to serve as a basis to learn generative networks, the ranking accuracy at test time degrades substantially, indicating that the subtle errors in the reconstruction loss do not necessarily correlate with errors in the ranking of frames. This also illustrates the contribution of the Rank Loss at the task of defining the form of the DR, allowing for a better generalization.

Meanwhile, It should be noted that the proposed DR is capable of accurately ranking both preceding and succeeding frames, something not possible with methods trained with explicit target representations. This is because the original dynamic image algorithm would only be able to rank either preceding or succeeding frames, and it would need to be given the temporal information at test time. An example is shown in Fig. 6.4. This clearly illustrates the capacities of the proposed still image-based approach at the task of sorting the surrounding frames of a given image relative to their position to it.

Influence of the encoder pre-training: This section also evaluates the effects of the proposed emotion-guided encoder pre-training for the subsequent learning of facial dynamics in terms of frame ranking accuracy. In particular, four different settings are used to pre-train the encoder: 1. pre-training it with arousal and valence labels and freezing its weights afterwards; 2. pre-training it and then reducing the learning rate to 5×10^{-5} as the initial learning rate for the subsequent training; 3. pre-training it and then keeping the initial learning for the subsequent training; 4. No pre-training is conducted. The evaluation was done using different time-window hyperparameters, w.r.t. window length and stride, for training networks and generating corresponding DRs. The number of frames used in the time-window around each input image is $N = 2T + 1$ (T preceding

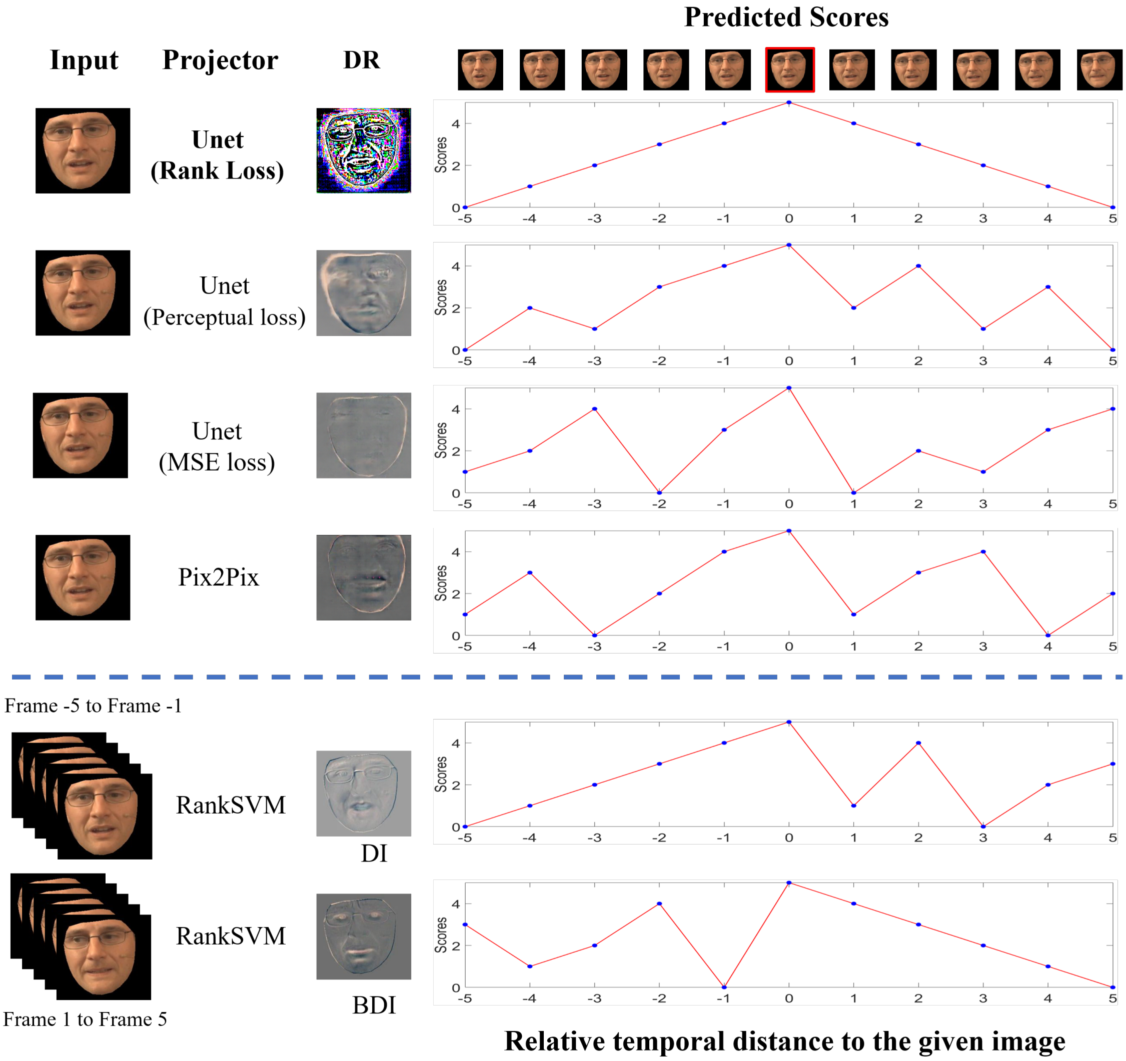


Figure 6.4: Examples of ranking frames using DRs generated by different methods. The networks of Pix2Pix, Unet(P) and Unet(MSE) were trained using Seq DI as the target.

frames, T proceeding frames and the given frame), which are the same to the settings described above. At test time, frames are chosen following the same sampling procedure as that used during training of the model. The results in Fig. 6.5 show that the proposed method with the encoder of setting 4 achieved results similar to those given by RankSVM (Smola & Schölkopf, 2004). Remarkably, this setting yields around 75% accuracy even for the longest cases (i.e., when ranking 19 frames with a stride of 4 frames). When comparing the four different settings, it is clear that settings 3 and 4 achieve similar results, which are better than settings 1 and 2. In addition, it can be observed that when the size of the time-window is less than 25 frames (0.5s for preceding frames and 0.5s for proceeding frames), the ranking accuracy remains mostly stable and rises slightly with increasing stride. However, when the time-window becomes large, the ranking performance drops significantly. This suggests that the proposed rank loss allows DFNN to efficiently learn short-term facial dynamics but is not suitable for modelling long-term dynamics.

These results also indicate that the emotion-guided encoder pre-training did not have a significant impact on ranking accuracy. On the other hand, the learning rate of the encoder plays an essential role in learning general facial dynamics. Reducing or freezing the learning rate of the encoder resulted in lower-ranking accuracy. This section assumes that the majority of emotion-related dynamics are not crucial for frame ranking, and thus the relatively large learning rate can force both emotion-related part of encoder and emotion-independent part of encoder to learn temporal cues that relate to frame ranking.

Dimensional affect estimation results

This section evaluates the proposed approach on the dimensional affect estimation task, i.e., predicting the intensities of valence and arousal for each frame. In particular, this section compares the results achieved by the proposed dynamic representations (DFI/DFR/DR/MDR) against the static aligned faces extracted from videos. As introduced in Sec. 6.2, the VGG-16 network (Parkhi, Vedaldi, Zisserman, et al., 2015) is fine-tuned for each of the alternatives aforementioned. More specifically, a VGG-16 network is employed to predict valence and arousal intensities by processing the generated DFI/DFR/DR/MDR, as well

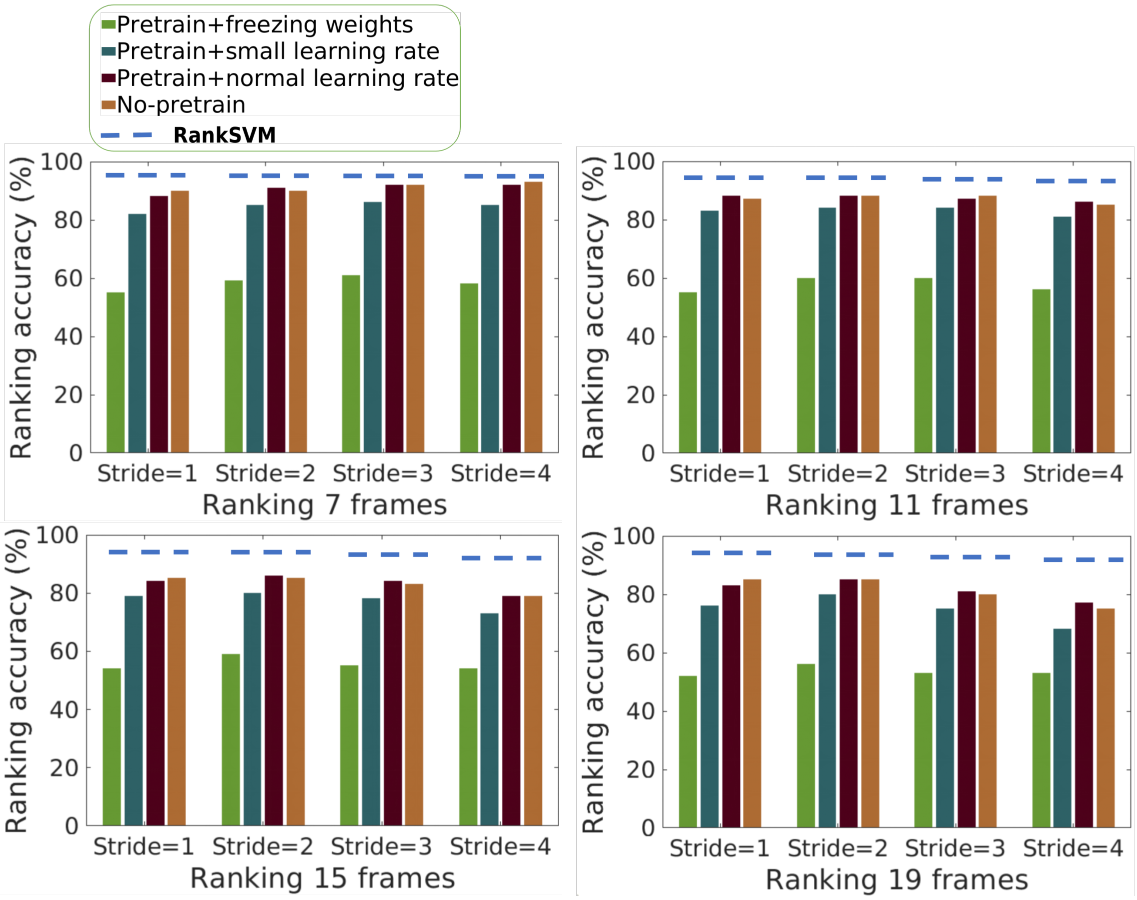


Figure 6.5: Average frame ranking accuracy obtained by four encoder settings.

	Arousal		Valence	
Method	CCC	MSE	CCC	MSE
Static	0.201	0.098	0.185	0.113
Pix2Pix* (Isola et al., 2017)	0.082	0.166	0.076	0.195
Unet(P)* (Johnson et al., 2016)	0.145	0.125	0.102	0.172
Unet(MSE)*	0.015	0.181	0.029	0.189
DFI	0.293	0.077	0.276	0.079
BDFI	0.291	0.077	0.276	0.080
DFR	**0.358**	**0.042**	**0.320**	**0.061**
SDR	0.289	0.078	0.273	0.082
MDR	0.316	0.058	0.299	0.072

Table 6.5: Affect estimation results on the SEMAINE dataset. SI denotes the still face image; $*$ denotes our own implementation

as for the per-frame output of other generative methods, e.g., UNet(MSE), UNet(P), and Pix2Pix. The results on SEMAINE and Affwild2 dataset are shown in Table 6.5 and Table 6.6. Firstly, it is clear that all proposed dynamic representations outperformed the static faces on both datasets (except that SDR generated slightly worse arousal predictions and BDFI achieved slightly worse valence predictions), illustrating that the facial dynamics encoded by the proposed dynamic representations can provide additional useful information to static face for dimensional affect estimation. Secondly, the DR/MDR generated from DFNNs yielded better results than all other generative approaches. Again, this indicates that the proposed rank loss is superior to the reconstruction loss in capture generic dynamics from unlabelled face videos. Finally, the DR/MDR generated by the DFNN achieved comparable results to the sequence-based dynamic representations under most conditions, showing that the facial dynamics learned by the proposed self-supervised learning approach have similar impacts as the dynamics learned from the sequence-based approach on dimensional affect estimation tasks.

	Arousal		Valence	
Method	CCC	MSE	CCC	MSE
Static	0.333	0.114	0.297	0.156
Pix2Pix*+VGG (Isola et al., 2017)	0.091	0.166	0.088	0.195
Unet(P)*+VGG (Johnson et al., 2016)	0.192	0.125	0.134	0.172
Unet(MSE)*+VGG	0.063	0.181	0.082	0.189
DFI	0.351	0.109	0.297	0.151
BDFI	0.352	0.110	0.295	0.153
DFR	**0.399**	**0.102**	**0.358**	**0.136**
SDR	0.326	0.119	0.299	0.151
MDR	0.345	0.111	0.316	0.145

Table 6.6: Affect estimation results on the Affwild 2 dataset. SI denotes the still face image; $*$ denotes our own implementation

Ablation studies on personality analysis tasks

Encoder pre-training settings: Fig. 6.6 compares the personality prediction results returned by person-specific representation encoded in PALs, based on the four encoder settings described earlier. The performance was computed on both VHQ and ChaLearn datasets. The comparison is made in terms of the PCC measure, averaged over all five personality traits. For both datasets, it is clear that the results achieved by the third setting outperformed the fourth setting. The difference in performance on the two datasets may be caused by: 1. the large differences in the number of training examples, i.e., 54 in VHQ dataset and 6000 in Chalearn dataset; 2. the different annotations: while Chalearn provides apparent personality traits labels, the VHQ dataset provides self-reported personality traits labels. However, it can be observed that the relative difference in performance between different encoder settings are almost the same. In summary, the self-supervised training of the pre-trained encoder can still lead the learned model to retain some emotion information that is positively associated with personality, resulting in better personality performance when keeping the initial learning rate.

Size of time-window: Fig.6.7 and Fig.6.8 compare the personality prediction results of the person-specific descriptors, produced by three different time-windows with

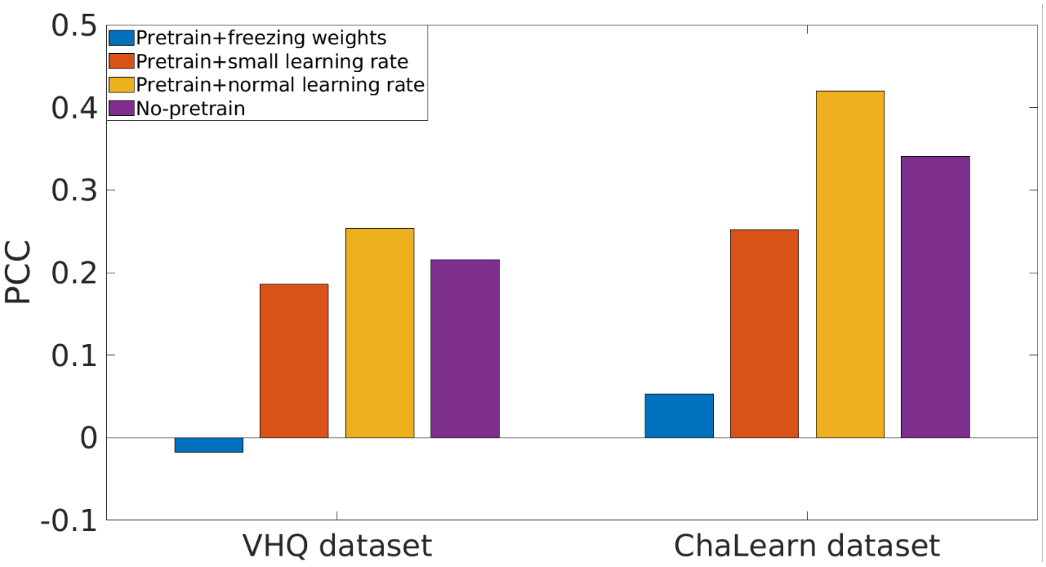


Figure 6.6: Average personality prediction results obtained by four encoder settings.

stride 2: $N = 7$ (0.52s), $N = 11$ (0.84s) and $N = 15$ (1.16s). It can be observed that the best self-reported personality predictions (average PCC of five traits is 0.2552) were achieved when the time-window size was 0.84s while the best apparent personality prediction results (average PCC of five traits is 0.4192) were achieved when the time-window size was 1.16s. It also can be observed that self-reported personality prediction is more sensitive than the apparent personality prediction, to the time-window size. More importantly, when conducting decision-level fusion of multi-scale predictions, i.e., the final predictions were made by applying linear regression to combine the predictions from all models, the performance clearly exceeds the ones from single-scale models (please see the following paragraph for more details on different fusion strategies). The hypothesis of this result is that the most relevant features for personality prediction may occur at different temporal scales. Therefore, using multi-scale dynamics can help learn more optimum models for personality prediction.

Fusion strategy: In real world applications, multiple videos may be available for each person. Therefore, it is interesting to explore how to optimally combine them to generate more accurate predictions. As three videos were recorded for each participant in the VHQ dataset, Fig.6.9 evaluates the following three ways to combine them for personality analysis: video-level fusion (combine three videos to a single video to train PALs and generate person-specific descriptors, where frames located at the borders were not used

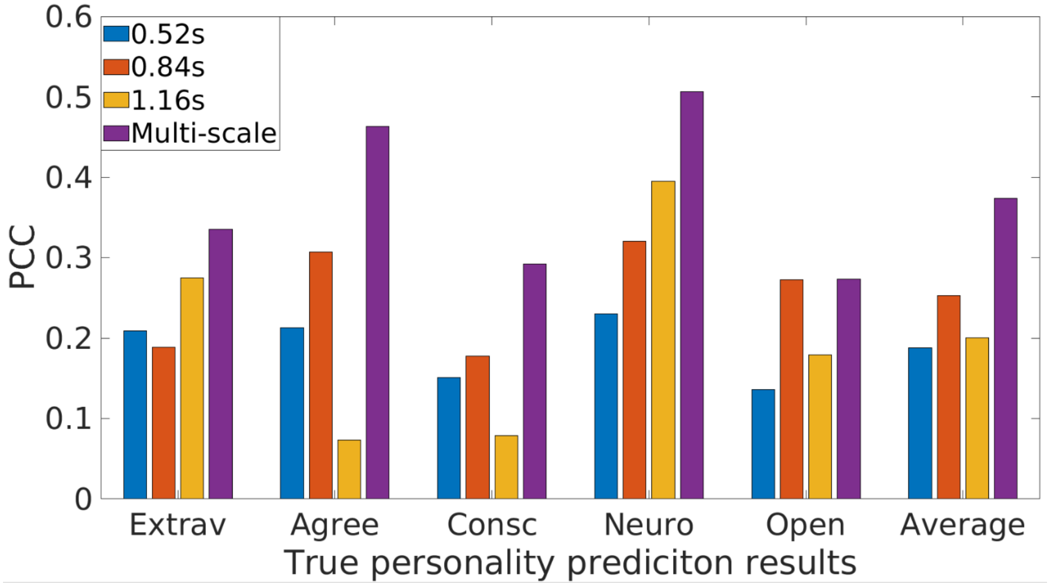


Figure 6.7: Self-reported personality prediction results obtained by four time-window settings on VHQ dataset, using person-specific representations. The average PCC for five traits are: 0.1877 (0.52s), 0.2532 (0.84s), 0.2002 (1.16s), and 0.3740 (multi-scale).

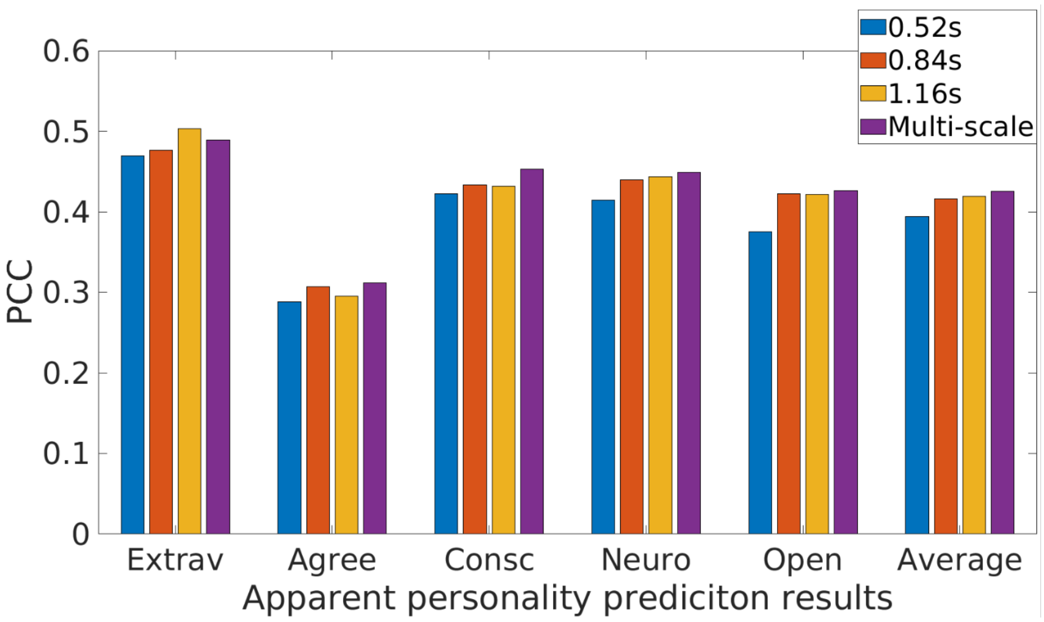


Figure 6.8: Apparent personality prediction results obtained by four time-window settings on ChaLearn dataset, using person-specific representations. The average PCC for five traits are: 0.3942 (0.52s), 0.4157 (0.84s), 0.4192 (1.16s), and 0.4258 (multi-scale).

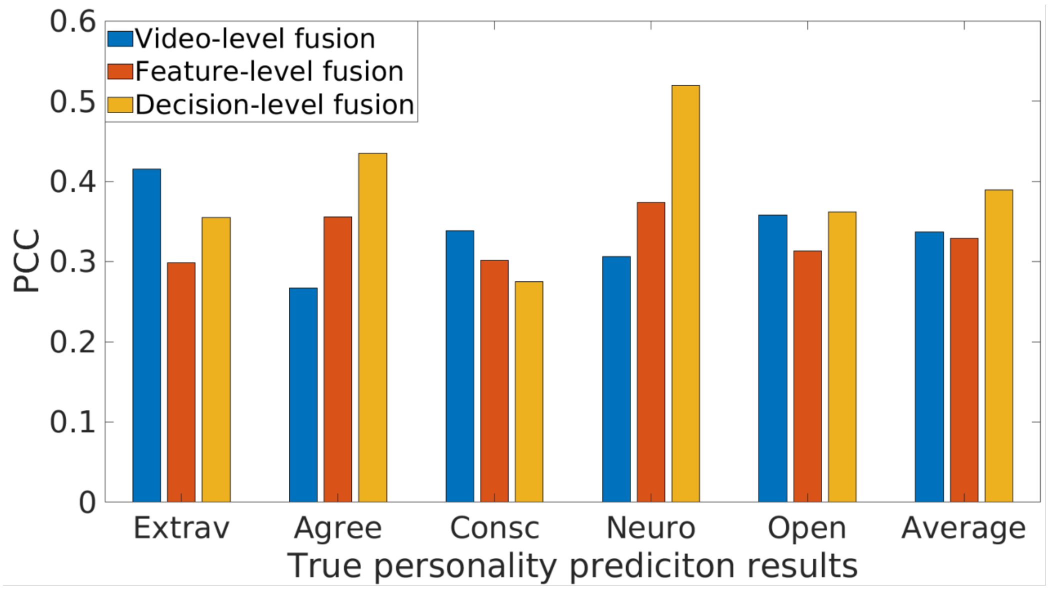


Figure 6.9: Self-reported personality prediction results obtained by three fusion strategies on VHQ dataset. The average PCC for five traits are: 0.3370 (video-level fusion), 0.3287 (feature-level fusion), and 0.3894 (decision-level fusion).

for the PAL training.), feature-level fusion (concatenating the person-specific descriptors produced by each video) and decision-level fusion (combining the predictions generated from each video using linear regression). It can be observed from the Fig.6.9 that the decision-level fusion produced the best performance for three traits (agreeableness, neuroticism and openness), and achieved the best average performance (PCC= 0.3894) over five traits. Additionally, video-level fusion outperformed the feature-level fusion in terms of average performance.

Task contents and video length: The VHQ dataset recorded three videos for each participant under three different interview scenarios, i.e., verbally answering PHQ-9, GAD-7 and BFI-10 questionnaires, respectively. To investigate the influence of the task contents on automatic personality traits analysis, Fig. 6.10 compares the personality traits prediction results obtained by person-specific representations from each type of videos in the VHQ dataset. As we can see, the best performance of four traits, i.e., Extraversion, Agreeableness, Conscientiousness and Neuroticism, were obtained from GAD-7 videos, while the models trained from BFI-10 videos performed best for the Openness trait. These results suggest that under the recording conditions of VHQ dataset, the GAD-7 questionnaire-based interview triggers facial behaviours that are more informative for

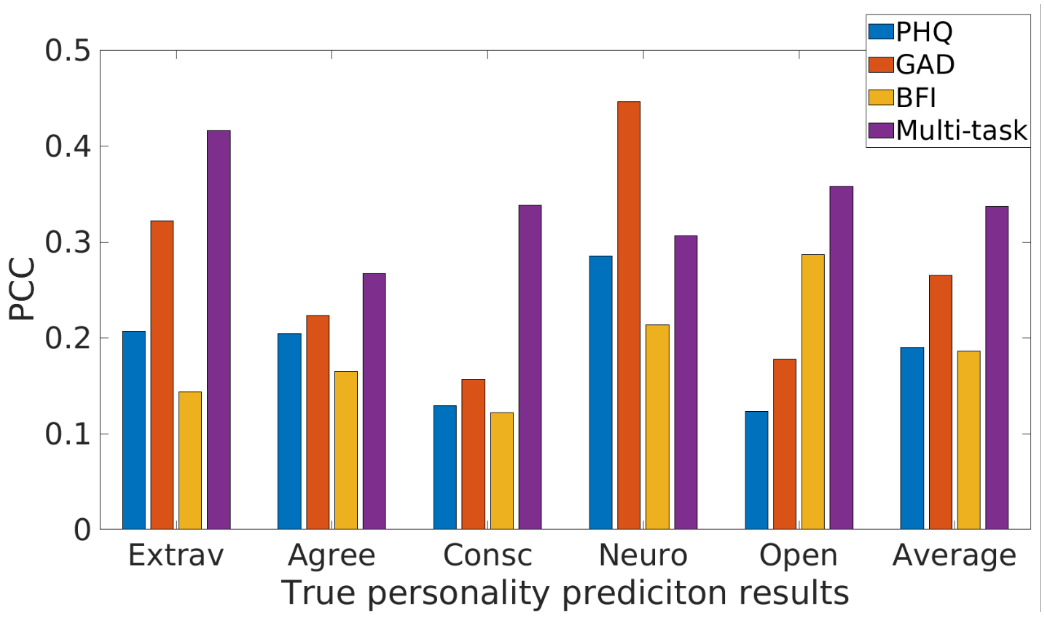


Figure 6.10: Self-reported personality prediction results obtained by videos of four task settings on VHQ dataset, using person-specific representations. The average PCC for five traits are: 0.1898 (PHQ videos), 0.2652 (GAD videos), 0.1861 (BFI videos), and 0.3371 (Multi-task).

E, A, C, N traits, while behaviours triggered by BFI-10 questionnaire-based interviews are more relevant to Openness. In addition, there are large variations on prediction performance across the three types of videos, indicating that the behaviour triggered by different tasks may have considerable impact on the performance of personality prediction.

Since an individual's personality traits are expected to be relatively stable over time, this section also investigates the impact of the duration of interview on personality analysis, i.e. this section assumes longer videos that contain facial behaviours triggered by multiple diverse tasks should result in PALs with better generalization capability. To test this hypothesis, Fig. 6.10 also reported the personality prediction results obtained by person-specific representations when PALs are trained by combining the three types of videos in the VHQ dataset, (Video-level fusion). The resulting videos are then longer than any of the individual PHQ/BFI/GAD videos, and contains three interviews. Fig. 6.10 illustrates that the video-level fusion significantly improves the performance for four personality traits, and achieved the best average PCC of 0.3371 over all the five traits, clearly outperforming the PCC results obtained from PHQ-9 videos (0.1898), GAD-7 videos (0.2652) and BFI-10 videos (0.1861). These results suggest that using longer videos containing different tasks, forces PALs to learn behaviour patterns which occur frequently under different conditions i.e. patterns stable over time are indicative of personality.

6.3.3 Ablation studies of the spectral representation

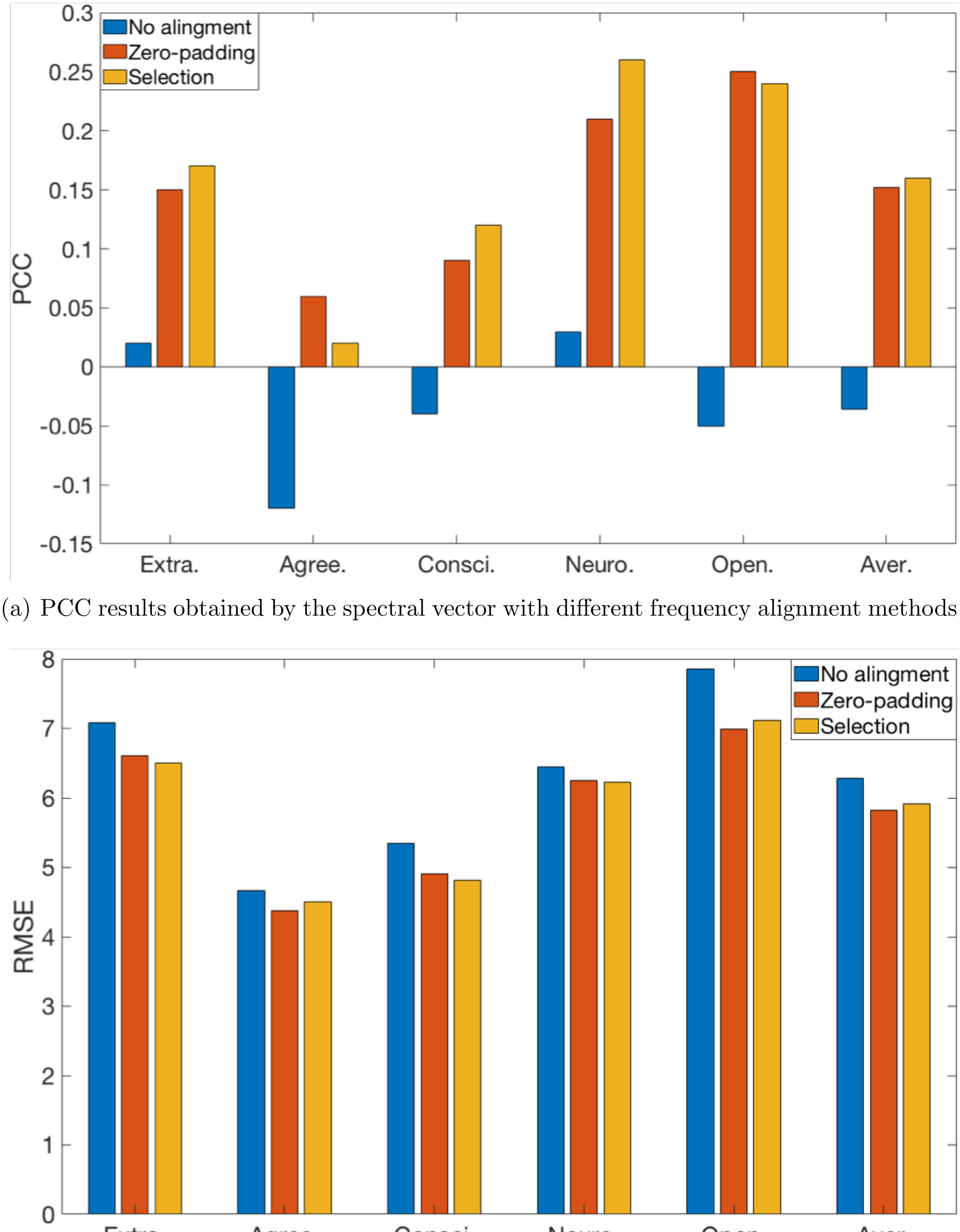
This section evaluates two settings of the spectral approach on personality traits estimation. Firstly, it compares the performance achieved by two frequency alignments. Then, it shows that the spectral vector achieved better performance than the spectral heatmap in VHQ dataset.

Comparison of frequency alignment methods:

As described in Section 5.3, two frequency alignment methods, i.e. zero-padding and frequency selection, have been employed. Fig.6.11 compares them against the models where no alignment was done. The reported results are the average values achieved by spectral vectors on three types of VHQ videos. The results demonstrate that frequency alignment is necessary as both alignment methods achieved enhanced performance compared to models with no alignment. It can also be observed that the two frequency alignment methods achieved similar performance, with the proposed frequency selection method performing slightly better than zero-padding. Both methods have their own advantages: while zero-padding can increase the resolution of the spectral signals, the use of frequency selection prevents the original signal from distortion.

Comparison of spectral representations:

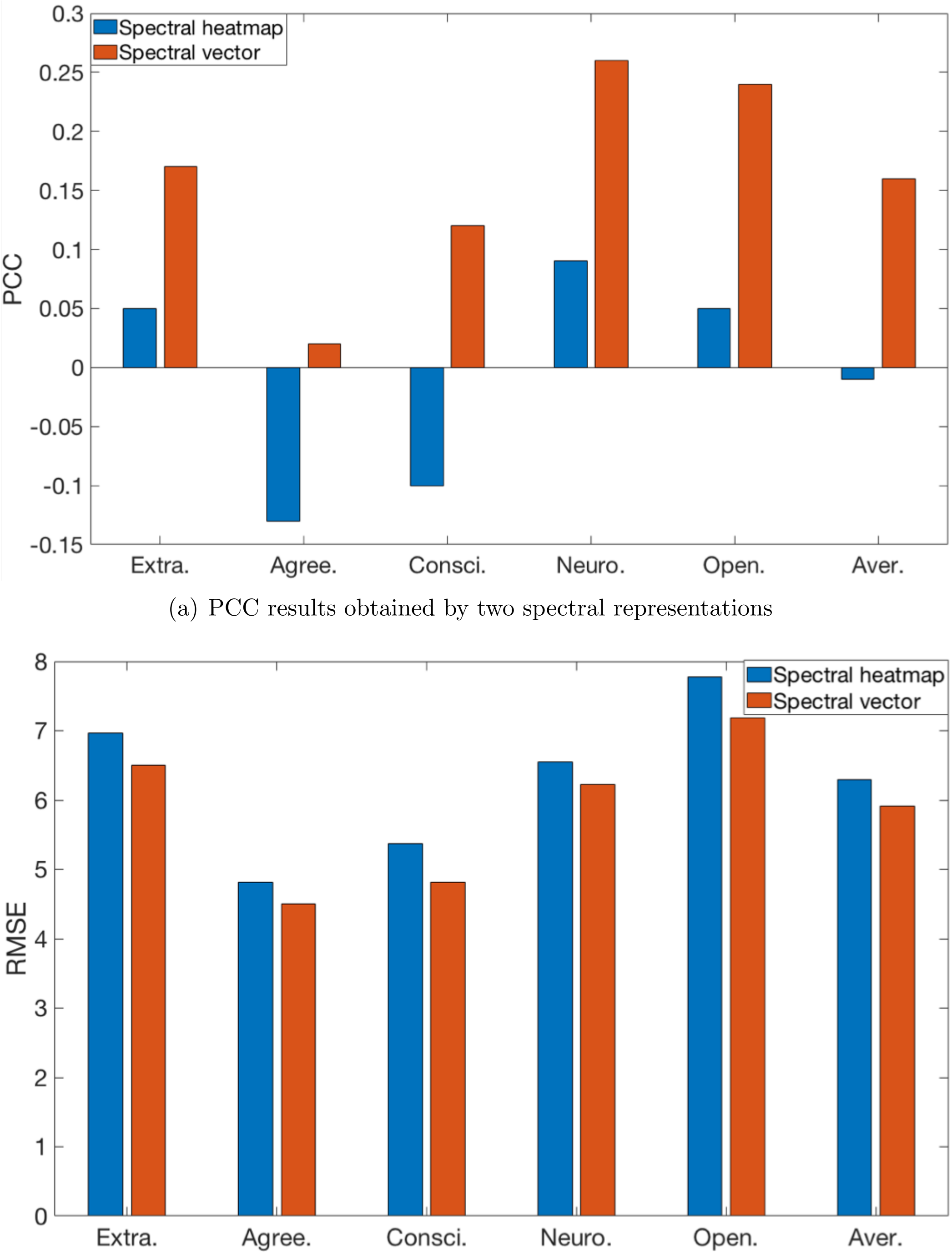
This section also compares the performance obtained by two proposed spectral representations, i.e., spectral heatmap and spectral vector. Fig.6.12 shows the average results achieved by them on three types of VHQ videos. Additionally, these results represent the average performance of two frequency alignments. It is clear that the performance of spectral vectors is much better compared to the performance of the spectral heatmaps, across all five personality traits. One possible reason behind this outcome is that the feature selection process removes a large part of less informative behaviour information



(a) PCC results obtained by the spectral vector with different frequency alignment methods

(b) RMSE results obtained by the spectral vector with different frequency alignment methods

Figure 6.11: The results obtained by the spectral vector with different frequency alignment methods on VHQ dataset



(a) PCC results obtained by two spectral representations

(b) RMSE results obtained by two spectral representations

Figure 6.12: The results obtained by two spectral representations on VHQ dataset

before feeding spectral vectors to ANNs, and thus the reduced data is more compact and less noisy.

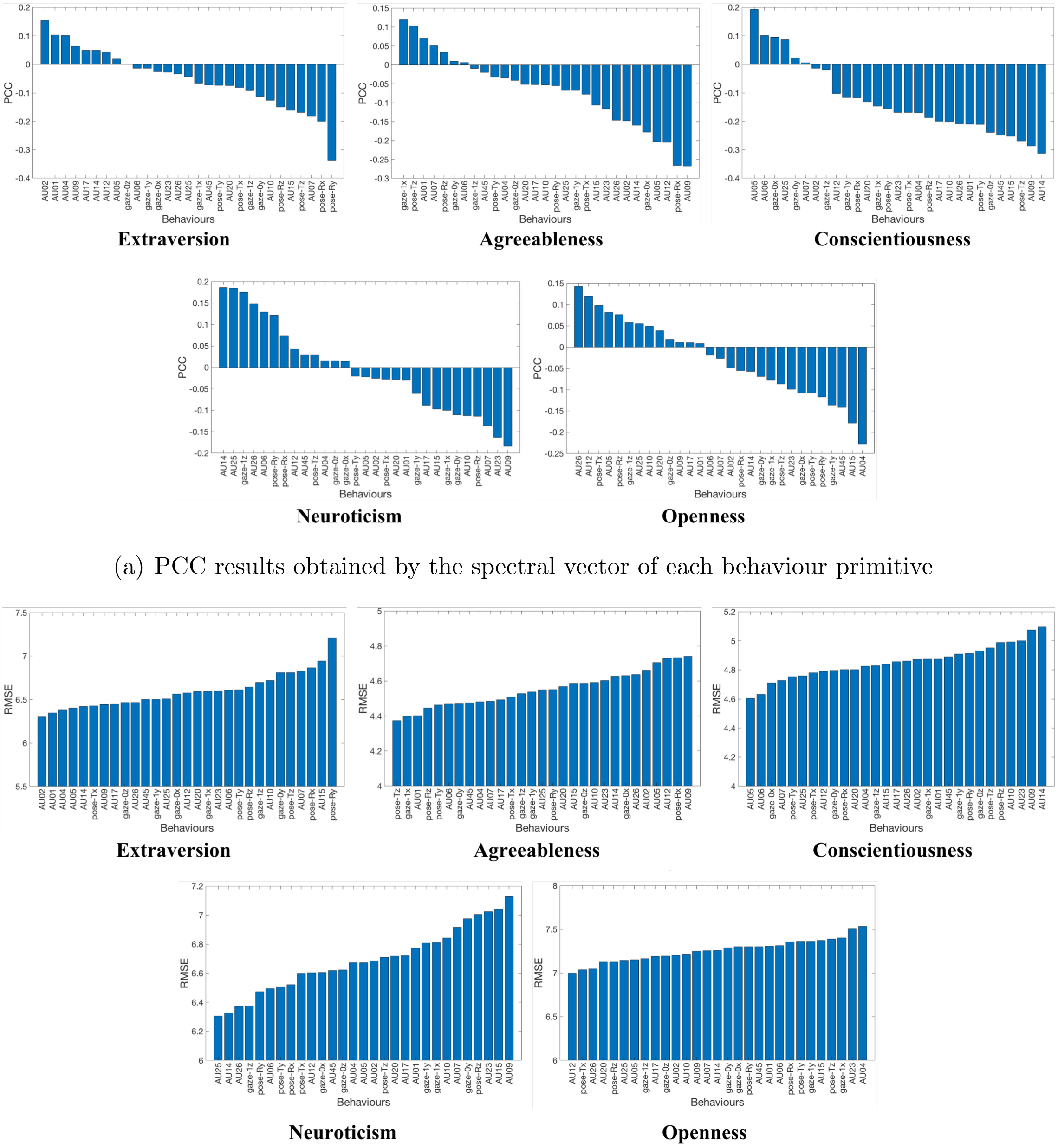
6.3.4 Analysis of facial behaviour primitives

This section independently evaluates the performance of each facial behaviour primitive on each personality trait estimation individually, using the proposed spectral vector. To do so, models were trained from the spectral vectors of each behaviour primitive, respectively. Here, Here, after the feature selection, the dimension of the final feature vector for each behaviour primitive (AU/gaze/head pose) ranges from 10 to 29. The results are achieved on the VHQ dataset by averaging the results obtained from three interview tasks. Please note that the reported results only indicate the relationship between automatically detected behaviours and personality traits intensities, which may be slightly different to the result achieved by using human annotated behaviour information. This is because the tool (OpenFace 2.0) that used for behaviour detection is not 100% accurate and the errors in detection may affect the personality traits analysis results.

As shown in Fig.6.13 and Fig.6.14, not only the five personality traits have completely different relationship with different facial behaviour primitives, but also each the same apparent personality trait and self-reported personality trait are reflected by different facial behaviour primitives. For self-reported personality traits estimation, AU02, AU01 and AU04 provided most important information for Extraversion trait; $gaze - 1x$, $pose - 1z$ and AU01 achieved the best predictions on Agreeableness trait; AU05, AU06 and $gaze - 0x$ predicted best Conscientiousness trait; AU14 and AU25 contain most relevant information for predicting Neuroticism trait; and finally, AU12, AU26 and $pose - 1x$ generated the best predictions for Openness trait. Meanwhile, For apparent personality traits estimation, AU04, AU09 and AU05 provided most important information for Extraversion trait; AU09, AU01 and AU04 achieved the best predictions on Agreeableness trait; AU05, AU07 and AU09 predicted best Conscientiousness trait; AU05 and AU04 contain most relevant information for predicting Neuroticism trait; and finally, AU12, AU25 and AU20 generated the best predictions for Openness trait. It can be observed that AU01 is informative to predict both types of Agreeableness trait; AU05 is informative to predict both types of Conscientiousness trait, and AU12 is informative to predict both types of Openness trait.

6.4 Summary

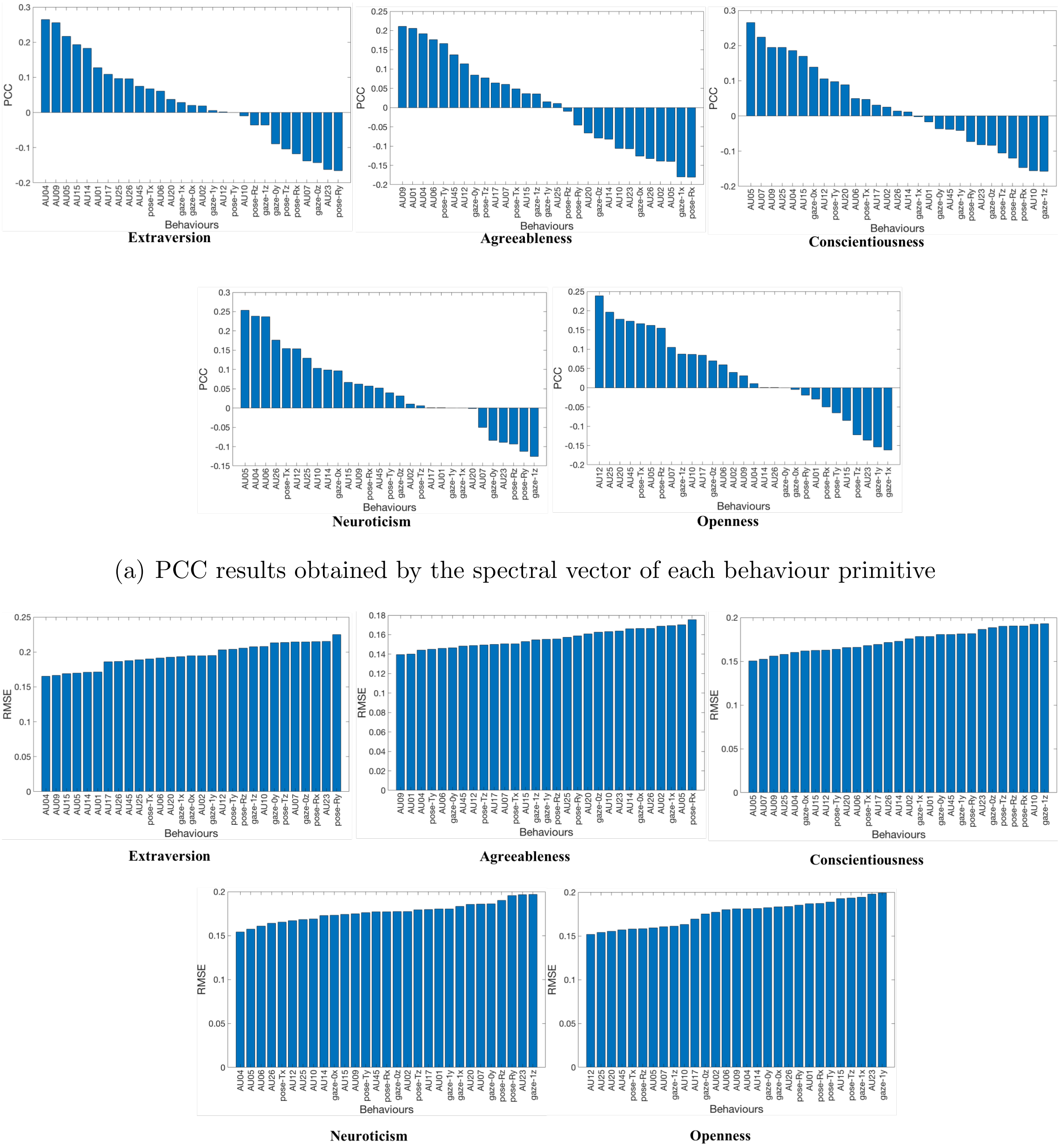
This section presents the detailed experimental settings and results achieved by the proposed person-specific representation and spectral representation on both self-reported and apparent personality traits analysis. Based on the results, it can be observed that person-specific representation generated better results than the spectral representation in personality traits recognition. Specifically, many factors such as the encoder pre-training strategies, video contents and length, the size of the time window, as well as the fusion strategy can largely affect the performance of the person-specific representation. For spectral representation, the spectral vector outperforms the spectral heatmap. This can be explained by the fact that the CNN model employed to learn spectral heatmap has the larger number of weights than the ANN used for spectral vector, and thus the limited training data is not enough to train a good CNN model. Another finding of this section is that different behaviour primitives (e.g., AUs, head poses and gazes) have completely different relationships with each personality trait.



(a) PCC results obtained by the spectral vector of each behaviour primitive

(b) RMSE results obtained by the spectral vector of each behaviour primitive

Figure 6.13: Personality traits estimation results obtained by spectral vector of each facial behaviour primitive on VHQ dataset.($gaze - 0x$, $gaze - 0y$, $gaze - 0z$ represent the gaze direction for left eye; $gaze - 1x$, $gaze - 1y$, $gaze - 1z$ represent the gaze direction for right eye; $pose - Tx$, $pose - Ty$, $pose - Tz$ represent the location of the head; $pose - Rx$, $pose - Ry$, $pose - Rz$ represent the rotation of the head. Please see (Baltrusaitis et al., 2018) for details.



(a) PCC results obtained by the spectral vector of each behaviour primitive

(b) RMSE results obtained by the spectral vector of each behaviour primitive

Figure 6.14: Personality traits estimation results obtained by spectral vector of each facial behaviour primitive on ChaLearn dataset.

Chapter 7

Personality-guided automatic depression analysis

As discussed in Chapter 1, the standard clinical depression assessment methods depending on verbal questionnaires are subjective and time-consuming. Thus it is necessary to develop automatic and objective assessment methods to aid depression monitoring and diagnosis. As reviewed in Sec.2.6, there is convergent psychological evidence suggested that depression is marked by non-verbal objective visual cues such as head movements, facial displays and gaze (S. Scherer et al., 2013; Goldstein, 1964), which can be automatically detected and analyzed without the intervention of clinicians. Therefore, building an automatic system based on such cues would not only provide an objective and repeatable evaluation but also would help alleviate key problems around cost and time requirements (Rana et al., 2019).

While there are many approaches devoted to automatically understanding depression status from non-verbal cues, an important factor that has not received much attention in the affective computing field is the relationship between personality traits and depression. As discussed in Sec.2.7, depression status is associated with personality traits (Kendler et al., 2006, 2007; Hettema et al., 2006; Klein et al., 2011; Krueger et al., 1996; Takahashi, Roberts, Yamagata, & Kijima, 2015). In short, these studies claimed that people with certain personality traits are more likely to be affected by depression. As a result, personality traits can act as strong priors and provide valuable information in addition to the facial behaviours for automatic depression analysis. Motivated by this, this chapter focuses on investigating: 1. whether the personality information can help ML models to learn better hypothesis of the depression status; 2. what are the better ways to combine facial behaviour descriptors and personality traits information for automatic depression severity estimation.

7.1 Novelty and contributions

Compared to previous automatic depression analysis studies that only used facial behaviour descriptors for prediction, this study explores several ways to combine facial behaviour descriptors with personality traits information for automatic depression analysis. Firstly, a new questionnaire-based audio-visual dataset is introduced (VHQ dataset). This dataset consists of recordings of participants undergoing 3 interview sessions. During each session, each participant was asked to answer a set of questions verbally based on one of the questionnaires: PHQ-9, GAD-7, or BFI-10. Meanwhile, the interview scenario can be one of the human face-to-face interaction, human video conference or human-to-robot interaction scenario. The Big-Five personality traits label for each participant is the original scores obtained by asking the participant to complete the BFI-44 questionnaire online. In this dataset, the average duration and standard deviation of duration, for each type of videos are 221.63s and 99.69s (BFI), 145.63s and 54.14s (GAD), 201.58s and 75.97s (PHQ), 568.9s and 97.97s (combined), respectively. The main contribution of the thesis for this dataset is that: 1. it describes the interaction methods that are used for the data collection; 2. it presented the statistics of the processed dataset (the final released dataset); 3. it provides several baselines depression analysis results as well as several personality-guided depression analysis results. To the best of author's knowledge, this is the first dataset that can be jointly used for the tasks of automatic personality traits, depression and anxiety analysis.

Then, this chapter investigates three ways to combine facial behavioural descriptors and personality descriptors (Sec.7.3.2). The experimental result shows that the feature level fusion of facial behaviour representation and personality traits descriptors improved the depression severity estimation performance under most interview scenarios. To the best of the author's knowledge, this work presents the first study that systematically investigates the feasibility of applying various personality traits information to vision-based automatic depression analysis.

7.2 Datasets

VHQ dataset: This section introduces a new audio-visual database called Virtual Human Questionnaire (VHQ) database (Jaiswal, Song, & Valstar, 2019; Jaiswal, Valstar, et al., 2019). The VHQ database consists of 165 face videos (55 videos for each session) recorded from a total of 55 participants. This database was collected to demonstrate the use of virtual human agents to simulate social interaction during which a user's facial behaviour can be analyzed. The study consisted of participants going through 3 interview sessions. In each session, 55 participants were asked to verbally answer a set of questions according to the following three standard questionnaires:

- Patient Health Questionnaire - 9 (PHQ-9): It is a self-administered questionnaire used for scoring 9 DSM-IV criteria for depression. It is widely used as a tool for monitoring the severity of depression. The range of total scores from this questionnaire varies from a minimum 0 (no depression) to a maximum 27 (severe depression). This is detailed discussed in Sec.2.5.

- Generalised Anxiety Disorder Assessment (GAD-7): It is a self-administered questionnaire consisting of 7 items. It has been widely used as a screening tool and for measuring the severity of generalized anxiety disorder. The range of scores from this questionnaire varies from a minimum 0 (no anxiety) to a maximum 21 (severe anxiety).

- Big Five Inventory (John et al., 1991): This is a self-administered questionnaire widely used for measuring 'big-five' personality traits that commonly used to describe personality and psyche (detailed discussed in Sec.2.1). Two versions of BFI

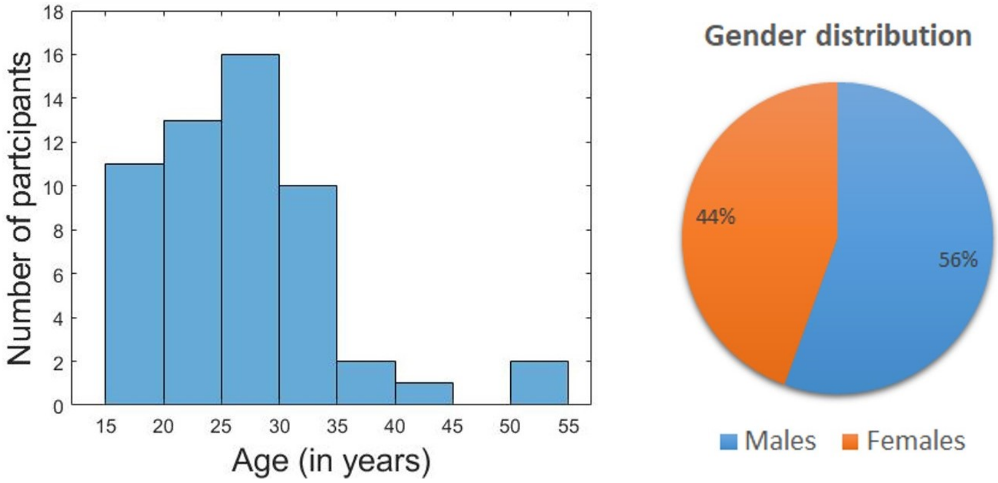


Figure 7.1: Age (left) and gender (right) distribution among the participants.

were used in this study. To conduct the interviews, the BFI-10 (Rammstedt & John, 2007) consisting of only 10 items was used. This was done because the original 44 item BFI questionnaire was considered too long for the interviews. However, to train our models, scores from the original 44 item BFI questionnaire (obtained using an electronic form) were employed.

During each interview session, the questions were taken verbatim from one of the above questionnaires. The interview sessions were conducted in one of three interaction modes: a real human interviewer sitting directly in front of the participant, a real human conducting interview over a video-conferencing link or a virtual human agent interviewer implemented based on the ARIA-VALUSPA Platform (Jaiswal, Valstar, et al., 2019). The purpose of these settings was to analyze the effect of different interaction modes on the participants' answers. The scores from the self-administered questionnaires were used as ground-truth labels for training and evaluation of models. The distribution of participants' age, gender and depression severity of this dataset are displayed in Fig. 7.1 and Fig. 7.2.

7.3 Methodology

This section explains the facial behaviour descriptors, personality descriptors, fusion strategies, feature selection methods and depression estimation models used for the study.

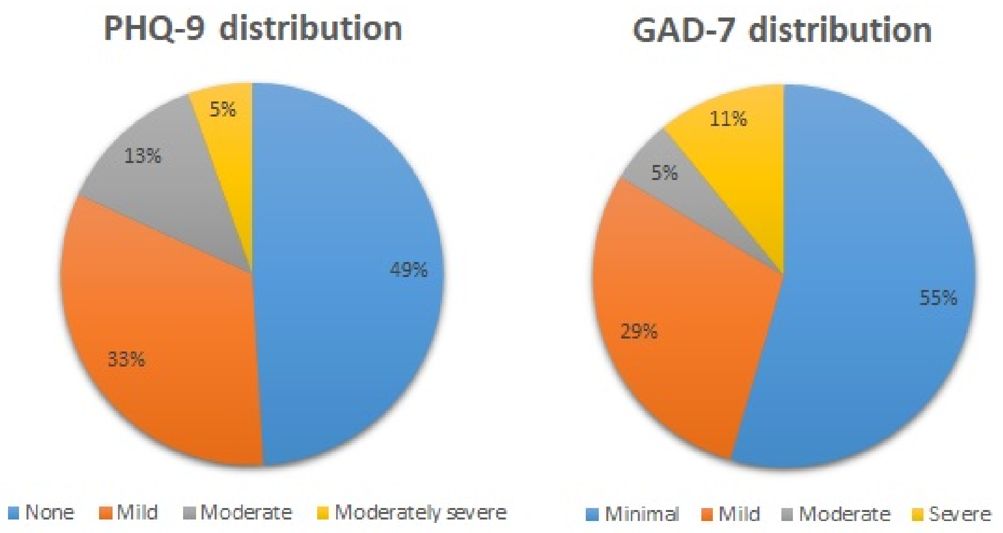


Figure 7.2: Depression (left) and anxiety (right) severity distribution among participants, according to PHQ-9 and GAD-7 scores respectively. For depression, the PHQ-9 scores were grouped into none (0-4), mild (5-9), moderate (10-14) and moderately severe (15-19) category. For anxiety, the GAD-7 scores were grouped into minimal (0-4), mild (5-9), moderate (10-14) and severe (15-21) category.

7.3.1 Non-verbal expressive facial behavioral descriptors

In this study, three video-level baseline facial descriptors are extracted from each video to provide facial behaviour information, including the proposed spectral vectors, spectral heatmaps and the person-specific representations. Here, only the automatically detected facial behaviour primitives (AUs, head poses and gazes) were used as the per-frame descriptor to produce the spectral vectors and spectral heatmaps, as they has much lower dimensions than the per-frame latent descriptor produced by DFNNs (According to experimental results, the spectral representations of facial behaviour primitives performed much better than spectral representations of the latent descriptor in both personality and depression analysis tasks.)

- **Spectral vectors**: Based on the process introduced in 5, the raw spectral vector of length 29×128 (29 facial behaviour primitives and 128 frequencies) for each subject is a 1-D vector obtained from multi-channel time-series signals consisting of 29 behaviour primitives. The final spectral vector (usually less than 100 components) is produced by applying CFS to the raw spectral vector, in order to reduce its dimensionality.

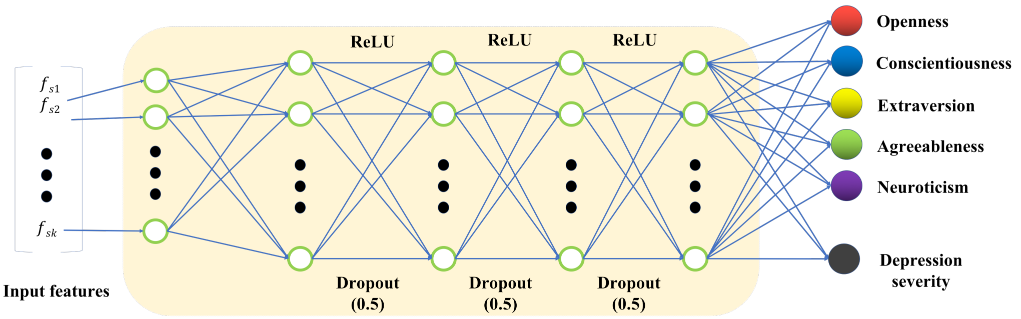


Figure 7.3: The network architecture for training-level fusion. This fusion strategy is achieved in a multi-task learning manner.

- **Spectral heatmaps**: Based on the process introduced in 5, the final produced spectral heatmap for each subject contains two matrices of size 29 × 128 (29 facial behaviour primitives and 128 frequencies), representing amplitude and phase information, which are obtained from $29 \times n$ multi-channel behaviour primitives time-series signals.

- **Person-specific representations**: Based on the process introduced in 4.3.2, this section also employs the personal video of each individual to train a set of PALs, whose weights and bias are then concatenated as a 1984-D person-specific behaviour descriptor for each person. The final person-specific vector (usually less than 50 components) is produced by applying CFS to the raw person-specific vector, in order to reduce its dimensionality.

7.3.2 Personality descriptors and fusion strategy

This section presents three ways of incorporating personality descriptors as the additional information to facial behaviour descriptors for automatic depression severity estimation.

- **Training-level fusion**: Firstly, the personality labels are employed to guide the depression severity estimation model's training. This can be achieved by adding five extra neurons in the output layer, paired with the big five personality traits (extraversion, conscientiousness, agreeableness, neuroticism and openness) labels during the training. This way, the model learns personality-related information in addition to the depression feature in a multi-task learning manner (shown in

Fig.7.3). Besides, this section also investigates the impact of each personality trait individually, by adding only a single extra output neuron in addition to the depression severity output during the training, which paired with each personality trait label, respectively.

- **Feature-level fusion with personality annotations**: The personality descriptor was represented in the form of scores corresponding to the big five personality traits labels obtained from the BFI-44 questionnaire. The resulting five dimensional personality feature vector was then combined with the facial behavioural descriptor obtained from video data, as the final descriptor.
- **Feature-level fusion with personality predictions**: The personality descriptor were represented in the form of scores corresponding to the big five personality trait predictions generated by the best system (please see Sec.7.4.2), which was concatenated with the facial behavioural descriptor obtained from video data, as the final descriptor.

7.3.3 Feature selection and normalization

The two feature-level fusion strategies would result in the high dimensionality of the combined descriptor (much higher compared to the amount of training data, i.e., in this chapter only 54 training examples, and the dimensions of original feature sets are usually more than 500). This may lead to models' overfitting. To avoid this, Correlation-based Feature Selection (CFS) (M. A. Hall, 1999) is again introduced to reduce the dimensionality, which selects a subset of features that are most relevant for our task, where Pearson's linear correlation coefficient is employed to measure the correlations. Then, each selected feature X was normalized by computing its Z-score given by the following equation:

$$Z = \frac{X - \mu}{\sigma} \tag{7.1}$$

where μ and σ represents the mean and standard deviation of the feature values from the corresponding dimension, over the entire training data.

7.3.4 Depression analysis model

To train regression models for depression severity (PHQ-9 scores) estimation, the ANN architecture described in Sec.4.3.2 has also been employed here to process behavioural and personality descriptors. The utilised ANN architecture consists of four hidden layers and an output layer that yields the estimated depression severity. Here, for the training level fusion setting (Sec.7.3.2), the output layer has either six or two neurons during the training, where five or one of them paired with either big five personality traits annotations or one of them. Meanwhile, the other output neuron corresponds to the depression severity. Same to the settings described in Sec.4.3.2, dropout with probability 0.5 and ReLU activation are attached at each hidden layer. MSE is employed as the loss function for training and RMSProp gradient descent method is used as the training method.

7.4 Experimental Settings

7.4.1 Implementation details

Datasets: As described in Sec.7.2, the VHQ database contains both personality traits and depression severity annotations for each individual, and thus it is suitable to the purpose of this Chapter, i.e. applying personality information to automatic depression analysis.

Pre-processing and training detalis: Following the same settings described in Sec. 6.2, this section also employed the OpenFace 2.0 toolkit to obtain aligned face images and facial behaviour primitives (e.g., AUs, gaze directions, and head poses). Meanwhile, all training settings of ANN models are the same to the settings described in Sec. 6.2.4, i.e., this thesis conducts leave-one-subject-out cross-validation on VHQ dataset. When implementing the training-level fusion (multi-task learning), both corresponding depression severity and personality traits labels were used to compute the loss.

Evaluation Metrics: In this study, RMSE, PCC and CCC measurements introduced in Sec.6.2.5 were also employed to evaluate the depression severity estimation performance. In addition, mean absolute error (MAE) is also utilised, which is defined as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |f_i - y_i| \tag{7.2}$$

where f_i is the predicted depression severity and y_i is the corresponding ground-truth.

7.4.2 Experiments

To systematically investigate the feasibility of applying personality information to automatic depression analysis, a series of experiments have been conducted.

- Firstly, the proposed three baseline representations, e.g., spectral vectors, spectral heatmaps and person-specific representations were produced to generate baseline results, where spectral vectors and person-specific representations were processed by the CFS and then fed to ANN models. Meanwhile, spectral heatmaps were processed by 1-D CNN (introduced in Sec.5.5) models. In these experiments, only facial behaviour dynamic representations were used to predict depression levels.

- Secondly, systems of the training-level fusion (multi-task learning) were compared on the VHQ dataset. Six systems were evaluated, including the five multi-task learning systems each of which learn the depression severity and a personality trait intensity, respectively, as well as multi-task learning of the depression severity and all five personality traits intensities. In the test phase, only the outputs corresponding to the depression severity are used to predict depression levels.

- Thirdly, systems of feature-level fusion with personality labels were conducted and evaluated on the VHQ dataset. Firstly, each personality trait label was attached to the facial behaviour descriptor to construct the behaviour-personality descriptor, respectively, which is then fed to ANN for the depression severity estimation task. After that, this section also adds annotations of all five personality traits to the behaviour descriptor to construct the behaviour-personality descriptor for depression severity estimation.

- Finally, representations were produced and evaluated on the VHQ dataset under the feature-level fusion with the personality predictions setting (Sec.7.3.2). This experiment firstly attaches each personality trait prediction to the facial behaviour descriptor to construct the behaviour-personality descriptor for depression severity estimation, respectively. Then, five personality traits predictions were jointly attached to the behaviour descriptor for depression severity estimation.

7.5 Results and discussions

7.5.1 Facial behaviour descriptor VS behaviour-personality descriptor

This section specifically evaluates both advantages and disadvantages of the proposed fusion strategies and compares them against the baselines. All the reported results are the average results over three interview tasks of the VHQ dataset.

Baselines: Table.7.1 compared the average results yielded by three baselines, where the spectral vector achieved the best performance, and person-specific representations outperformed the spectral heatmap. A potential reason for the spectral heatmap's poor performance can be explained as the limited number of training data (54), which is not enough to train a good CNN model, i.e., the CNN model is underfitting. Since the spectral vectors and person-specific representations are 1-D vectors and processed by feature selection, their dimensions are much smaller than spectral maps. As a result, they can be learned by ANNs with fewer parameters that are needed to be optimized. However, while person-specific representations only encodes a single scale of short-term dynamics, the spectral vector encodes multiple-scale dynamics of facial behaviour primitives, which can be the reason for the best depression analysis performance achieved by the spectral vector baseline.

Training-level fusion: Fig.7.4 compared the results achieved by two baselines (spectral vector and person-specific representations) against their corresponding training-level fusion systems (multi-task learning). It can be observed that multitask learning

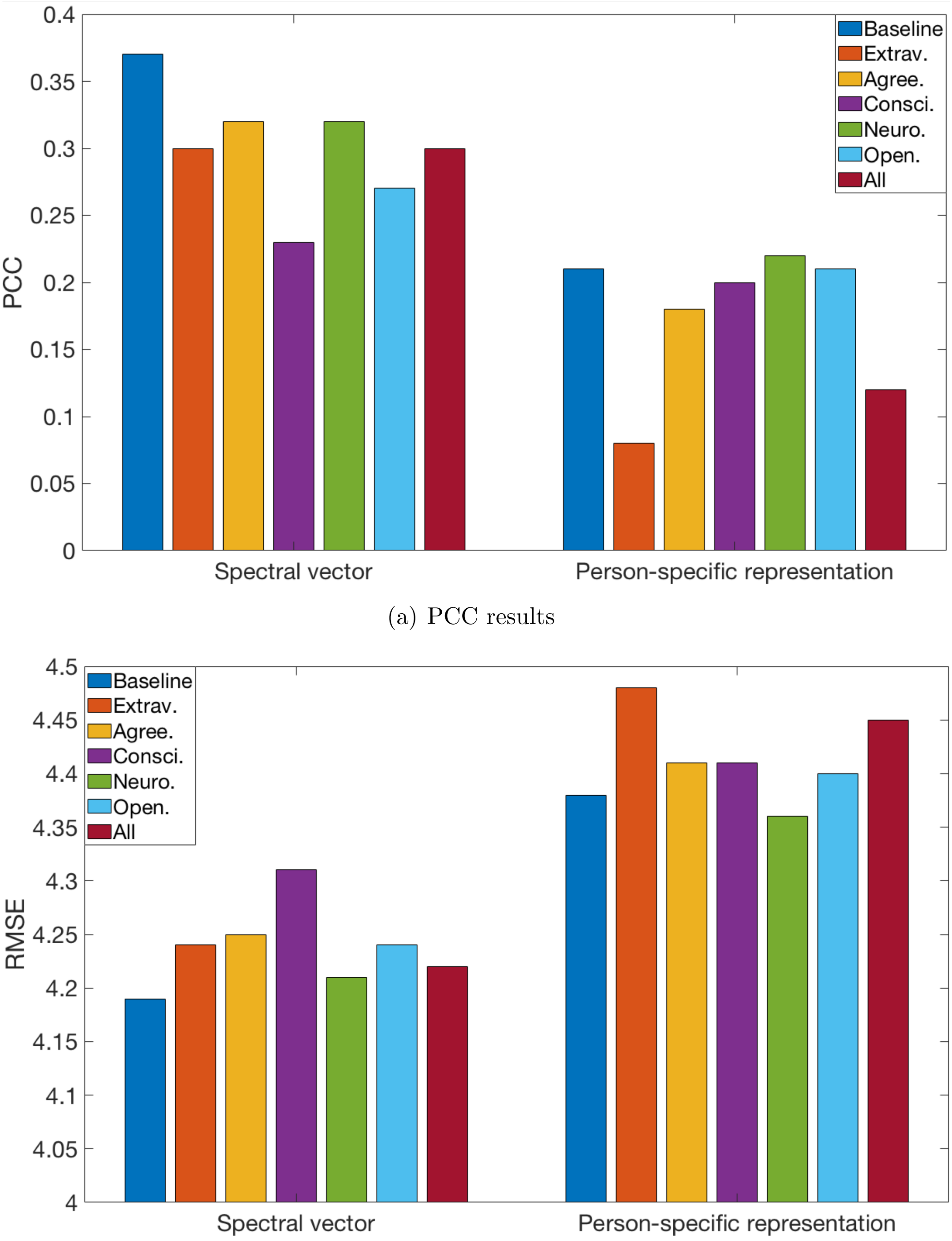
Descriptor	MAE	RMSE	PCC
Spectral Heatmap	3.79	4.75	0.09
Spectral vector	3.13	4.19	0.37
Person-specific	3.29	4.38	0.21

Table 7.1: The average results of leave-one-subject-out cross-validation achieved by three baselines on three tasks of VHQ dataset

depression severity and personality traits generally yielded worse depression severity estimation performance. In particular, all spectral vector-based training-level fusion systems could not achieve a comparable result to the baseline. Meanwhile, for person-specific representations-based systems, only multitask learning depression with neuroticism trait achieved a slightly better result than the baseline. These results indicate that using a simple ANN to multitask learn personality information is not a superior way to learn better depression models.

Feature-level fusion: Fig.7.5 and Fig.7.6 compared the results achieved by spectral vector and person-specific behaviour representations against their corresponding behaviour-personality descriptors, i.e., feature-level fusion of behaviour descriptor and personality labels/predictions, where the personality predictions were obtained from the best systems presented in Sec.6.3.1. As we can see from Fig.7.5, when individually adding a label or prediction of agreeableness, conscientiousness or openness trait as an additional descriptor to spectral vectors, the produced results are the same to the baseline. This is because the attached 1-D personality features have not made any difference in the feature selection process, i.e., they haven't been selected during this process. In other words, the final selected feature sets of these three settings are exactly the same to the feature set selected from baselines. Meanwhile, the extraversion and neuroticism traits' labels/predictions either slightly enhanced or degraded the baseline performance. Importantly, when combining all five personality traits' labels/predictions, the depression severity performance has been increased dramatically.

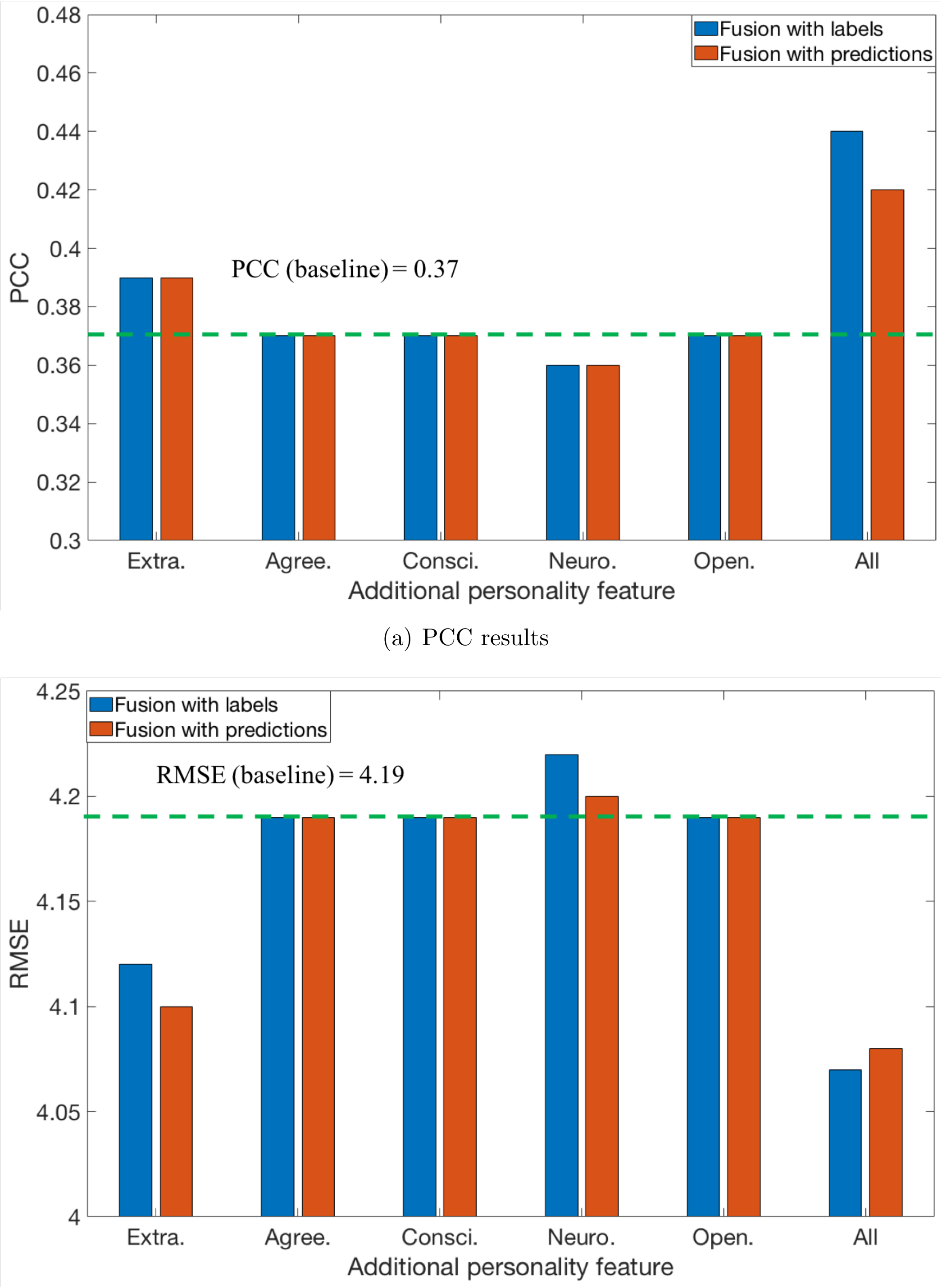
Meanwhile, according to Fig.7.6(a), individually adding the label or prediction of extraversion, agreeableness, conscientiousness or openness as the additional descriptor to



(a) PCC results

(b) RMSE results

Figure 7.4: The average results achieved by training-level fusion strategies on three tasks



(a) PCC results

(b) RMSE results

Figure 7.5: The average results achieved by feature-level fusion strategies on three tasks using spectral vectors

person-specific behaviour descriptors yielded the same result to the baseline result due to the same reason explained above. In contrast to the spectral vector, adding neuroticism traits' labels/predictions slightly enhanced the person-specific representation baseline performance. Again, when combining all five personality traits' labels/predictions, the depression severity estimation performance has been enhanced.

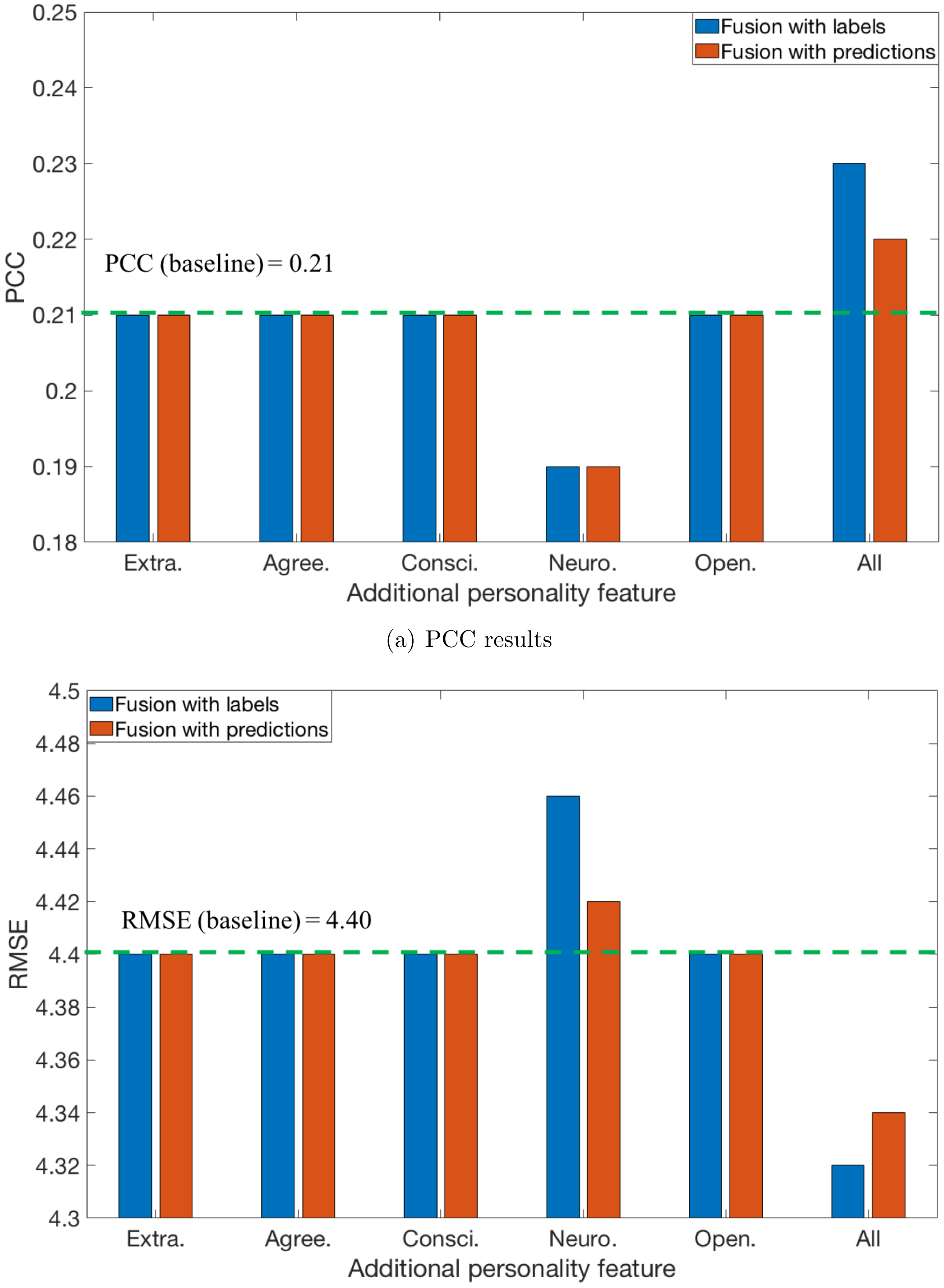
In summary, both personality traits predictions that have been automatically detected using the proposed approaches, and personality traits labels, can positively impact depression severity estimation performance under some tested conditions, indicating that personality traits can provide useful clues in addition to the facial behaviour descriptors for inferring depression levels. However, using personality traits as additional targets for multitask learning ANNs may not be an efficient way to fuse personality and behaviour information for depression analysis.

7.5.2 Interactive Behaviour Studies

Analysis of facial behaviour primitives

This section independently evaluates the performance of each facial behaviour primitive on depression severity estimation. To do so, models were trained and tested using the spectral vectors of each behaviour primitive, respectively.

As shown in Fig.7.7, individually using spectral vectors of gaze-0y, AU09, and AU45 intensities achieved decent estimation performance, where PCC results are over 0.2, and RMSE results are less than 4.5. Particularly, gaze-0y yielded the best result among 29 behaviour primitives, with PCC more than 0.25 and RMSE less than 4.4. The predictions obtained from the rest head poses and gaze directions have lower correlations with the ground-truth, as four features corresponding to these ranked at the bottom part of both CCC and RMSE. Besides, this section also reports the performance of each human behaviour primitive on depression severity estimation achieved in AVEC 2013 dataset. According to Fig.7.8, it is clear that AU15, AU17, AU12, AU04 and AU09 intensities achieved the best performance, where AU15 generated the best result among all behaviour primitives, with PCC of more than 0.65. Apparently, the evaluations on two datasets re-



(a) PCC results

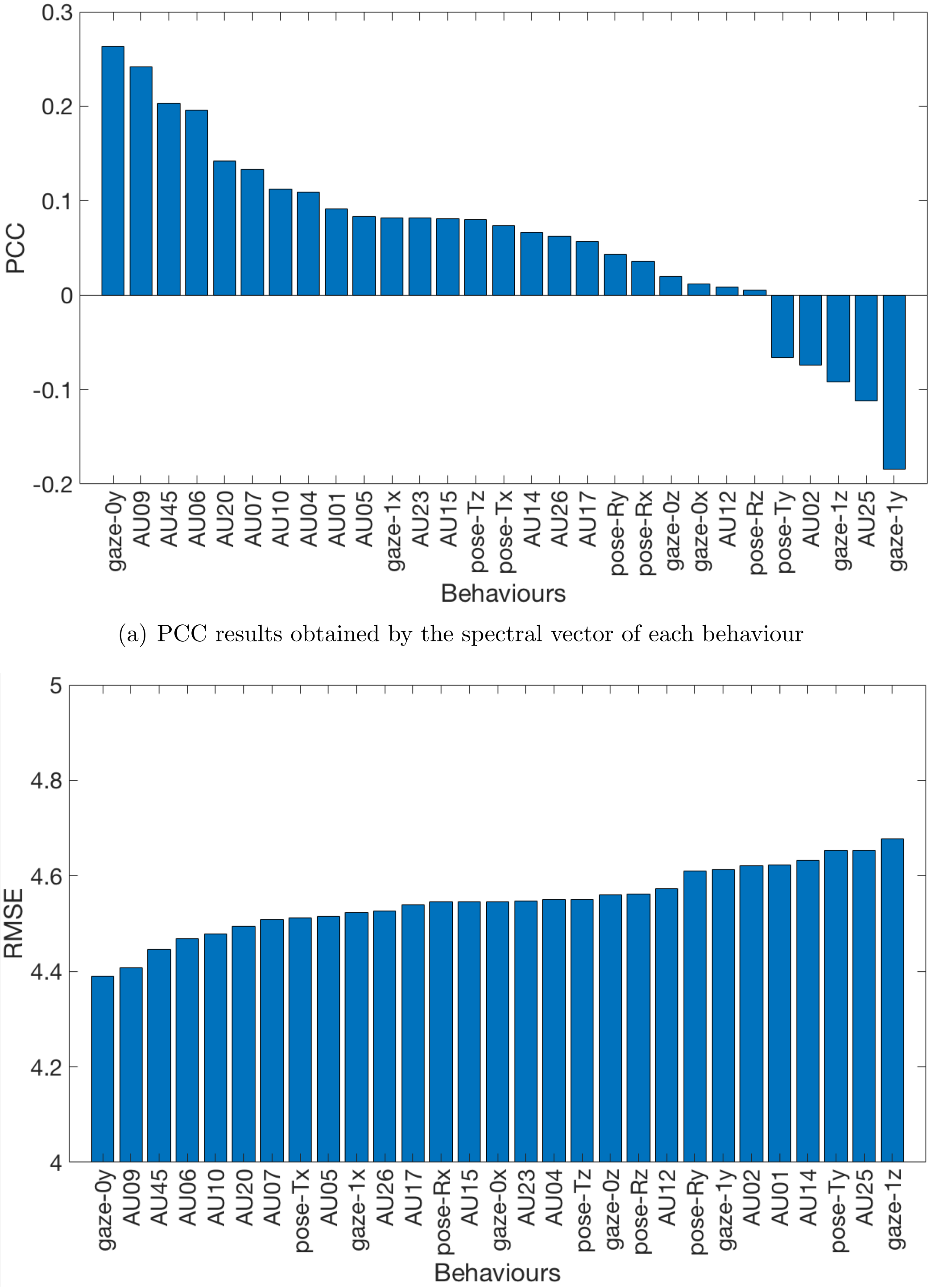
(b) RMSE results

Figure 7.6: The average results achieved by feature-level fusion strategies on three tasks using person-specific representations

sulted in different conclusions in terms of the most informative behaviour primitives on depression severity estimation, except that both evaluations showed that most head poses and gaze directions behaviours are less informative to depression, as they achieved low PCC and high RMSE results on both datasets. The different results in most informative behaviour primitives may be due to that: 1. the tasks and scenarios are entirely different in two datasets; 2. their labels were recorded by different questionnaires, which may lead them not to measure the same aspect of depression.

Task contents

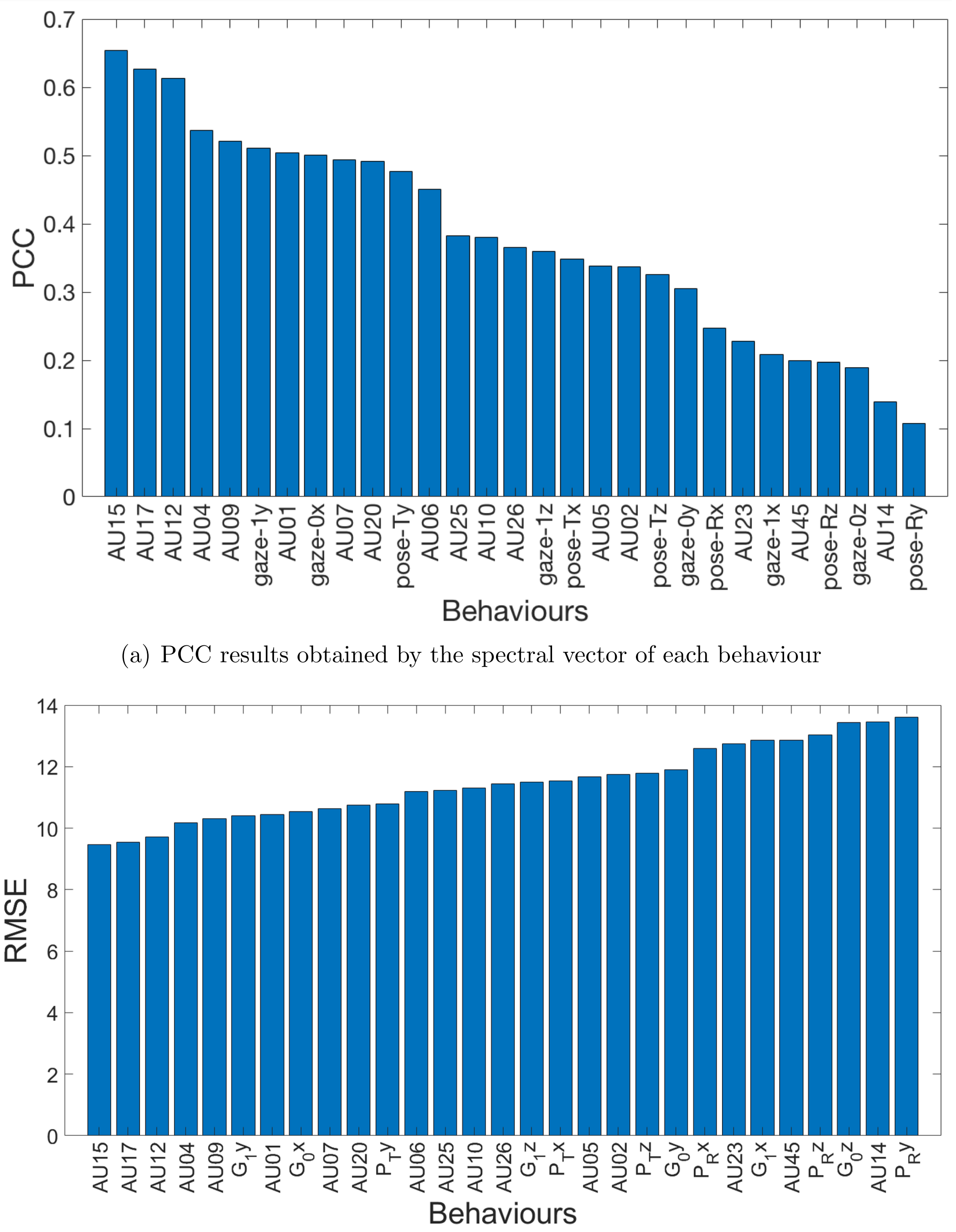
Since the differences in task contents resulted in differences in personality traits prediction performance, this section also investigates the influence of different interview tasks on depression severity estimation using baselines and two feature-level fusion conditions. As illustrated in Fig.7.9, it is clear that the depression severity estimation results varied a lot when using videos from different interview tasks. Here, only the best result achieved by each condition is reported (all results were achieved using spectral vectors). In VHQ dataset, the GAD-7 questionnaire-based interview task seems triggered the most depression-informative behaviours, with PCC of 0.49, 0.36, 0.59 under three conditions, while the BFI-10 questionnaire-based interviews produced the worst prediction performance, with PCC of 0.24, 0.23, 0.34, respectively, for depression severity estimation. These large differences indicate that the task contents have great impacts on depression analysis performance. This is because different tasks triggered different behaviour responses in people's faces. Meanwhile, the feature-level fusion with personality traits labels generated the best depression analysis performance with PCC of 0.59 on GAD-7 task, showing that combining behaviour descriptors with self-reported personality traits information can increase the upper-bound of the depression severity estimation application.



(a) PCC results obtained by the spectral vector of each behaviour

(b) RMSE results obtained by the spectral vector of each behaviour

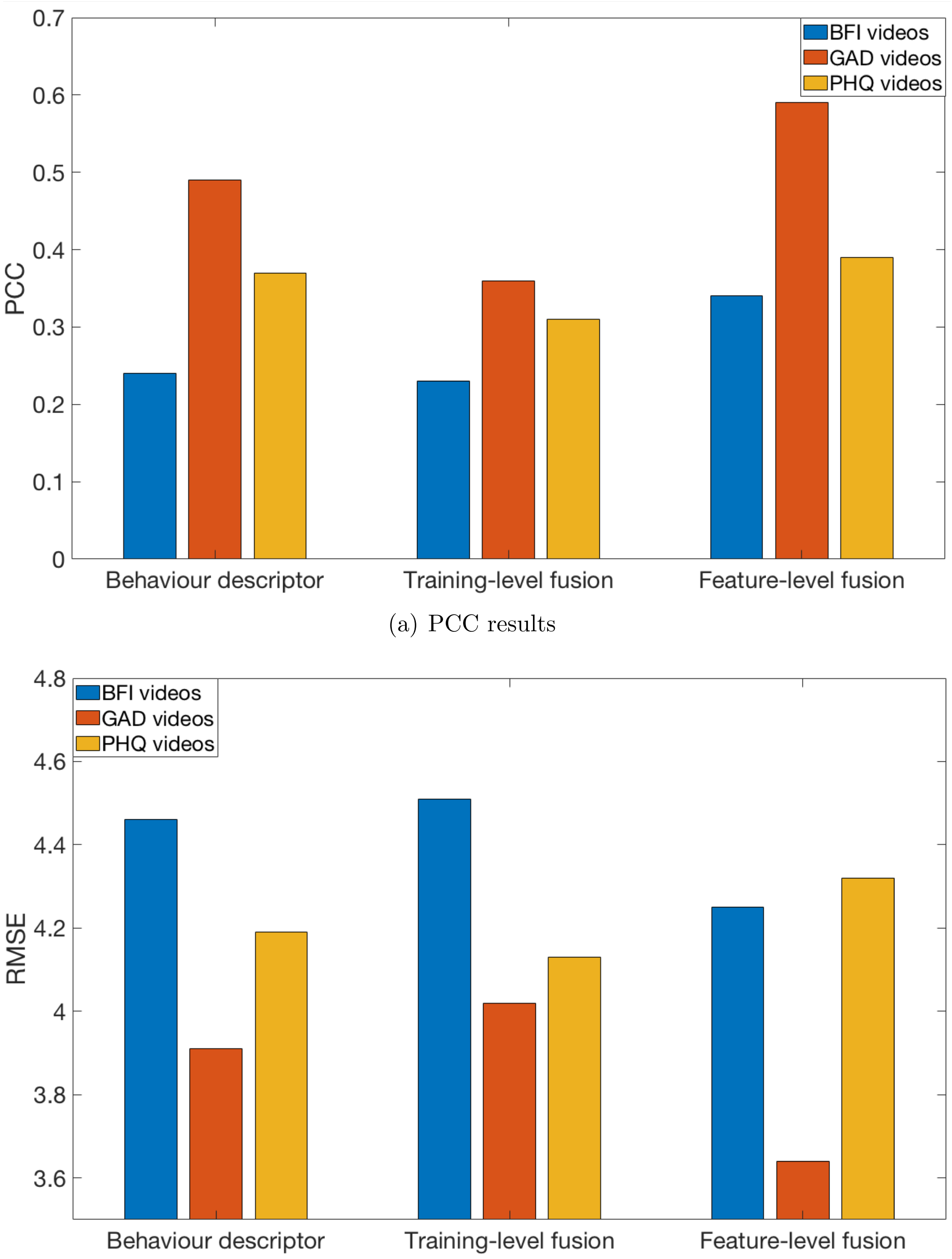
Figure 7.7: Depression severity estimation results obtained by spectral vector of each facial behaviour primitive, on VHQ dataset. $gaze-0x$, $gaze-0y$, $gaze-0z$ represent the gaze direction for left eye; $gaze-1x$, $gaze-1y$, $gaze-1z$ represent the gaze direction for right eye; $pose-Tx$, $pose-Ty$, $pose-Tz$ represent the location of the head; $pose-Rx$, $pose-Ry$, $pose-Rz$ represent the rotation of the head. Please see (Baltrusaitis et al., 2018) for details.



(a) PCC results obtained by the spectral vector of each behaviour

(b) RMSE results obtained by the spectral vector of each behaviour

Figure 7.8: Depression severity estimation results obtained by spectral vector of each facial behaviour primitive, on AVEC 2013 dataset.



(a) PCC results

(b) RMSE results

Figure 7.9: Best results obtained from three tasks using different fusion strategies

7.6 Summary

This section introduces a new audio-visual dataset that can be used for automatic personality, depression and anxiety analysis studies, based on which it investigates how to use personality information for depression severity estimation. Differing from most existing approaches that only used behavioural descriptors to predict depression severity, this section focuses on exploiting the underlying relationship between depression and personality traits. In particular, several standard ways that combine behavioural descriptors and personality information are evaluated for automatic depression severity estimation. The results show that even simply concatenating behavioural descriptors with self-reported personality traits or personality traits predictions can improve the depression severity estimation performance. This finding is consistent with previous biological and psychological studies(Summarized in Sec. 2.7) that depression is well associated with one's personality traits. The main limitation of this Chapter is that only several standard fusion methods were investigated. In other words, superior results may be achieved if more advanced methods are extended to combine behavioural and personality information.

Chapter 8

Conclusions and future work

This chapter summarizes the proposed person-specific adaptation approach and spectral analysis approach, making conclusions based on the obtained video-based automatic personality and depression analysis results. Besides, this section also discusses the limitations of the proposed approaches and gives directions for the future work.

8.1 Person-specific facial dynamics modelling

The personality traits primarily focus on evaluating the aspects of personality, which are relatively stable over time but differ across individuals (Kassin, 2003). Thus, this thesis first proposed a person-specific adaptation approach to encode person-specific video-level facial dynamic descriptors for automatic personality traits estimation. The person-specific approach builds on the assumption of a sequence dynamic encoding approach: dynamic image algorithm (Bilen et al., 2017). To evaluate the capability of such approach on facial dynamics modelling, this thesis extended the dynamic image algorithm to the face domain with facial shape information (facial landmarks), allowing a single raster image to encode short-term temporal evolution from a face image sequence. The main advantage of this approach is that it encodes facial dynamics in the context of the face, and the produced DFR is length-independent.

Since dynamic facial representations achieved better performance than static face images on dimensional affect estimation. This thesis further proposed to self-supervised learning a similar dynamic representation that also aims to infer short-term facial dynamics but from a single face image. In particular, this thesis proposed a Rank Loss allowing networks to learn generic facial dynamics from unlabelled face videos. As a result, the well-trained generic network (DFNN) can infer a DR from any previously unseen test image, which effectively summarizes the facial dynamics surround it. Based on this self-supervised learning approach, the person-specific adaptation approach starts with pre-training an emotion-guided encoder, and then employed the rank loss to train generic network in a self-supervised manner, which learns generic short-term facial dynamics from a set of unlabelled face videos. The generic model is then frozen, and a set of intermediate filters are incorporated into this architecture. The proposed self-supervised learning is then resumed with only person-specific videos. This way, the learned filters' weights will be person-specific, making them a valuable source of person-specific dynamic modelling. Finally, the weights of the learned filters are concatenated as a person-specific representation, which can be directly used to predict the personality/depression status without needing other parts of the network.

The experiments were first conducted to evaluate the dynamic encoding abilities of the proposed approach. In particular, the dynamic representations generated by sequence-based approach and the self-supervised learned generic net were tested on frame ranking task and dimensional affect estimation task, whose results suggest the following conclusions: 1. the DFNN trained with the proposed rank loss function generalizes better ranking ability to unseen face images than models trained with reconstruction loss using pre-defined representations, demonstrating the ability of the proposed still image-based self-supervised learning approach can properly learn facial dynamics without the need of human annotations; 2. the DFR generated by RankSVM and multi-scale DR generated by DFNN can effectively encode spatio-temporal facial patterns, as both of them achieved better dimensional affect estimation results than only using static images as the input; 3. The time-window of the frames used for training decides the temporal scale of learned dynamics. The results indicate that dynamics at each scale can provide some useful and unique information for valence and arousal estimation. This was evident from the results where the fusion of multi-scale temporal dynamics provided better performance than single-scale models.

Meanwhile, the experiments were also conducted on automatic self-reported/apparent personality traits estimation tasks to evaluate the impacts of different settings on the proposed person-specific approach. The results suggested that: 1. While the conducted experiments kept the initial learning rate of the DFNN's encoder during the self-supervised training, the emotion guided pre-training still enhanced the automatic personality traits prediction performance. The hypothesis is that the pre-training may provide good initial weights for the encoder; 2. facial behaviours triggered by different tasks may have considerable impact on the performance of personality prediction models; 3. longer videos that contain more tasks, can force PALs to learn behaviour patterns which occur frequently under different conditions i.e., patterns stable over time are indicative of personality; 4. facial dynamics at each scale can provide some useful and unique information regarding personality. This was evident from the results where the fusion of multi-scale temporal dynamics provided better performance than single-scale models; 5. person-specific approach achieved similar apparent personality traits estimation performance to the current state-of-the-art on the Chalearn dataset, and the state-of-the-art self-reported personality traits estimation result on the VHQ dataset.

Limitations: In this thesis, person-specific facial dynamics are represented by the unique weights and biases of PALs. However, there are several limitations of this approach: 1. the differences in weights and biases of only a set of intermediate layers may not be enough to represent the differences in people's facial behaviours; 2. the proposed approach can only encode short-term facial dynamics as the frame ranking performance degrades heavily when the time-window is longer than 1 s, i.e., it lacks the ability to encode long-term facial dynamics; 3. it requires training a set of intermediate layers for each test video at the inference stage, which is very time-consuming; 4. the UNet used in this thesis may not the optimum network for personality/depression analysis tasks; 5. this thesis only evaluated the person-specific representation of facial behaviours, while other modalities, e.g. audio, body poses, EEG, etc. may also informative to personality/depression analysis;

6. there is a lack of publicity available large-scale audio-visual self-reported personality datasets, which restrict the application of deep learning approaches in this field.

8.2 Multi-scale facial dynamics modelling

The spectral approach is proposed to encode multi-scale long-term facial dynamics. It firstly employs Fourier Transform to convert multi-channel time-series signals (produced by multiple per-frame facial behaviour primitives) to the frequency domain as spectral signals. Each spectral component encodes a unique scale of dynamic (frequency) information of the entire video. As a result, the produced spectral signals contain multi-scale video-level facial temporal information of multiple behaviour primitives. However, due to the variation in original videos' length, their corresponding spectral signals' sizes are also different. Then, two frequency alignment methods were proposed to deal with such problems and generate two length-independent spectral representations, i.e., spectral heatmap and spectral vector, which encode multi-scale video-level temporal information, and can be easily learned by CNNs or ANNs.

This approach was applied to the task of automatic personality traits estimation. In the thesis, three measurements: PCC, RMSE and ACC, were used to evaluate the personality recognition performance. Although only ACC was widely used in previous related studies, the additional usage of PCC and RMSE can evaluate not only the errors but also the correlation between predictions and ground-truth. The generated results indicate the following conclusions: 1. The spectral approach can significantly reduce the dimension of the original time-series data while retaining most information; 2. each personality trait is associated with a unique set of behaviour primitives, where $AU02$, $AU01$ and $AU04$ provided most important information for the Extraversion trait; $gaze-1x$, $pose-1z$ and $AU01$ achieved the best predictions on the Agreeableness trait; AU05, AU06 and $gaze-0x$ predicted the best Conscientiousness trait; $AU14$ and $AU25$ contain most relevant information for predicting Neuroticism trait; and finally, $AU12$, $AU26$ and $pose-1x$ generated the best predictions for Openness trait; 3. The proposed two frequency alignment methods achieved similar results while the spectral vector showed superior performance than spectral heatmap for personality tratis analysis; 4. The person-specific representation produced better performance than the spectral vector in the automatic personality traits estimation task. The underlying reason can be explained that person-specific representation specifically encodes personalized dynamics that stable over time for the given person but different from others, which has not been considered in spectral vector. 5. The combination of spectral vector and person-specific representation achieved the best results than using either of them, showing that both multi-scale facial dynamics and person-specific short-term facial dynamics encoded in the proposed two approaches contain unique information that is informative to personality traits estimation.

Limitations: While experiments showed that the spectral approach can significantly reduce the dimension of the original data and retain most information, a limitation of this approach is that it requires manual selection of the settings and heavy pre-processing. Meanwhile, the spectral approach only used the automatically detected AUs, gaze and head pose as the frame-level facial representation, ignoring some other useful information (e.g., microexpressions, speech, etc.). In addition, this approach may still discard some useful dynamics (frequencies) during the frequency alignment process. Finally, the performance of using CNN to train from spectral heatmaps can be potentially improved if more training data is available, i.e., the lack of data problems prevents training a good deep model.

8.3 Personality-guided depression analysis

Chapter .7 investigated several ways of applying personality traits information to depression analysis. This Chapter combines video-level behaviour descriptors (spectral representations and person-specific representations) with different self-reported personality traits descriptors, including training-level fusion (multi-task learning personality traits and depression levels), feature-level fusion with personality traits labels, and feature-level fusion with personality traits predictions. The experiment results showed that: 1. the spectral vector baseline generated superior results than the person-specific representation in the automatic depression severity estimation task. This can be explained by that only personalized dynamics may not carry enough information for depression analysis, as depressed individuals can show some similar generic behaviours, e.g., lack of facial expressions. Therefore, spectral vectors that encode both generic and personalized multi-scale facial dynamics, can provide more valuable clues; 2. the most informative facial behaviour primitives on depression severity estimation can be different on different datasets but AU09 showed decent results on both datasets; 3. multi-task learning personality traits and depression using simple ANN model may not be a proper way to enhance the depression severity estimation performance; 4. the feature-level fusion of personality and behaviour descriptors provided clear improvements in the depression severity estimation performance, indicating that the personality information can indeed help the automatic depression analysis from the perspective of affective computing; 5. both personality traits labels and predictions can improve the depression analysis performance, where the labels-based fusion generated even better results. This is because the personality predictions are not 100% accurate, and their errors may further negatively impact the depression analysis.

Limitations: While the experimental results showed that personality information can help the automatic depression analysis, this thesis only evaluated several basic fusion strategies. Meanwhile, the VHQ dataset only contains a limited number of valid subjects (55) and the dimensions of the final feature vectors used in experiments of this thesis are relatively large (more than 15). Thus, the models' training may suffer from the curse of dimensionality. Thus, the models' training may suffer from the curse of dimensionality. Specifically, the CFS used in the thesis may not be the best choice to reduce the dimensions of the person-specific representation and the spectral representation. We assume that there is a potential to further improve their performance if other proper dimensionality reduction approaches are employed, such as other feature selection methods or proper Principal component analysis (PCA) methods. In Chapter 7, only depression severity estimation studies were conducted. However, in some real-world applications, a binary prediction (depressed or not) and a categorical prediction (minimal depressive symptoms, mild depressive symptoms, moderate depressive symptoms and severe depressive symptoms) may also useful.

8.4 Future work

According to the limitations summarized above, the correlation and error between personality/depression predictions and ground-truth are not optimal. To further improve the performance, one main future direction is to address the drawbacks of the proposed person-specific approach and spectral approach. For person-specific approach, the future work will focus on: 1. using a unique network that has a unique typology and weights to better represent one's complex person-specific behaviours, and finding good target functions to encode long-term facial behavioural patterns; 2. combining the current work with audio-visual and verbal information to predict personality, i.e., investigating the optimum network structure for jointly learning the personality traits from audio-visual and verbal signals in a self-supervised manner. Regarding the spectral approach, one potential future work will be building a spectral pooling method inserted to CNNs for end-to-end training, allowing super long and variable length time-series signals to be directly fed to CNN models, i.e., multimodal audio-visual data (or even other data such as EEG) can be learned in an end-to-end manner. Another future work will be devoted to improving the frequency alignment algorithm, avoiding the important frequency information being lost during the frequency alignment process. Finally, in terms of the personality-guided depression analysis, a potential research direction is to apply and extend more advanced fusion algorithms such as attention mechanisms for the better fusion of behaviour and personality descriptor. In addition, for both approaches, this thesis only employed CFS to reduce the dimensionality of the spectral/person-specific features. It can be assumed that introducing more advanced dimensionality reduction algorithms would further improve the performance.

Besides improving the proposed approaches themselves, the future work also aims to extend the current VHQ dataset with: 1. more subjects; 2. more recordings for each participant; 3. data recorded from multiple sensors (audio, video and physiological signals such as skin conductance, heart rate, etc.); 4. more scenarios. It can be expected that an enhanced dataset is collected, the curse of dimensionality issue of the proposed approach can be better addressed. Moreover, with more modalities and more data, well-trained models could have stronger generalization capability.

Finally, this thesis only applied the proposed approaches to personality/depression intensities estimation. However, in some real-world applications, binary predictions or categorical predictions may be more useful. Thus, from the perspective of application, future work can be devoted to: 1. applying the proposed approaches to generate binary or categorical predictions; 2. applying the proposed approaches to analyze other human internal states such as anxiety, state of the mind, etc.

References

Al-gawwam, S., & Benaissa, M. (2018). Depression detection from eye blink features. In *2018 ieee international symposium on signal processing and information technology (isspit)* (pp. 388–392).

Al Jazaery, M., & Guo, G. (2018). Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Transactions on Affective Computing*.

Al-Samarraie, H., Eldenfria, A., & Dawoud, H. (2017). The impact of personality traits on users' information-seeking behavior. *Information Processing & Management*, *53*(1), 237–247.

Ambridge, B. (2014). *Psy-q: You know your iq-now test your psychological intelligence.* Profile Books.

Anis, K., Zakia, H., Mohamed, D., & Jeffrey, C. (2018). Detecting depression severity by interpretable representations of motion dynamics. In *2018 13th ieee international conference on automatic face & gesture recognition (fg 2018)* (pp. 739–745).

Aran, O., & Gatica-Perez, D. (2013a). Cross-domain personality prediction: from video blogs to small group meetings. In *Proceedings of the 15th acm on international conference on multimodal interaction* (pp. 127–130).

Aran, O., & Gatica-Perez, D. (2013b). One of a kind: Inferring personality impressions in meetings. In *Proceedings of the 15th acm on international conference on multimodal interaction* (pp. 11–18).

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *Applications of computer vision (wacv), 2016 ieee winter conference on* (pp. 1–10).

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th ieee international conference on automatic face & gesture recognition (fg 2018)* (pp. 59–66).

Barnett, T., Pearson, A. W., Pearson, R., & Kellermanns, F. W. (2015). Five-factor model personality traits as predictors of perceived and actual usage of technology. *European Journal of Information Systems*, *24*(4), 374–390.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, *44*(1), 1–26.

Batrinca, L., Lepri, B., & Pianesi, F. (2011). Multimodal recognition of personality during short self-presentations. In *Proceedings of the 2011 joint acm workshop on human gesture and behavior understanding* (pp. 27–28).

Bekhouche, S. E., Dornaika, F., Ouafi, A., & Taleb-Ahmed, A. (2017). Personality traits and job candidate screening via analyzing facial videos. In *Computer vision and pattern recognition workshops (cvprw), 2017 ieee conference on* (pp. 1660–1663).

Bernstein, I. H., Garbin, C. P., & McClellan, P. G. (1983). A confirmatory factoring of the california psychological inventory. *Educational and Psychological Measurement*, *43*(3), 687–691.

Biel, J.-I., Aran, O., & Gatica-Perez, D. (2011). You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *Icwsm.*

Biel, J.-I., & Gatica-Perez, D. (2010). Voices of vlogging. In *Fourth international aaai conference on weblogs and social media.*

Biel, J.-I., & Gatica-Perez, D. (2013). The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, *15*(1), 41–55.

Biel, J.-I., Teijeiro-Mosquera, L., & Gatica-Perez, D. (2012). Facetube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings* *of the 14th acm international conference on multimodal interaction* (pp. 53–56).

Bilen, H., Fernando, B., Gavves, E., & Vedaldi, A. (2017). Action recognition with dynamic image networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., & Gould, S. (2016). Dynamic image networks for action recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3034–3042).

Bishay, M., Palasek, P., Priebe, S., & Patras, I. (2019). Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *IEEE Transactions on Affective Computing*.

Bishay, M., Priebe, S., & Patras, I. (2019). Can automatic facial expression analysis be used for treatment outcome estimation in schizophrenia? In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1632–1636).

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, *23*(3), 257–267.

Bond, M. H., Nakazato, H., & Shiraishi, D. (1975). Universality and distinctiveness in dimensions of japanese person perception. *Journal of Cross-Cultural Psychology*, *6*(3), 346–357.

Borkenau, P., Brecke, S., Möttig, C., & Paelecke, M. (2009). Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, *43*(4), 703–706.

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of personality and social psychology*, *62*(4), 645.

Botwin, M. D., & Buss, D. M. (1989). Structure of act-report data: Is the five-factor model of personality recaptured? *Journal of Personality and social Psychology*, *56*(6), 988.

Brook, J. S., Brook, D. W., De La Rosa, M., Whiteman, M., Johnson, E., & Montoya,

I. (2001). Adolescent illegal drug use: The impact of personality, family, and environmental factors. *Journal of behavioral medicine*, *24*(2), 183–203.

Brozgold, A. Z., Borod, J. C., Martin, C. C., Pick, L. H., Alpert, M., & Welkowitz, J. (1998). Social functioning and facial emotional expression in neurological and psychiatric disorders. *Applied Neuropsychology*, *5*(1), 15–23.

Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres Torres, M., Pelachaud, C., … Valstar, M. (2017). The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th acm international conference on multimodal interaction* (pp. 350–359).

Calder, A. J., Ewbank, M., & Passamonti, L. (2011). Personality influences the neural responses to viewing facial expressions of emotion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1684–1701.

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? judging emotion from the face in context. *Journal of personality and social psychology*, *70*(2), 205.

Cavallera, G., Passerini, A., & Pepe, A. (2013). Personality and gender in swimmers in indoor practice at leisure level. *Social Behaviour and Personality*, *41*, 693–704.

Celiktutan, O., Eyben, F., Sariyanidi, E., Gunes, H., & Schuller, B. (2014). Maptraits 2014-the first audio/visual mapping personality traits challenge-an introduction: Perceived personality and social dimensions. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 529–530).

Celiktutan, O., & Gunes, H. (2014). Continuous prediction of perceived traits and social dimensions in space and time. In *Image processing (icip), 2014 ieee international conference on* (pp. 4196–4200).

Celiktutan, O., & Gunes, H. (2015). Computational analysis of human-robot interactions through first-person vision: Personality and interaction experience. In *Robot and human interactive communication (ro-man), 2015 24th ieee international symposium on* (pp. 815–820).

Celiktutan, O., Sariyanidi, E., & Gunes, H. (2015). Let me tell you about your personality!†: Real-time personality prediction from nonverbal behavioural cues. In *Automatic face and gesture recognition (fg), 2015 11th ieee international conference and workshops on* (Vol. 1, pp. 1–1).

Celiktutan, O., Skordos, E., & Gunes, H. (2017). Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, *10*(4), 484–497.

Celli, F., Bruni, E., & Lepri, B. (2014). Automatic personality and interaction style recognition from facebook profile pictures. In *Proceedings of the 22nd acm international conference on multimedia* (pp. 1101–1104).

Chen, J., Chen, Z., Chi, Z., & Fu, H. (2018). Facial expression recognition in video with multiple feature fusion. *IEEE Transactions on Affective Computing*, *9*(1), 38–50.

Chentsova-Dutton, Y. E., Tsai, J. L., & Gotlib, I. H. (2010). Further evidence for the cultural norm hypothesis: Positive emotion in depressed and control european american and asian american women. *Cultural Diversity and Ethnic Minority Psychology*, *16*(2), 284.

Chioqueta, A. P., & Stiles, T. C. (2005). Personality traits and the development of depression, hopelessness, and suicide ideation. *Personality and individual differences*, *38*(6), 1283–1291.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint*.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of abnormal psychology*, *100*(3), 316.

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., ... De la Torre, F. (2009). Detecting depression from facial actions and vocal prosody.

In *Affective computing and intelligent interaction and workshops, 2009. acii 2009. 3rd international conference on* (pp. 1–7).

Conley, J. J. (1985). Longitudinal stability of personality traits: A multitrait–multimethod–multioccasion analysis. *Journal of personality and social psychology*, *49*(5), 1266.

Corr, P. J., & Matthews, G. (2009). *The cambridge handbook of personality psychology.* Cambridge University Press Cambridge, UK:.

Correa, J. A. M., Abadi, M. K., Sebe, N., & Patras, I. (2018). Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*.

Costa Jr, P. T., & McCrae, R. R. (1988). From catalog to classification: Murray's needs and the five-factor model. *Journal of personality and social psychology*, *55*(2), 258.

Craddock, N., & Mynors-Wallis, L. (2014). Psychiatric diagnosis: impersonal, imperfect and important. *The British Journal of Psychiatry*, *204*(2), 93–95.

Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., & Epps, J. (2013). Diagnosis of depression by behavioural signals: a multimodal approach. In *Proceedings of the 3rd acm international workshop on audio/visual emotion challenge* (pp. 11–20).

D., C., P., Darwin, Charles, Maudsley, & Henry. (1981). The expression of emotions in man and animals. *The American Journal of Psychology*.

DeBruine, L. M., Jones, B. C., Little, A. C., Boothroyd, L. G., Perrett, D. I., Penton-Voak, I. S., ... Tiddeman, B. P. (2006). Correlated preferences for facial masculinity and ideal or actual partner's masculinity. *Proceedings of the Royal Society of London B: Biological Sciences*, *273*(1592), 1355–1360.

de Melo, W. C., Granger, E., & Hadid, A. (2019). Combining global and local convolutional 3d networks for detecting depression from facial expressions. In *2019 14th ieee international conference on automatic face & gesture recognition (fg 2019)* (pp. 1–8).

De Moor, M., Beem, A., Stubbe, J., Boomsma, D., & De Geus, E. (2006). Regular exercise, anxiety, depression and personality: a population-based study. *Preventive* *medicine*, *42*(4), 273–279.

DENNIS, B., CHARNEY, M., NELSON, J. C., QUINLAN, D. M., et al. (1981). Personality traits and disorder in depression. *Am J Psychiatry*, *138*, 1601.

DePaulo, B. M. (1992). Nonverbal behavior and self-presentation. *Psychological bulletin*, *111*(2), 203.

Dhall, A., & Goecke, R. (2015). A temporally piece-wise fisher vector approach for depression analysis. In *2015 international conference on affective computing and intelligent interaction (acii)* (pp. 255–259).

Dhall, A., & Hoey, J. (2016). First impressions-predicting user personality from twitter profile images. In *International workshop on human behavior understanding* (pp. 148–158).

Dibeklioğlu, H., Hammal, Z., & Cohn, J. F. (2018). Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics*, *22*(2), 525–536.

Diederik P. Kingma, J. B. (2015). Adam: A method for stochastic optimization. In *Int'l conference for learning representations (iclr).*

Digman, J. M. (1989). Five robust trait dimensions: Development, stability, and utility. *Journal of personality*, *57*(2), 195–214.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, *41*(1), 417–440.

Digman, J. M., & Inouye, J. (1986). Further specification of the five robust factors of personality. *Journal of personality and social psychology*, *50*(1), 116.

Digman, J. M., & Takemoto-Chock, N. K. (1981). Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies. *Multivariate behavioral research*, *16*(2), 149–170.

Douglas-Cowie, E., Cowie, R., Cox, C., Amir, N., & Heylen, D. (2008). The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *Lrec workshop on corpora for research on emotion and affect* (pp. 1–4).

Duta, I. C., Ionescu, B., Aizawa, K., & Sebe, N. (2017). Spatio-temporal vlad encoding for human action recognition in videos. In *International conference on multimedia modeling* (pp. 365–378).

Edgar, C., McRorie, M., & Sneddon, I. (2012). Emotional intelligence, personality and the decoding of non-verbal expressions of emotion. *Personality and Individual Differences*, *52*(3), 295–300.

Edition, F., Association, A. P., et al. (1994). *Diagnostic and statistical manual of mental disorders.* Washington, American Psychological Association.

Egede, J., Valstar, M., & Martinez, B. (2017). Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. In *Automatic face & gesture recognition (fg 2017), 2017 12th ieee international conference on* (pp. 689–696).

Egede, J. O., Song, S., Olugbade, T. A., Wang, C., Williams, A., Meng, H., ... Bianchi-Berthouze, N. (2020). *Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions.*

Eisenberg, N., Miller, P. A., Schaller, M., Fabes, R. A., Fultz, J., Shell, R., & Shea, C. L. (1989). The role of sympathy and altruistic personality traits in helping: A reexamination. *Journal of Personality*, *57*(1), 41–67.

Ekman, & Paul. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384-392.

Ekman, P. (1976). Pictures of facial affect. *Consulting Psychologists Press*.

Ekman, P. (1989). The argument and evidence about universals in facial expressions. *Handbook of social psychophysiology*, 143–164.

Ekman, P. (1992a). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200.

Ekman, P. (1992b). *Facial expressions of emotion: New findings, new questions.* SAGE Publications Sage CA: Los Angeles, CA.

Ekman, P., & Friesen, W. V. (1977). Facial action coding system.

Ellgring, H. (2007). *Non-verbal communication in depression.* Cambridge University

Press.

Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Gucluturk, Y., Guclu, U., ... others (2018). Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos. *arXiv preprint arXiv:1802.00745*.

Essa, I. A., & Pentland, A. P. (2002). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *19*(7), 757-763.

Eysenck, H. J., & Eysenck, S. (1965). The eysenck personality inventory.

Eysenck, H. J., & Furnham, A. (1993). Personality and the barron-welsh art scale. *Perceptual and Motor Skills*, *76*(3), 837–838.

Fang, S., Achard, C., & Dubuisson, S. (2016). Personality classification and behaviour interpretation: An approach based on feature categories. In *Proceedings of the 18th acm international conference on multimodal interaction* (pp. 225–232).

Farkas, L. G., & Munro, I. R. (1987). *Anthropometric facial proportions in medicine.* Charles C Thomas Pub Limited.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 6202–6211).

Ferwerda, B., Schedl, M., & Tkalcic, M. (2016). Using instagram picture features to predict users' personality. In *International conference on multimedia modeling* (pp. 850–861).

Fisch, H.-U., Frey, S., & Hirsbrunner, H.-P. (1983). Analyzing nonverbal behavior in depression. *Journal of abnormal psychology*, *92*(3), 307.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology*, *44*(3), 329.

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social Psychology*, *80*(6), 1011.

Friesen, E., & Ekman, P. (1978). Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, *3*.

Gaebel, W., & Wölwer, W. (2004). Facial expressivity in the course of schizophrenia and depression. *European archives of psychiatry and clinical neuroscience*, *254*(5), 335–342.

Gehricke, J.-G., & Shapiro, D. (2000). Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion. *Psychiatry Research*, *95*(2), 157–167.

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S., & Rosenwald, D. P. (2013). Social risk and depression: Evidence from manual and automatic facial expression analysis. In *Automatic face and gesture recognition (fg), 2013 10th ieee international conference and workshops on* (pp. 1–8).

Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S. M., Hammal, Z., & Rosenwald, D. P. (2014). Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, *32*(10), 641–647.

Goldstein, I. B. (1964). Role of muscle tension in personality theory. *Psychological Bulletin*, *61*(6), 413.

Gong, Y., & Poellabauer, C. (2017). Topic modeling based multi-modal depression detection. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge* (pp. 69–76).

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, *37*(6), 504–528.

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., . . . others (2014). The distress analysis interview corpus of human and computer interviews. In *Lrec* (pp. 3123–3128).

Greenstein, F. I. (1967). The impact of personality on politics: An attempt to clear away underbrush. *American Political Science Review*, *61*(3), 629–641.

Güçlütürk, Y., Güçlü, U., van Gerven, M. A., & van Lier, R. (2016). Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *European conference on computer vision* (pp. 349–358).

Guntuku, S. C., Qiu, L., Roy, S., Lin, W., & Jakhetiya, V. (2015). Do others perceive you as you want them to?: Modeling personality based on selfies. In *Proceedings of the 1st international workshop on affect & sentiment in multimedia* (pp. 21–26).

Gupta, R., Malandrakis, N., Xiao, B., Guha, T., Van Segbroeck, M., Black, M., ... Narayanan, S. (2014). Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (pp. 33–40).

Hall, J. A., Andrzejewski, S. A., Murphy, N. A., Mast, M. S., & Feinstein, B. A. (2008). Accuracy of judging others' traits and states: Comparing mean levels across tests. *Journal of Research in Personality*, *42*(6), 1476–1489.

Hall, M. A. (1999). Correlation-based feature selection for machine learning.

Han, J., Zhang, Z., Cummins, N., Ringeval, F., & Schuller, B. (2016). Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals. *Image and Vision Computing*.

Hao, F., Pang, G., Wu, Y., Pi, Z., Xia, L., & Min, G. (2019). Providing appropriate social support to prevention of depression for highly anxious sufferers. *IEEE Transactions on Computational Social Systems*.

Haque, A., Guo, M., Miner, A. S., & Fei-Fei, L. (2018). Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.

Hasani, B., & Mahoor, M. H. (2017). Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 30–40).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

He, L., Jiang, D., & Sahli, H. (2018). Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding. *IEEE Transactions on Multimedia*.

Heinström, J. (2000). The impact of personality and approaches to learning on information behaviour. *Information research*, *5*(3), 5–3.

Hettema, J. M., Neale, M. C., Myers, J. M., Prescott, C. A., & Kendler, K. S. (2006). A population-based twin study of the relationship between neuroticism and internalizing disorders. *American journal of Psychiatry*, *163*(5), 857–864.

Hirschfeld, R. M., Klerman, G. L., Lavori, P., Keller, M. B., Griffith, P., & Coryell, W. (1989). Premorbid personality assessments of first onset of major depression. *Archives of general psychiatry*, *46*(4), 345–350.

Hogan, R. (1982). A socioanalytic theory of personality. In *Nebraska symposium on motivation.*

Hogan, R., Johnson, J. M., Johnson, J. A., & Briggs, S. R. (1997). *Handbook of personality psychology.* Elsevier.

Huang, I.-C., Lee, J. L., Ketheeswaran, P., Jones, C. M., Revicki, D. A., & Wu, A. W. (2017). Does personality affect health-related quality of life? a systematic review. *PloS one*, *12*(3), e0173806.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967-5976.

Izard, C. E. (1977). Human emotions. emotions, personality, and psychotherapy. *New York: PlenumPress*.

Izard, C. E., Dougherty, L., Hembree, E., & Izard, C. (1983). A system for identifying affect expressions by holistic judgments (affex).

Jain, V., Crowley, J. L., Dey, A. K., & Lux, A. (2014). Depression estimation using audiovisual features and fisher vector encoding. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (pp. 87–91).

Jaiswal, S., Song, S., & Valstar, M. (2019). Automatic prediction of depression and anxiety from behaviour and personality attributes. In *2019 8th international conference on affective computing and intelligent interaction (acii)* (p. 1-7). doi: 10.1109/ACII.2019.8925456

Jaiswal, S., & Valstar, M. (2016). Deep learning the dynamic appearance and shape of facial action units. In *Applications of computer vision (wacv), 2016 ieee winter conference on* (pp. 1–8).

Jaiswal, S., Valstar, M., Kusumam, K., & Greenhalgh, C. (2019). Virtual human questionnaire for analysis of depression, anxiety and personality. In *Proceedings of the 19th acm international conference on intelligent virtual agents* (pp. 81–87).

Jan, A., Meng, H., Gaus, Y. F. B. A., & Zhang, F. (2018). Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(3), 668–680.

Jenkins, J. M. (1993). Self-monitoring and turnover: The impact of personality on intent to leave. *Journal of Organizational Behavior*, *14*(1), 83–91.

John, O. P. (1990). The search for basic dimensions of personality. In *Advances in psychological assessment* (pp. 1–37). Springer.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory—versions 4a and 54.* Berkeley, CA: University of California, Berkeley, Institute of Personality . . . .

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (pp. 694–711).

Joo, J., Steen, F. F., & Zhu, S.-C. (2015). Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the ieee international conference on computer vision* (pp. 3712–3720).

Joshi, J., Goecke, R., Parker, G., & Breakspear, M. (2013). Can body expressions contribute to automatic depression analysis? In *2013 10th ieee international conference and workshops on automatic face and gesture recognition (fg)* (pp. 1–7).

Joshi, J., Gunes, H., & Goecke, R. (2014). Automatic prediction of perceived traits using visual cues under varied situational context. In *2014 22nd international conference on pattern recognition (icpr)* (pp. 2855–2860).

Junior, J., Jacques, C., Güçlütürk, Y., Pérez, M., Güçlü, U., Andujar, C., . . . others (2018). First impressions: A survey on computer vision-based apparent personality trait analysis. *arXiv preprint arXiv:1804.08046*.

Kalimeri, K., Lepri, B., & Pianesi, F. (2010). Causal-modelling of personality traits: extraversion and locus of control. In *Proceedings of the 2nd international workshop on social signal processing* (pp. 41–46).

Kaltwang, S., Todorovic, S., & Pantic, M. (2016). Doubly sparse relevance vector machine for continuous facial behavior estimation. *IEEE transactions on pattern analysis and machine intelligence*, *38*(9), 1748–1761.

Kampman, O., Barezi, E. J., Bertero, D., & Fung, P. (2018). Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (Vol. 2, pp. 606–611).

Kassin, S. M. (2003). *Essentials of psychology.* Prentice Hall.

Kaya, H., Gurpinar, F., & Ali Salah, A. (2017). Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 1–9).

Kaya, H., Gürpınar, F., & Salah, A. A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, *65*, 66–75.

Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of personality and social psychology*, *68*(3), 441.

Keltner, D. (1996). Facial expressions of emotion and personality. In *Handbook of emotion, adult development, and aging* (pp. 385–401). Elsevier.

Kendler, K. S., Gardner, C. O., Gatz, M., & Pedersen, N. L. (2007). The sources of comorbidity between major depression and generalized anxiety disorder in a swedish national twin sample. *Psychological medicine*, *37*(3), 453–462.

Kendler, K. S., Gatz, M., Gardner, C. O., & Pedersen, N. L. (2006). Personality and major depression: a swedish longitudinal, population-based twin study. *Archives of* *general psychiatry*, *63*(10), 1113–1120.

Kendler, K. S., Kuhn, J., & Prescott, C. A. (2004). The interrelationship of neuroticism, sex, and stressful life events in the prediction of episodes of major depression. *American Journal of Psychiatry*, *161*(4), 631–636.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis.* Guilford Press.

Khorrami, P., Le Paine, T., Brady, K., Dagli, C., & Huang, T. S. (2016). How deep neural networks can improve emotion recognition on video data. In *Image processing (icip), 2016 ieee international conference on* (pp. 619–623).

Kimura, S., & Yachida, M. (1997). Facial expression recognition and its degree estimation. In *Proceedings of ieee computer society conference on computer vision and pattern recognition* (pp. 295–300).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, D. N., Kotov, R., & Bufferd, S. J. (2011). Personality and depression: explanatory models and review of the evidence. *Annual review of clinical psychology*, *7*, 269–295.

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). *Nonverbal communication in human interaction.* Cengage Learning.

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, *20*(3), 165–182.

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., … Zafeiriou, S. (2019a). Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 1–23.

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., … Zafeiriou, S. (2019b, Feb 13). Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*. Retrieved from `https://doi.org/10.1007/s11263-019-01158-4` doi: 10.1007/s11263-019-01158-4

Kollias, D., & Zafeiriou, S. (2018). Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*.

Komulainen, E., Meskanen, K., Lipsanen, J., Lahti, J. M., Jylhä, P., Melartin, T., … Ekelund, J. (2014). The effect of personality on daily life emotional processes. *PLoS One*, *9*(10), e110907.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2017). Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Krueger, R. F., Caspi, A., Moffitt, T. E., Silva, P. A., & McGee, R. (1996). Personality traits are differentially linked to mental disorders: a multitrait-multidiagnosis study of an adolescent birth cohort. *Journal of abnormal psychology*, *105*(3), 299.

Lane, W., & Manner, C. (2011). The impact of personality traits on smartphone ownership and use. *International Journal of Business and Social Science*, *2*(17).

Lei, H., & Skinner, H. A. (1982). What difference does language make? structural analysis of the personality research form. *Multivariate Behavioral Research*, *17*(1), 33–46.

Lepri, B., Mana, N., Cappelletti, A., Pianesi, F., & Zancanaro, M. (2009). Modeling the personality of participants during group interactions. In *International conference on user modeling, adaptation, and personalization* (pp. 114–125).

Lepri, B., Subramanian, R., Kalimeri, K., Staiano, J., Pianesi, F., & Sebe, N. (2012). Connecting meeting behavior with extraversion—a systematic study. *IEEE Transactions on Affecive Computing*, *3*(4), 443–455.

Li, Y., Wan, J., Miao, Q., Escalera, S., Fang, H., Chen, H., … Guo, G. (2020). Cr-net: A deep classification-regression network for multimodal apparent personality analysis. *International Journal of Computer Vision*, 1–18.

Lien, J. J. J., Kanade, T., Cohn, J. F., Li, C. C., & Zlochower, A. J. (1998). Subtly different facial expression recognition and expression intensity estimation. In *Ieee computer society conference on computer vision & pattern recognition.*

Lindsten, F. (2010). A remark on zero-padding for increased frequency resolution. *Sitio web: http://goo. gl/uBMFTw*.

Liu, R., Reimer, B., Song, S., Mehler, B., & Solovey, E. (2021). Unsupervised fnirs feature extraction with cae and esn autoencoder for driver cognitive load classification. *Journal of Neural Engineering*, *18*(3), 036002.

Lo, M.-T., Hinds, D. A., Tung, J. Y., Franz, C., Fan, C.-C., Wang, Y., ... others (2017). Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nature genetics*, *49*(1), 152.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3431–3440).

Ma, X., Yang, H., Chen, Q., Huang, D., & Wang, Y. (2016). Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 35–42).

MANSTEAD, A. R. (1991). Expressiveness as individual difference. *fundamentals of Nonverbal Behavior*.

Martinez, B., Valstar, M. F., Jiang, B., & Pantic, M. (2017). Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*.

Matsudaira, T., & Kitamura, T. (2006). Personality traits as risk factors of depression and anxiety among japanese students. *Journal of clinical psychology*, *62*(1), 97–109.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, *52*(1), 81.

McCrae, R. R., & Costa, P. T. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of personality and social psychology*, *56*(4), 586.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17.

Meng, H., Huang, D., Wang, H., Yang, H., Ai-Shuraifi, M., & Wang, Y. (2013). Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd acm international workshop on audio/visual emotion challenge* (pp. 21–30).

Munro, J., & Damen, D. (2020). Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 122–132).

Nasir, M., Jati, A., Shivakumar, P. G., Nallan Chakravarthula, S., & Georgiou, P. (2016). Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 43–50).

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and social psychology bulletin*, *35*(12), 1661–1671.

Nguyen, L. S., Marcos-Ramiro, A., Marrón Romera, M., & Gatica-Perez, D. (2013). Multimodal analysis of body communication cues in employment interviews. In *Proceedings of the 15th acm on international conference on multimodal interaction* (pp. 437–444).

Nicolaou, M. A., Gunes, H., & Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, *2*(2), 92–105.

Nicolle, J., Bailly, K., & Chetouani, M. (2015). Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In *2015 11th ieee international conference and workshops on automatic face and gesture recognition (fg)* (Vol. 6, pp. 1–6).

Nicolle, J., Rapp, V., Bailly, K., Prevost, L., & Chetouani, M. (2012). Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th acm international conference on multimodal interaction* (pp. 501–508).

Noller, P., Law, H., & Comrey, A. L. (1987). Cattell, comrey, and eysenck personality factors compared: More evidence for the five robust factors? *Journal of Personality and Social Psychology*, *53*(4), 775.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, *66*(6), 574.

Okada, S., Aran, O., & Gatica-Perez, D. (2015). Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 acm on international conference on multimodal interaction* (pp. 15–22).

Ormel, J., Oldehinkel, A. J., & Vollebergh, W. (2004). Vulnerability before, during, and after a major depressive episode: a 3-wave population-based study. *Archives of general psychiatry*, *61*(10), 990–996.

Ortony, A., Norman, D. A., & Revelle, W. (2005). Effective functioning: A three level model of affect, motivation, cognition, and behavior. *Who needs emotions*, 173–202.

Palmero, C., Selva, J., Smeureanu, S., Junior, J., Jacques, C., Clapés, A., … others (2020). Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1–12).

Pampouchidou, A., Simantiraki, O., Fazlollahi, A., Pediaditis, M., Manousos, D., Roniotis, A., … others (2016). Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 27–34).

Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition. In *Bmvc* (Vol. 1, p. 6).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., … Lerer, A. (2017). Automatic differentiation in pytorch.

Pentland, A. (2006). A computational model of social signalin..

People with mental health problems still waiting over a year for talking treatments. (2013). https://www.mind.org.uk/news-campaigns/news/people-with-mental-health-problems-still-waiting-over-a-year-for-talking-treatments/. ([Online; accessed 25-July-2019])

Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., & Zancanaro, M. (2008). Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on multimodal interfaces* (pp. 53–60).

Pintea, S. L., van Gemert, J. C., & Smeulders, A. W. M. (2014). Deja vu: Motion prediction in static images. In *Eccv.*

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., ... Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European conference on computer vision* (pp. 400–418).

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the european conference on computer vision (eccv)* (pp. 818–833).

Qin, R., Gao, W., Xu, H., & Hu, Z. (2018). Modern physiognomy: an investigation on predicting personality traits and intelligence from the human face. *Science China Information Sciences*, *61*(5), 058105.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, *41*(1), 203–212.

Rana, R., Latif, S., Gururajan, R., Gray, A., Mackenzie, G., Humphris, G., & Dunn, J. (2019). Automated screening for distress: A perspective for the future. *European journal of cancer care*, e13033.

Reed, L. I., Sayette, M. A., & Cohn, J. F. (2007). Impact of depression on response to comedy: A dynamic facial coding analysis. *Journal of abnormal psychology*, *116*(4), 804.

Renneberg, B., Heyn, K., Gebhard, R., & Bachmann, S. (2005). Facial expression of emotions in borderline personality disorder and depression. *Journal of behavior therapy and experimental psychiatry*, *36*(3), 183–196.

Revelle, W., & Scherer, K. R. (2009). Personality and emotion. *Oxford companion to* *emotion and the affective sciences*, 304–306.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., . . . Pantic, M. (2019). Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop* (p. 3–12). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3347320.3357688` doi: 10.1145/3347320.3357688

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., . . . Pantic, M. (2017). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge* (pp. 3–9).

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic face and gesture recognition (fg), 2013 10th ieee international conference and workshops on* (pp. 1–8).

Roberts, B. W., & Jackson, J. J. (2008). Sociogenomic personality psychology. *Journal of personality*, *76*(6), 1523–1544.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Rottenberg, J., Kasch, K. L., Gross, J. J., & Gotlib, I. H. (2002). Sadness and amusement reactivity differentially predict concurrent and prospective functioning in major depressive disorder. *Emotion*, *2*(2), 135.

Russell, J. A. (1991a). Culture and the categorization of emotions. *Psychological bulletin*, *110*(3), 426.

Russell, J. A. (1991b). Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin*, *115*(1), 102.

Sanchez-Cortes, D., Aran, O., Mast, M. S., & Gatica-Perez, D. (2011). A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions* *on Multimedia*, *14*(3), 816–832.

Scherer, K., & Ekman, P. (1982). *Handbook of methods in nonverbal behavior research.* Cambridge University Press.

Scherer, S., Stratou, G., & Morency, L.-P. (2013). Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th acm on international conference on multimodal interaction* (pp. 135–140).

Schuller, B., Valster, M., Eyben, F., Cowie, R., & Pantic, M. (2012). Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th acm international conference on multimodal interaction* (pp. 449–456).

Senoussaoui, M., Sarria-Paja, M., Santos, J. F., & Falk, T. H. (2014). Model fusion for multimodal depression classification and level detection. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (pp. 57–63).

Shevlin, M., Walker, S., Davies, M. N., Banyard, P., & Lewis, C. A. (2003). Can you judge a book by its cover? evidence of self–stranger agreement on personality at zero acquaintance. *Personality and Individual Differences*, *35*(6), 1373–1383.

Sloan, D. M., Strauss, M. E., Quirk, S. W., & Sajatovic, M. (1997). Subjective and expressive emotional responses in depression. *Journal of affective disorders*, *46*(2), 135–141.

Sloan, D. M., Strauss, M. E., & Wisner, K. L. (2001). Diminished response to pleasant stimuli by depressed women. *Journal of abnormal psychology*, *110*(3), 488.

Smith, G. M. (1967). Usefulness of peer ratings of personality in educational research. *Educational and Psychological measurement*, *27*(4), 967–984.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*(3), 199–222.

Song, S., Jaiswal, S., Sanchez, E., Tzimiropoulos, G., Shen, L., & Valstar, M. (2021). Self-supervised learning of person-specific facial dynamics for automatic personality recognition. *IEEE Transactions on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2021.3064601

Song, S., Jaiswal, S., Shen, L., & Valstar, M. (2020). Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2020.2970712

Song, S., Sanchez, E., Shen, L., & Valstar, M. (2021). Self-supervised learning of dynamic representations for static images. In *2020 25th international conference on pattern recognition (icpr)* (pp. 1619–1626).

Song, S., Sánchez-Lozano, E., Kumar Tellamekala, M., Shen, L., Johnston, A., & Valstar, M. (2019). Dynamic facial models for video-based dimensional affect estimation. In *Proceedings of the ieee international conference on computer vision workshops* (pp. 0–0).

Song, S., Sánchez-Lozano, E., Shen, L., Johnston, A., & Valstar, M. (2019). Inferring dynamic representations of facial actions from a still image. *arXiv preprint arXiv:1904.02382*.

Song, S., Shen, L., & Valstar, M. (2018). Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *2018 13th ieee international conference on automatic face gesture recognition (fg 2018)* (p. 158-165). doi: 10.1109/FG.2018.00032

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Staiano, J., Lepri, B., Subramanian, R., Sebe, N., & Pianesi, F. (2011). Automatic modeling of personality states in small group interactions. In *Proceedings of the 19th acm international conference on multimedia* (pp. 989–992).

Stepanov, E. A., Lathuiliere, S., Chowdhury, S. A., Ghosh, A., Vieriu, R.-L., Sebe, N., & Riccardi, G. (2018). Depression severity estimation from multiple modalities. In *2018 ieee 20th international conference on e-health networking, applications and services (healthcom)* (pp. 1–6).

Stuhrmann, A., Suslow, T., & Dannlowski, U. (2011). Facial emotion processing in major depression: a systematic review of neuroimaging findings. *Biology of mood & anxiety disorders*, *1*(1), 10.

Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., & Mittal, A. (2016). Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *European conference on computer vision* (pp. 337–348).

Subramanian, R., Wache, J., Abadi, M. K., Vieriu, R. L., Winkler, S., & Sebe, N. (2016). Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, *9*(2), 147–160.

Subramanian, R., Yan, Y., Staiano, J., Lanz, O., & Sebe, N. (2013). On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *Proceedings of the 15th acm on international conference on multimodal interaction* (pp. 3–10).

Sun, B., Zhang, Y., He, J., Yu, L., Xu, Q., Li, D., & Wang, Z. (2017). A random forest regression method with selected-text feature for depression assessment. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge* (pp. 61–68).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Takahashi, Y., Roberts, B. W., Yamagata, S., & Kijima, N. (2015). Personality traits show differential relations with anxiety and depression in a nonclinical sample. *Psychologia*, *58*(1), 15–26.

Teijeiro-Mosquera, L., Biel, J.-I., Alba-Castro, J. L., & Gatica-Perez, D. (2015). What your face vlogs about: expressions of emotion and big-five traits impressions in youtube. *IEEE Transactions on Affective Computing*, *6*(2), 193–205.

Tian, Y.-L., Kanade, T., Cohn, J. F., Li, S. Z., & Jain, A. K. (2005). *Facial expression analysis.* Springer London.

Tsai, J. L., Pole, N., Levenson, R. W., & Muñoz, R. F. (2003). The effects of depression on the emotional responses of spanish-speaking latinas. *Cultural Diversity and Ethnic Minority Psychology*, *9*(1), 49.

Tsao, W.-C., & Chang, H.-R. (2010). Exploring the impact of personality traits on online shopping behavior. *African Journal of Business Management*, *4*(9), 1800–1812.

Tu, Z., Li, H., Zhang, D., Dauwels, J., Li, B., & Yuan, J. (2019). Action-stage emphasized spatiotemporal vlad for video action recognition. *IEEE Transactions on Image Processing*, *28*(6), 2799–2812.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., ... Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 3–10).

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., ... Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (pp. 3–10).

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., ... Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd acm international workshop on audio/visual emotion challenge* (pp. 3–10).

Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., ... Cohn, J. F. (2015). Fera 2015-second facial expression recognition and analysis challenge. In *Automatic face and gesture recognition (fg), 2015 11th ieee international conference and workshops on* (Vol. 6, pp. 1–8).

Ventura, C., Masip, D., & Lapedriza, A. (2017). Interpreting cnn models for apparent personality trait regression. In *Computer vision and pattern recognition workshops (cvprw), 2017 ieee conference on* (pp. 1705–1713).

Vernon, R. J., Sutherland, C. A., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353–E3361.

Vinciarelli, A., & Mohammadi, G. (2014). A survey of personality computing. *IEEE Transactions on Affective Computing*, *5*(3), 273–291.

Vukasović, T., & Bratko, D. (2015). Heritability of personality: a meta-analysis of behavior genetic studies. *Psychological bulletin*, *141*(4), 769.

Walker, J., Doersch, C., Gupta, A., & Hebert, M. (2016). An uncertain future: Forecasting from static images using variational autoencoders. In *European conference on computer vision* (pp. 835–851).

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20–36).

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Ieee conference on computer vision and pattern recognition (cvpr)* (Vol. 1, p. 5).

Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In *Neural networks (ijcnn), 2017 international joint conference on* (pp. 1578–1585).

Watson, D., & Clark, L. A. (1992). On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. *Journal of personality*, *60*(2), 441–476.

Wei, X.-S., Zhang, C.-L., Zhang, H., & Wu, J. (2018). Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, *9*(3), 303–315.

Wen, L., Li, X., Guo, G., & Zhu, Y. (2015). Automated depression diagnosis based on facial dynamic analysis and sparse coding. *IEEE Transactions on Information Forensics and Security*, *10*(7), 1432–1441.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., & Mehta, D. D. (2014). Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th international workshop on audio/visual emotion challenge* (pp. 65–72).

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., & Mehta, D. D. (2013). Vocal biomarkers of depression based on motor incoordination. In *Proceed-*

ings of the 3rd acm international workshop on audio/visual emotion challenge (pp. 41–48).

Yan, Y., Nie, J., Huang, L., Li, Z., Cao, Q., & Wei, Z. (2016). Exploring relationship between face and trustworthy impression using mid-level facial features. In *International conference on multimedia modeling* (pp. 540–549).

Yang, C., Xu, Y., Shi, J., Dai, B., & Zhou, B. (2020). Temporal pyramid network for action recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 591–600).

Yang, L., Jiang, D., He, L., Pei, E., Oveneke, M. C., & Sahli, H. (2016). Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 89–96).

Yang, L., Jiang, D., & Sahli, H. (2018). Integrating deep and shallow models for multimodal depression analysis—hybrid architectures. *IEEE Transactions on Affective Computing*.

Yi, Z., Zhang, H. R., Tan, P., & Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. In *Iccv* (pp. 2868–2876).

Zebrowitz, L. A., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin*, *28*(2), 238–249.

Zhang, K., Huang, Y., Du, Y., & Wang, L. (2017). Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, *26*(9), 4193–4203.

Zhang, L., Peng, S., & Winkler, S. (2019). Persemon: A deep network for joint analysis of apparent personality, emotion and their relationship. *IEEE Transactions on Affective Computing*.

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *European conference on computer vision* (pp. 649–666).

Zhou, X., Jin, K., Shang, Y., & Guo, G. (2018). Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*.

Zhu, Y., Shang, Y., Shao, Z., & Guo, G. (2017). Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing*.

Zimmerman, R. D. (2008). Understanding the impact of personality traits on individuals'turnover decisions: A meta-analytic path model. *Personnel Psychology*, *61*(2), 309–348.

Zlochower, A. J., Cohn, J. F., Lien, J. J.-J., & Kanade, T. (1998). Automated face coding: A computer-vision based method of facial expression analysis in parent-infant interaction. *Infant Behavior and Development*(21), 16.