Personality Recognition by Modelling Person-specific Cognitive Processes using Graph Representation

Zilong Shao[1,2,3], Siyang Song[4]*, Shashank Jaiswal[5], Linlin Shen[1,2,3]*, Michel Valstar[5], Hatice Gunes[4]
[1]Computer Vision Institute, School of Computer Science and Software Engineering, Shenzhen University, China
[2]Shenzhen Institute of Artificial Intelligence of Robotics of Society, Shenzhen, China
[3]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China
[4]Department of Computer Science and Technology, University of Cambridge, United Kingdom
[5]Computer Vision Lab, University of Nottingham, United Kingdom
shaozilong2019@email.szu.edu.cn,ss2796@cam.ac.uk,shashank@blueskeye.com,llshen@szu.edu.cn
michel.valstar@nottingham.ac.uk,Hatice.Gunes@cl.cam.ac.uk

ABSTRACT

Recent research shows that in dyadic and group interactions individuals' nonverbal behaviours are influenced by the behaviours of their conversational partner(s). Therefore, in this work we hypothesise that during a dyadic interaction, the target subject's facial reactions are driven by two main factors: (i) their internal (person-specific) cognition, and (ii) the externalised nonverbal behaviours of their conversational partner. Subsequently, our novel proposition is to simulate and represent the target subject's (i.e., the listener) cognitive process in the form of a person-specific CNN architecture whose input is the audio-visual non-verbal cues displayed by the conversational partner (i.e., the speaker), and the output is the target subject's (i.e., the listener) facial reactions. We then undertake a search for the optimal CNN architecture whose results are used to create a person-specific graph representation for recognising the target subject's personality. The graph representation, fortified with a novel end-to-end edge feature learning strategy, helps with retaining both the unique parameters of the person-specific CNN and the geometrical relationship between its layers. Consequently, the proposed approach is the first work that aims to recognize the true (self-reported) personality of a target subject (i.e., the listener) from the learned simulation of their cognitive process (i.e., parameters of the person-specific CNN). The experimental results show that the CNN architectures are well associated with target subjects' personality traits and the proposed approach clearly outperforms multiple existing approaches that predict personality directly from non-verbal behaviours. In light of these findings, this work opens up a new avenue of research for predicting and recognizing socio-emotional phenomena (personality, affect, engagement etc.) from simulations of person-specific cognitive processes.

*Corresponding author.

This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '21, October 20–24, 2021, Virtual Event, China.
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8651-7/21/10.
https://doi.org/10.1145/3474085.3475460

CCS CONCEPTS

• **Computing methodologies** → *Activity recognition and understanding*; • **Applied computing** → *Psychology*.

KEYWORDS

personality recognition, neural architecture search, graph neural network, edge feature learning

ACM Reference Format:
Zilong Shao, Siyang Song, Shashank Jaiswal, Linlin Shen, Michel Valstar, Hatice Gunes. 2021. Personality Recognition by Modelling Person-specific Cognitive Processes using Graph Representation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475460

1 INTRODUCTION

Personality computing [50] has attracted increasing research interest over the last decade due to its wide range of applications such as candidate screening for recruitment [38], personalised and adaptive human-agent and human-robot interactions (e.g., [19]). Recent advances in machine learning (ML) have enabled the development of non-invasive automatic personality traits analysers that can provide fast, objective, and repeatable personality assessments. As there is converging evidence demonstrating that nonverbal behaviours are significant predictors of personality, most existing automatic approaches attempt to predict personality traits from nonverbal audio-visual behaviours (e.g., facial expressions, vocal prosody, etc.) [10, 17, 29, 49, 53]. Majority of these works focus on analysing an individual's observable behaviours, disregarding the interpersonal interaction context and cues (e.g., interpersonal behaviours due to dyadic / triadic /group interactions). However, recent research shows that in dyadic and group interactions individuals' nonverbal behaviours and affect are influenced by, and therefore can be predicted from, the behaviours of their conversational partner(s) [31], [33]. Therefore, a recent trend in personality computing is to utilise and incorporate interpersonal features for automatic personality analysis of a target subject (e.g., [35], [11], [41]).

Metaphorically speaking, in these approaches, the ML model can be said to play the role of the external artificial observer that observes the target subject's nonverbal distal cues - i.e., audio signals (e.g., delta-mel-cepstral, speech duration, etc.) [18, 40, 51], visual

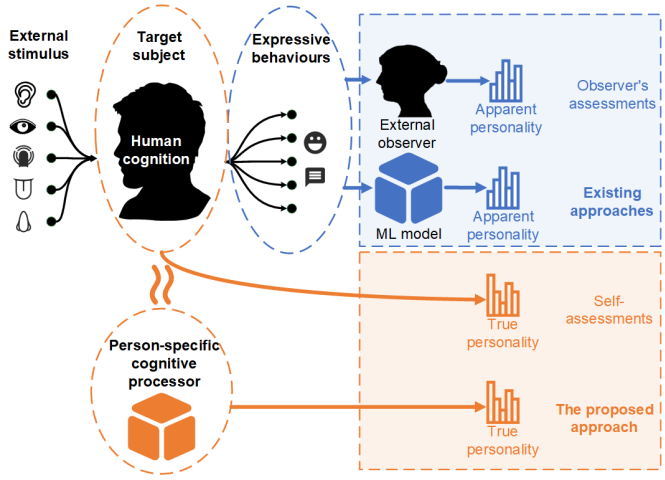


Figure 1: The difference between the proposed approach (depicted in orange) and existing approaches (depicted in blue). While the existing approaches attempt to achieve automatic personality perception from non-verbal behaviours, our approach learns to recognize true personality by modelling target subjects' cognition.

cues (e.g., facial actions and gestures) [18, 40, 51], and/or observable inter-personal cues [17, 41], [35], [11]. Individuals externalize their personality through distal cues (e.g., energy) but these cues undergo a perception bias based on what the observer actually perceives, and become proximal cues (e.g., loudness) [50]. The ML model trained with the personality labels provided by the external observers, outputs its *perception* of the corresponding human speaker's personality. In other words, the predictions yielded by these ML approaches are similar to the human listener's perceptions of the speaker's personality. These are referred to as Automatic Personality Perception (APP) solutions (the inference of assessments from proximal cues) [50]. However, in some contexts, the task is to infer the self-assessed (true) personality, i.e., Automatic Personality Recognition (APR), which targets the inference of true personalities from machine detectable distal cues [50]. While APP aims to predict apparent personality based on proximal behavioural cues, the goal of APR is to predict the true personality that impacts the generation of such behaviours (distal cues). Thus, utilising models that were trained as external observers or listeners may not be appropriate for inferring true personality traits **(Problem 1)**.

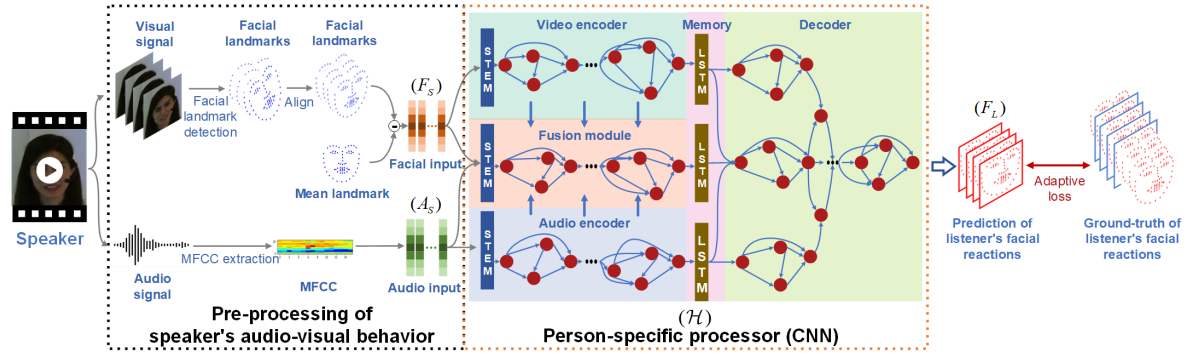
Moreover, majority of the automatic personality analysers attempted to infer personality traits from a single frame or thin slices of behaviour (short segments) [17, 24, 51]. These approaches tend to re-use clip-level personality labels as the labels for its constituent frames or short segments, and train ML models based on such frames/segments. The problem with these approaches is that at the level of a single frame or short segment, even people with different personality traits may display very similar non-verbal audio-visual behaviours. Therefore, these training strategies would end up utilising the same input pattern with multiple labels, making it practically impossible to train a model that has a good generalization capability [44, 45, 47] **(Problem 2)**. Although some approaches select a set of key frames to represent an entire video and infer personality from such video-level representations [4, 29, 53] , they ignore the details contained in the discarded frames **(Problem 3)**.

To address the shortcomings highlighted above, this paper aims to: (i) infer true personality based on the assumption that personality influences the individuals externalization of their distal cues [50] **(P1)**, and (ii) encode the information of an entire clip into a length-independent representation without discarding any frames **(P2)**. To achieve the first aim, we draw inspiration from the biological studies showing that personality is closely associated with human cognition [14, 22, 25, 27] that senses and processes incoming external information, and produces relevant behavioural responses. As CNNs are powerful tools to simulate various human cognitive processes (e.g., pattern recognition, classification, etc.), in this paper we investigate how a CNN can simulate human cognition and whether the simulated cognitive processes are informative for recognizing the target individual's personality. Accordingly, we hypothesise that during a dyadic interaction, the target subjects' (listeners) facial reactions are driven by two main factors: (i) their internal (person-specific) cognition, and (ii) the externalised nonverbal behaviours of their conversational partner (speakers). Subsequently, our novel proposition is to learn an individual's person-specific cognitive process from the non-verbal behaviours they display in response to their conversational partner, and then use the learned representation to infer their true personality. To the best of our knowledge, this is the first audio-visual approach that attempts to recognize the true (self-reported) personality of target subjects by modelling their cognitive processes (**Contribution 1**). In addition to the above, we propose three additional contributions. As CNNs capability for cognitive modelling are determined by their specific architecture and weights, we first introduce a multi-modal neural architecture search strategy to automatically search an optimised person-specific processor architecture that can reproduce the subject's facial reactions (**Contribution 2**). We then propose a novel graph representation encoding strategy to parameterize each CNN architecture into a fixed-size graph representation, where a weighting method is utilized to encode CNN edge parameters as vertex features for graph representation (**Contribution 3**). While most existing graph network-based approaches only consider the binary adjacency relationship between vertices as the edge features, we propose a novel end-to-end deep learning strategy to learn task-specific edge features that provide richer clues for personality recognition (**Contribution 4**). The encoded graph representation is then fed to a Graph Neural Network (GNN) for automatic personality traits recognition (addressing P1). In this paper, each person-specific processor is optimized by all available data of the target subject. Thus, the proposed approach is able to utilize the information from an entire clip (addressing P2 and P3). The full pipeline of our approach is illustrated in Fig. 2.

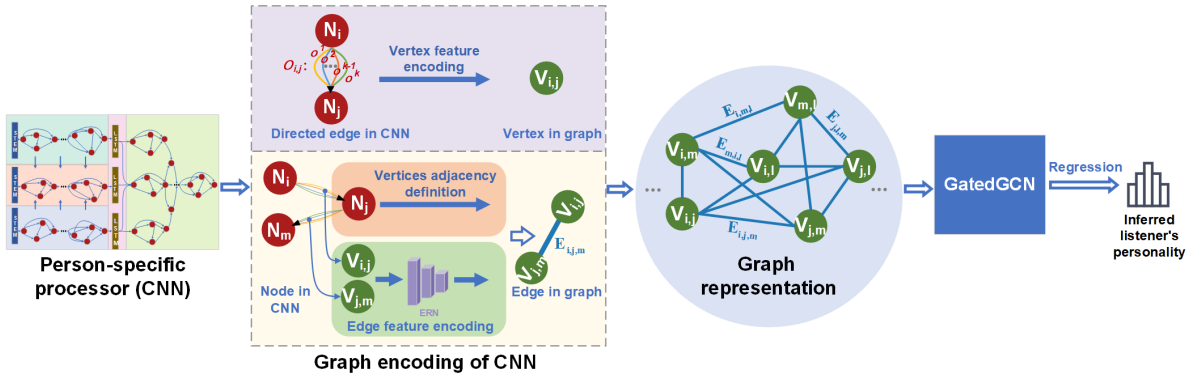
2 RELATED WORK

2.1 Automatic audio-visual personality analysis

Early automatic personality analysis studies usually starts with extracting low-level hand-crafted audio-visual non-verbal features [10, 34, 36, 48] or interpersonal descriptors [17, 41], and then feed



(a) Searching a person-specific processor to simulate a target subject's (listener's) cognition.



(b) Recognition of the target subject's (listener) personality from the graph representation of the target subject's (listener) person-specific CNN model.

Figure 2: The pipeline of the proposed approach. (a) Our approach starts with searching a person-specific processor (multi-modal CNN) architecture that can reproduce listener's facial reactions according to the speaker's audio-visual non-verbal signals (Sec. 3.1); (b) Then, we parameterize the person-specific processor as a graph representation to represent the listener's cognition and feed it to a graph neural network for listener's personality recognition (Sec. 3.2).

them to traditional regressors to predict personality traits. Recently, deep learning techniques have started to dominate this research area. In [18] and [51], two-stream audio-visual bi-modal networks are utilized to combine frame-level deep visual and audio features at the fully connected layer, and infer personality at a frame-level. Ventura et al. [49] proposed a Descriptor Aggregation Network (DAN) to extract frame-wise features at multiple spatial resolutions. The results show that facial information, especially eyes, nose and mouth, play a key role for personality trait prediction. Instead of inferring personality at frame-level, Principi et al.[40] proposed a multi-modal CNN to jointly learn audio and visual information from thin video segments, which are combined with predictions of attribute-specific models to predict personality.

Since personality trait models focus on evaluating the aspects of personality that are relatively stable over a long period of time (usually longer than the duration of a single audio-visual clip), the features extracted from a frame or thin slices of behaviours may not carry reliable information to infer personality [23]. Thus, some studies [3, 34] summarize frame/segment-level features of the entire clip into a global statistical descriptor. Meanwhile, other studies attempt to select a small number of key frames to represent the entire video. For example, Zhang et al. [53] selects a face image from each short segment. Then, a consensus strategy is employed to process all the selected frames, to produce video-level personality predictions. Li et al. [29] also select a face image and a face-background image from each segment and stacks them as the clip-level stream. This approach converts the acoustic wave of an entire clip to fixed-length vectors as the clip-level audio representation. Beyan et al. [4] proposed to generate multiple dynamic facial images [5, 46] to represent each video segment and then choose a set of dynamic images that have the highest spatiotemporal saliency as the key-dynamic images to construct the video-level representation. In summary, the aforementioned approaches do not utilise the full scale of the available information in the data, as they select a subset or key frames to represent an entire video to infer personality.

2.2 The relationship between personality and human cognition

There are exist biological and psychological studies claiming that people's cognition relates to their personality. Kumari et al. [27] examined the influence of extraversion and neuroticism with brain fMRI activity based on "n-back" task and found that brain responses

during cognitive activities are related to personality. More specifically, the personality has been found to be important in many cognitive processes such as risk taking [26], creativity [32], music learning [13]. Particularly, in [13], an exploratory factor analysis was conducted, where the experimental results indicate that creativity and primary process cognition are associated with the extraversion and psychoticism traits. In addition, a longitudinal study conducted by Schaie et al. [43] showed that some of the personality-cognition relations could last for over 35-years. As humans' person-specific cognition largely depends on their unique brain structure and activities [16, 52], a number of studies also show that some personality traits (e.g., Extraversion, Conscientiousness, and Neuroticism) can be reflected in the brain structure differences [25] and activities, such as brain local volumes [14] and gray and white matter [21].

As reviewed in Sec. 2.1, the difference between our approach, depicted in orange in Fig. 1, and the existing approaches (depicted in blue), is the fact that the existing approaches attempt to achieve automatic personality perception from non-verbal behaviours. Our approach instead, draws inspiration from the aforementioned works on the interrelationship between personality and human cognition, and learns to recognize true personality by simulating and modelling target subjects' cognitive processes.

3 METHODOLOGY

As discussed in [50], automatic personality recognition aims to predict true personality that is associated with human expressive distal cues (e.g., facial actions) governed by human cognition. In this sense, we propose to conduct automatic personality recognition by modelling person-specific cognitive processes using graph representation. To simulate and model a subject's cognitive processes, we search for an optimal person-specific multi-modal CNN architecture that can accurately reproduce this subject's facial reactions (Fig. 2(a), also visualized in our supplementary video)). We hypothesize that this CNN architecture represents the person-specific cognitive processes of the target subject. To use the person-specific CNN architecture as input for personality recognition, we represent the CNN parameters (operation parameters (OP) and layers' weights (LW)) in a graph format, which are then processed by the graph neural network (GNN) model for automatic personality recognition (Fig. 2(b)).

3.1 Simulating person-specific cognition

Input and Target: We first define the criteria for evaluating whether a CNN is able to simulate an individual's cognition accurately. Inspired by the human cognitive processor model [9] (shown in Fig. 3), which receives and processes external signals and generates corresponding behavioural reactions, we propose that when a CNN receives the same external signals, if it is able to generate the same reactions as the target subject, its architecture can represent the subject's cognitive processes. We formulate our target as follows. During a dyadic interaction, given an audio signal A_S and facial behaviours F_S of the speaker, the goal is to learn a CNN model $\mathcal{H}$ that can generate the facial reactions F_L of the listener, which can be denoted as:

$$F_L = \mathcal{H}(A_S, F_S) \tag{1}$$

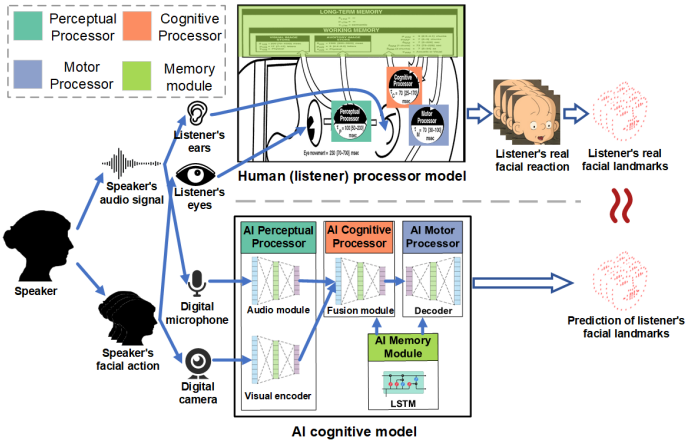


Figure 3: Using CNN to simulate human cognition.

Once $\mathcal{H}$ is learned well, it takes on the role of the corresponding listener's cognitive processor during a dyadic interaction. Consequently, we assume that $\mathcal{H}$ is *sufficiently* informative for modeling the listener's true personality traits not only because human personality relates to human cognition but also because true personality is a key factor in governing how human non-verbal behaviours are generated and displayed [50].

This paper uses speaker and listener's facial landmark sequences (in the form of 80 frames) as the input and target facial action representation, respectively. The aligned facial landmarks are obtained for each frame using OpenFace 2.0 [2], which are then transformed based on a pre-defined mean face shape in order to keep only facial behaviours without the identity information (as suggested by [15]). We use 64 bin log-mel spectra as the audio representation, where each audio frame is computed by a 40 ms hanning window with stride size of 40 ms. This way, the number of audio frames for each video is the same as the number of video frames. This is illustrated as the pre-processing part in Fig. 2(a).

Multi-modal cognitive processor model: The main idea behind our approach is to predict true personality traits from the simulated human cognition. To achieve this goal, we search for an optimal person-specific cognitive processor model architecture, represented as a multi-modal CNN, to represent the target subject's unique cognitive processes $\mathcal{H}$. We follow the previous neural architecture search (NAS) approaches [30, 37] to represent each CNN architecture as a directed acyclic graph where a node N_j represents a set of feature maps generated from its adjacent parent nodes $N_i, N_{i+1}, \cdots N_{j-1}$ $(i < j)$. A pair of adjacent nodes (N_i and N_j) are connected by a directed edge $DE_{i,j}$ which consists of a set of pre-defined operations $o^k_{i,j}$ (10 operations were used in this paper, explained in Sec. 3.2). As a result, the feature maps of N_j are jointly produced by all operations on all of its parent nodes, which can be formulated as

$$N_j = \sum_{i<j} (O_{i,j}(N_i) \times \text{adj}(i,j)) \tag{2}$$

where $\text{adj}(i,j))$ denotes the binary adjacency relationship (e.g., 0 or 1) between N_i and N_j. Here, each operation $o^k_{i,j}$ in $DE_{i,j}$ is learned

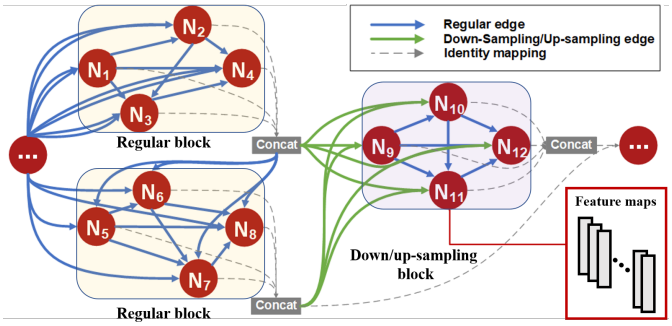


Figure 4: The details of the CNN internal connections. Each node in a regular block is influenced by the information coming from all its previous nodes, and each regular block takes the outputs from previous two regular blocks.

to have a unique parameter $\alpha_{i,j}^k$ to represent its importance.:

$$O_{i,j}(N_i) = \sum_{k=1}^{K} (\alpha_{i,j}^k \times o_{i,j}^k(N_i)) \tag{3}$$

Inspired by the Model Human Processor (MHP) [9], the main architecture of the person-specific processes is set to have a visual encoder and an audio encoder that simulate the human perceptual processor, a fusion module that simulates the human cognitive processor, and a decoder to simulate the human motor processor. Also, a Long-short-term-memory network (LSTM) is employed to process the latent features generated from two encoders and the fusion module, simulating the human working memory (illustrated in Fig. 3). Specifically, we set the audio encoder, visual encoder and fusion module to have the same number of blocks and each block contains multiple nodes and edges. The n_{th} $(n > 2)$ fusion block takes four inputs: the outputs of the n_{th} visual block and n_{th} audio block, the output of the $(n-1)_{th}$ fusion block, and the output of the $(n-2)_{th}$ fusion block. Consequently, the input audio and visual signals can be combined and jointly processed at multiple levels (illustrated as the person-specific processor in Fig. 2(a)) during the CNN processing, simulating the uncertainty and complexity of human cognitive and reaction processes. Specifically, we define blocks whose input and output feature maps have the same size as the regular blocks, and in such blocks, each node is connected to all of its previous nodes, allowing the extracted features (nodes) to be potentially influenced by the information of multiple previous features (nodes) during the CNN model learning (depicted in yellow in Fig. 4). In this paper, we set each CNN edge to have unique OPs and LWs.

Adaptive loss function: To search an optimal multi-modal architecture that can reproduce the facial reactions of a listener (i.e., achieving Eqa. 1), we consider that there is always a time delay for the listener to generate facial reactions as execution of the corresponding cognitive processes take some time (as stated in [9]). The time delay may vary not only for different individuals but also for the same individual depending upon external factors. In light of this, we introduce a novel adaptive factor τ to model such uncertainty. Let us define an audio-visual input $A_S(t_1, t_2)$ and $F_S(t_1, t_2)$ as the speaker's audio-visual non-verbal behaviours that were produced from time t_1 to t_2. We propose an adaptive MSE (A-MSE) loss function to measure the similarity between the predicted listener's facial reaction and the ground-truth:

$$\begin{aligned} & L_{A-MSE}(t_2, \tau) \\ & = L_{MSE}(F_L^p(t_1, t_2), F_L^g(t_1 + \tau, t_2 + \tau)) \\ & = \sum_{i=t_1}^{t_2} \sum_{j=1}^{68} \min((x_{i,j}^p - x_{i+\tau,j}^g)^2 + (y_{i,j}^p - y_{i+\tau,j}^g)^2, \varepsilon) \end{aligned} \tag{4}$$

where $F_L^p(t_1, t_2)$ and $F_L^g(t_1 + \tau, t_2 + \tau)$ are the predicted and the true (ground-truth) facial reaction landmarks of the listener; $(x_{i,j}^p, y_{i,j}^p)$ denotes the predicted coordinates of the jth facial landmark of the ith frame and $(x_{i+\tau,j}^g, y_{i+\tau,j}^g)$ is the corresponding ground-truth coordinate. Here, the ε is a constant value employed to avoid extremely large loss values caused by outliers (e.g. incorrectly detected face regions) which can lead to a misguided CNN search. To achieve the adaptive loss, in practice we use a sliding time-window to compare the prediction of listener's facial reactions and the set of ground truth candidates as illustrated in the last section of Fig. 2(a), e.g., $F_L(t_1 + r, t_2 + r), r = 1, 2, \cdots R$, and only choose $r = \tau$ that allows the loss $L_{A-MSE}(t_1, t_2, \tau)$ to have the lowest value, i.e., $\tau = \text{argmin} L(t_1, t_2, \tau)$. As a result, the delay period can be automatically adapted for each iteration.

Person-specific cognitive processor architecture optimization: In this paper, we conduct a single-level optimization strategy using the continuous relaxation algorithm [30] for person-specific processor architecture search. Particularly, our strategy adjusts both operation parameters (OPs) α and layers' weights (LWs) ω of the entire CNN simultaneously to minimize the A-MSE loss. In comparison to the popular bi-level optimization strategy [12, 30] which separately optimizes OPs in the validation set and LWs in the training set, the proposed single-level strategy firstly allows the OPs and LWs to be simultaneously optimized rather than freezing one of them while optimizing the other. This aims to replicate how human cognition works with all cognitive processes jointly activated during reaction generation - there is no evidence suggesting that some parts of the human model processor are frozen during reaction generation. Secondly, the proposed single-level strategy allows the parameters and weights to be optimized using the full clip instead of a sub-segment of it (**addressing the Problem 2**).

3.2 Graph representation for personality recognition

To allow for the person-specific processors to be modelled by standard ML regression techniques, we encode each processor's CNN architecture (α, ω) (i.e., OPs and LWs of the multi-modal CNN) into a graph representation $G(V, E)$ that consists of several vertices and edges, as CNNs have graphical topology.

Vertex features: A vertex $V_{i,j}$ of a graph representation $G(V, E)$ is a 1-D vector that represents a directed edge $DE_{i,j}$ in the searched CNN, which contains a set of operation parameters (OPs) $\alpha_{i,j}$ and layers' weights (LWs) $\omega_{i,j}$. This process is illustrated in Fig. 2(b) and is depicted in purple. We notice that the number of valid OPs are depending on the type of the directed edge, e.g. the up/down-sampling OPs are not used in directed edges of regular blocks while

convolution OPs are not used for up/down-sampling directed edges. To keep all directed edges have the same number of operations (10 in this paper), we set OPs of those invalid operations to zero. In addition, since the number of LWs for directed edges depend on the number of their input and output feature maps, these are different for different directed edges of a CNN (ranging from hundreds to tens of thousands). To allow each vertex feature have the same dimensionality, we follow the idea of [28] to select a subset of weights $S\omega_{i,j}^{k}$ from each operation o^k to represent its most important kernels (filters), i.e., we choose kernels with top-5 highest L1 values (sum of absolute weights). As a result, the number of selected weights are also fixed for all directed edges regardless of the number of input and output feature maps. Since each OP $\alpha_{i,j}^{k}$ represents the importance of k_{th} operations of the $DE_{i,j}$, we multiply each OP $\alpha_{i,j}^{kw}$ with the corresponding $S\omega_{i,j}^{kw}$, where $kw = kw_1, kw_2, \cdots, kw_\upsilon$ corresponds to the index of operations that have LWs (e.g., convolution). Finally, we concatenate the obtained multiplied OP-LW vector (the second term in Eqa. 5) with other κ OPs of operations (pooling, up/down-sampling, identity mapping) that do not have LWs (the first term in Eqa. 5), as the final vertex feature:

$$V_{i,j} = [\alpha_{i,j}^{kn_1,kn_2,\cdots,kn_\kappa}, \langle \alpha_{i,j}^{kw_1,kw_2,\cdots,kw_\upsilon}, S\omega_{i,j}^{kw_1,kw_2,\cdots,kw_\upsilon} \rangle] \quad (5)$$

where $\kappa + \upsilon = K$ ($\kappa = 5$, $\upsilon = 5$ and $K = 10$ in this paper). Specifically, we pre-define 5 operations that have LWs, which are separable convolution and dilated convolution with kernel sizes of 1×3 or 1×5, respectively, and transposed convolution with a kernel size of 1×3. We also pre-define 5 operations that do not have LWs, which are max pooling, average pooling (both with scale of 1×3), upsampling using linear/nearest strategy, and identity mapping. The pseudo code of the vertex and edge feature encoding are detailed demonstrated in the supplementary document.

Edge features: While each vertex in a produced graph may have underlying relationships with all other vertices, we only define that a pair of vertices $V_{i,j}$ and $V_{j,m}$ in $G(V, E)$ are adjacent if their corresponding directed edges $DE_{i,j}$ and $DE_{j,m}$ in the CNN model are connected to the same node N_j. This process is to prevent the produced graph representation from having overly large number of edges and is illustrated in Fig.2(b) in yellow. Instead of only using a single binary value (0 or 1) to define the adjacency (connectivity) between vertices in the graph representation $G(V, E)$, we propose to use multi-dimensional edge features that provide more information about their relationships. Inspired by the fact that all directed edge's OPs and LWs are jointly optimized during the CNN architecture searching process, we define $E(i, j, m)$ as the mutual influence between a pair of directed edges $DE_{i,j}$ and $DE_{m,j}$ during the optimization process, which is used as the edge feature for corresponding adjacent vertices $V_{i,j}$ and $V_{j,m}$ in the graph representation. In practice, we propose to learn $E_{i,j,m}$ directly from $V_{i,j}$ and $V_{j,m}$:

$$E(i, j, m) = ERN(V_{i,j}, V_{j,m}) \quad (6)$$

Specifically, for each pair of adjacent vertices $V_{i,j}$ and $V_{j,m}$ of $\mathbb{R}^{1\times k}$, their vertices features are concatenated as a new 2-D matrix $V_{i,j,m}$ of $\mathbb{R}^{2\times k}$. Then, we feed $V_{i,j,m}$ to an edge relationship extraction network (ERN) to learn their relationship, i.e., edge feature $E(i, j, m)$. It should be noted that all ERNs are jointly trained with the personality recognition model in an end-to-end manner. This way, all ERNs are learned to generate the task-specific edge features (personality-related feature in this paper) from corresponding vertices pair.

Personality recognition model: In this paper, we formulate the personality recognition problem as a graph regression task. Particularly, we employ the state-of-the-art residual gated graph convolution neural network (residual GatedGCN) [7] as the personality recognition model to process produced person-specific graph representations. We empirically employ a network that consists of six GatedGCN layers and two fully connected (FC) layers (attached to the last GatedGCN layer) with ReLU activation and dropout (0.3) to concatenate all produced vertices features. The size of output layer is set to 5 to jointly recognize the *five* personality traits of extroversion (Ext), agreeableness (Agr), openness (Ope), conscientiousness (Con), and neuroticism (Neu).

4 EXPERIMENTS

4.1 Dataset

We evaluate the proposed approach on the NoXi dataset [8], which is a multi-lingual human-to-human dyadic interaction dataset that was designed to generate spontaneous interactions with emphasis on adaptive behaviours in unexpected situations. The dataset consists of 84 sessions in which one participant acts as an Expert and another acts as a Novice interacting on a chosen topic of expertise via video conferences. They were allowed to continue the conversation until it reached a natural end. During the interaction, participants can interrupt each other for either changing the topic or inducing a mild debate whenever possible. The NoXi dataset contains 84 pairs of audio-visual clips with participants' ages ranging from 21 to 50 years. The average and standard deviation of clips' duration are 18 mins 6 seconds and 6 mins 28 seconds, respectively. All participants provided the self-assessments of their Big-five personality traits using the Saucier's Mini-Markers [42].

4.2 Implementation details

Neural architecture search: In this paper, all person-specific processor architectures have 6 blocks for each encoder (a pre-defined convolution layer, 3 down-sampling blocks and 2 regular blocks) and 5 decoder blocks (2 regular blocks and 3 up-sampling blocks), while the employed LSTMs have 3 hidden layers. During the neural architecture search, for each trial the input speaker's audio-visual signal lasts for 80 frames and the listener's candidate ground-truth consists of 105 frames, i.e., the delay factor r ranges from 0 to 25 frames. Meanwhile, the batch size was set to 60 audio-visual clips and 2 Adam optimizers were independently used to adjust OP α and LWs ω, with the learning rate of 0.05 and 0.001, respectively. In this paper, all experiments were conducted on Pytorch platform using Nvidia V100 GPUs.

GatedGCN and ERN: We evaluated graphical representations built on four types of vertex features: OP only, LW only, and two OP-LW combinations. For all experiments, the size of the first GatedGCN layer is set to be the same as the dimensionality of the input vertex feature d_{in}. The size of the rest of the five GatedGCN layers were set to be 11, 10, 10, 5 and 5 for OP features. For LW feature and OP-LW combination features, these were set to $\lfloor d_{in}/2 \rfloor$, $\lfloor d_{in}/4 \rfloor$, $\lfloor d_{in}/8 \rfloor$, $\lfloor d_{in}/16 \rfloor$ and $\lfloor d_{in}/32 \rfloor$, where $\lfloor\rfloor$ denotes the floor function. Each ERN used in our experiments is made up of two

	Methods	Ope	Con	Ext	Agr	Neu	Avg.
ACC	Spectral [45]	0.752	0.807	0.849	0.800	0.788	0.799
	DCC [18]	0.755	0.787	0.772	0.736	0.791	0.768
	NJU-LAMDA [51]	0.741	0.826	0.827	0.753	0.789	0.787
	CR-Net [29]	0.830	0.876	0.904	0.887	0.903	0.880
	PALs [44]	0.845	0.819	0.916	0.837	0.911	0.866
	Ours (A-MModal (S))	0.833	0.890	0.913	0.869	0.917	0.884
	Ours (MModal (M))	**0.889**	**0.925**	0.923	**0.913**	0.921	0.914
	Ours (A-MModal (M))	0.882	**0.925**	**0.931**	0.912	**0.925**	**0.915**
PCC	Spectral [45]	-0.010	0.059	0.135	0.071	0.024	0.056
	DCC [18]	-0.153	-0.078	0.037	-0.024	0.121	0.008
	NJU-LAMDA [51]	-0.110	0.118	0.115	-0.067	0.032	0.017
	CR-Net [29]	0.117	0.188	0.249	0.196	0.251	0.200
	PALs [44]	0.129	0.091	0.27	0.106	0.264	0.172
	Ours (A-MModal (S))	0.097	0.193	0.296	0.163	0.313	0.212
	Ours (MModal (M))	**0.181**	0.324	0.353	**0.265**	0.351	0.295
	Ours (A-MModal (M))	0.168	**0.335**	**0.371**	0.249	**0.382**	**0.301**

Table 1: Personality recognition results on the NoXi dataset. *MModal* denotes the graph representations of multi-modal processors. *(M)* and *(S)* represent the multi-level and single-level fusion, respectively. *A-* represents that the CNNs were trained with adaptive loss. For all our systems, graph representations were obtained using OP-LW (W) vertices features while GatedGCN with edge features was employed as the personality recognition model.

convolution blocks, each containing a 1-D convolution layer, a batch normalization layer, and a ReLU activation function. In our experiments, we set the dimension of the edge features the same as that of the vertex features and utilise a 7-fold subject-independent cross-validation strategy. For each fold, 154 videos were used for training and hyperparameter optimisation, and 14 videos for testing. We report the accuracy on the test sets averaged over 7 folds.

Evaluation metrics: Two common metrics are used to evaluate the personality recognition performance, the Pearson Correlation Coefficient (PCC) and the mean accuracy measurement (ACC), as adopted in the relevant ChaLearn Challenge [39].

4.3 Comparison to existing approaches

Table 1 compares the variations of the proposed model achieving the best results to existing state-of-the-art audio-visual personality analysis approaches (automatic personality perception (APP) solutions) - all were evaluated on the NoXi dataset. As can be observed, the predictions produced by our multi-modal systems are positively correlated with all self-reported personality traits, both of which achieved PCC>0.32 for Con, Ext and Neu traits. These results indicate that despite CNNs and humans having different cognitive mechanisms, if a CNN can simulate the cognitive process for generating facial reactions of a target subject, its architectural parameters are *well* associated with this subject's personality traits. Also, compared to existing solutions that directly predict personality traits from non-verbal behaviours, the proposed approach that recognizes personality traits from the simulated cognition, is more reliable for true (self-reported) personality traits recognition. In addition, we found that graph representations of our multi-modal

Cognitive model	PCC	MSE ($\times 10^{-5}$)
Audio-to-face	0.769	2.882
Face-to-face	**0.770**	2.894
PS-Multi-to-face (S)	**0.781**	2.390
IP-Multi-to-face (S)	0.775	2.503
A-IP-Multi-to-face (S)	**0.781**	2.260
PS-Multi-to-face (M)	0.794	2.362
IP-Multi-to-face (M)	0.798	2.245
A-IP-Multi-to-face (M)	**0.802**	2.331

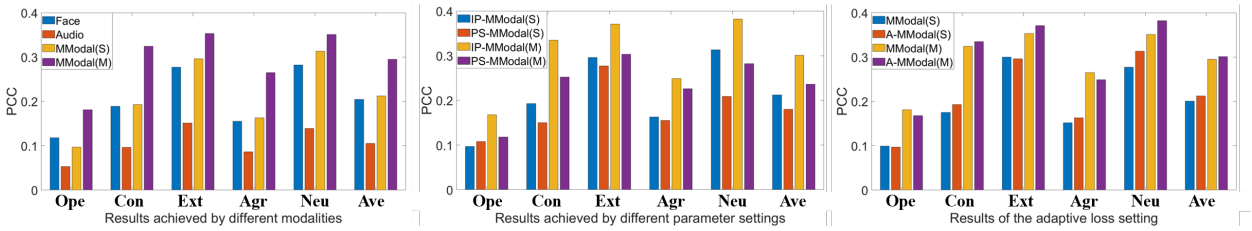
Table 2: Facial reactions prediction performance. *PS-* and *IP-* denote the parameter sharing and independent parameter strategy, respectively; *Multi-* refers to the multi-modal audio and face information extracted from the speaker; *A-* re

systems achieved the best ACC performance for all five traits. Also, it can be observed that the approaches which predict personality using clip-level features (e.g., CR-Net, PALs and Spectral approach) have clear advantages over the approaches that infer personality from a single frame or thin slices (DCC and NJU-LAMDA), demonstrating that long-term information may be more informative for modelling personality traits.

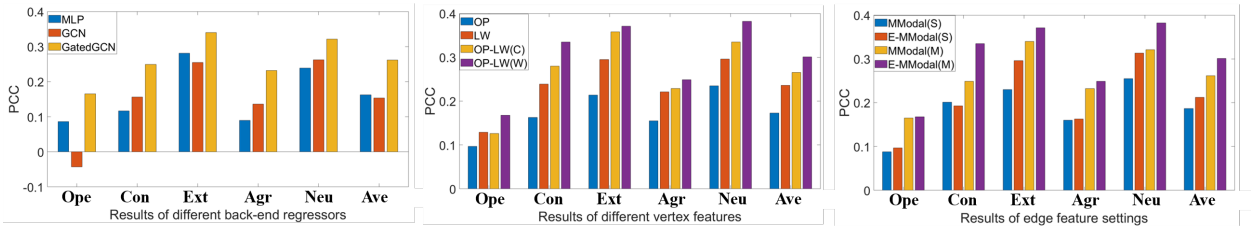
4.4 Ablation studies

Person-specific CNN: We show the personality recognition performance achieved by the proposed approach with different person-specific CNN settings in Fig. 5(a) and the facial reaction results in Table.2. It can be observed that the predictions generated by all model variations are positively correlated with the self-reported values across all five traits. In general, multi-modal systems achieved better results than single-modal systems for both personality recognition and facial reaction prediction tasks. This demonstrates that both audio and facial behaviours of the speaker have impact on the listener's facial reactions, and when modelling the cognitive processes for reaction generation, each modality provides unique and useful clues for personality recognition. For multi-modal systems, multi-level fusion not only allows the searched person-specific architectures to simulate better facial reactions than the single-level fusion setting, but also yields better personality recognition performance. This is due to the fact that the multi-level fusion strategy allows the CNN internal feature maps (cognitive processes triggered by the speakers' audio and facial behaviours) to interact at multiple cognitive levels.

Meanwhile, the **independent parameter (IP)** setting largely improved the personality recognition performance over the widely-used **parameter sharing (PS)** strategy [30, 37]. In particular, the prediction performance of Ope, Con and Neu traits benefited from the IP setting with 4.4%, 2.4% and 1.6% ACC improvements and 42.4%, 32.9% and 35.5% PCC improvements, respectively. This validated our assumption that we should use different CNN blocks to simulate the different functions of human cognitive processes. While most of our systems have achieved good performance in recognizing Con, Ext and Neu traits, the use of the adaptive loss provided further improvements for the recognition of these traits. Since a similar conclusion, i.e., Ext and Neutraits are well associated with human cognition, has been frequently claimed by previous studies, we assume that the proposed adaptive loss allows the



(a) Personality recognition results of different person-specific cognitive processor settings.



(b) Personality recognition results of different recognition model settings.

Figure 5: Ablation studies results, where the definition of *MModal, S, M, PS-, IP-, A-* can be found in the captions of Table. 1 and Table. 2. *E-* represents the use of end-to-end learned edge features.

searched CNN simulate better the real human cognitive processes. Based on the architectural parameters of multi-level fusion of the multi-modal person-specific processors, Fig. 5(b) compares the performance achieved by different personality recognition models and graph representation settings.

Graph representation: We found that when using LW or the simple concatenation of OP and LW (OP-LW (C)) as the vertex feature, their graph representations have very similar capability for recognizing personality, both of which outperformed the graph representations that only use OP as the vertex feature. This can be explained by the fact that the OP vertex feature ignores all layer weights which are crucial in deciding CNNs' generalization capabilities. In other words, clues for inferring personality traits reside in both OP and their LW. More importantly, the graph representation constructed by the proposed operation parameter-based weighted layers' weights (OP-LW (W)) outperformed the OP-LW (C) setting across all five traits, demonstrating that such weighting strategy is a superior way to construct vertex features.

Personality recognition model: Almost all models achieved positive correlations across all traits. Moreover, even the system that simply uses an MLP with four hidden layers to process person-specific representations (i.e., 1-D vectors that are produced by concatenating all OP and LW of each person-specific processor, respectively and then processed by correlation-based feature selection (CFS) [20].) can make predictions that show clear positive correlations ($PCC > 0.1$) with the self-reported Con, Ext and Neu traits. This demonstrates that the optimal CNNs architectures found by our approach are *well* associated with target subjects' personality traits. Although the standard GCN model only generated a similar performance to the MLP, the GatedGCN model achieved a large improvement. In particular, the proposed end-to-end edge feature learning strategy (edge-GatedGCN) further enhanced the personality recognition performance over the GatedGCN that only uses binary adjacency matrix to define edges. We conclude that the proposed end-to-end edge feature learning strategy helps GatedGCN extract additional task-specific clues for personality recognition.

5 CONCLUSION

Our approach recognizes true personality from the simulated person-specific human cognition, which is represented as a graph of person-specific CNN architecture. Our experimental results show that CNNs' architectural parameters are positively associated with the self-reported personality traits. Additional modalities might enable better CNN model learning and can in principle be combined via the proposed fusion module. One limitation of this work is that searching a person-specific CNN for each subject takes a relatively longer period of time compared to existing approaches, and therefore it may not be suitable for *fast* personality assessment requirements. Nevertheless, this work opens up a new avenue of research for recognizing socio-emotional phenomena (personality, affect, engagement, mental health) from the simulations of person-specific cognitive processes that will have further implications for relevant fields including neuroscience, cognitive and behavioural sciences, and social robotics [1], [6] (e.g., for creating data-driven robot coaches that can express personalized behaviours during dyadic interactions.).

ACKNOWLEDGMENTS

Linlin Shen and Zilong Shao are supported by the National Natural Science Foundation of China under Grant 91959108. Siyang Song and Hatice Gunes are partially supported by the European Union's Horizon 2020 Research and Innovation Programme, under the WorkingAge Project (grant agreement No.826232). Hatice Gunes is also supported by the EPSRC (Grant Ref: EP/R030782/1).

REFERENCES

[1] Minja Axelsson, Indu P. Bodala, and Hatice Gunes. 2021. Participatory Design of A Robotic Mental Well-being Coach. In *Proc. IEEE Int'l Conference on Robot and Human Interactive Communication (RO-MAN 2021)*.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[3] Salah Eddine Bekhouche, Fadi Dornaika, Abdelkrim Ouafi, and Abdelmalik Taleb-Ahmed. 2017. Personality traits and job candidate screening via analyzing facial videos. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 1660–1663.

[4] Cigdem Beyan, Andrea Zunino, Muhammad Shahid, and Vittorio Murino. 2019. Personality Traits Classification Using Deep Visual Activity-based Nonverbal Features of Key-Dynamic Images. *IEEE Transactions on Affective Computing* (2019).

[5] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. 2016. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3034–3042.

[6] Indu P. Bodala, Nikhil Churamani, and Hatice Gunes. 2021. Teleoperated Robot Coaching for Mindfulness Training: A Longitudinal Study. In *Proc. IEEE Int'l Conference on Robot and Human Interactive Communication (RO-MAN 2021)*.

[7] Xavier Bresson and Thomas Laurent. 2017. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553* (2017).

[8] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 350–359.

[9] Stuartk Card, THOMASP MORAN, and Allen Newell. 1986. The model human processor- An engineering model of human performance. *Handbook of perception and human performance*. 2, 45–1 (1986).

[10] Oya Celiktutan and Hatice Gunes. 2017. Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability. *IEEE Transactions on Affective Computing* 8, 1 (2017), 29–42.

[11] O. Celiktutan, E. Skordos, and H. Gunes. 2019. Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement. *IEEE Transactions on Affective Computing* 10, 4 (2019), 484–497. https://doi.org/10.1109/TAFFC.2017.2737019

[12] Benoît Colson, Patrice Marcotte, and Gilles Savard. 2007. An overview of bilevel optimization. *Annals of operations research* 153, 1 (2007), 235–256.

[13] Kathleen A Corrigall, E Glenn Schellenberg, and Nicole M Misura. 2013. Music training, cognition, and personality. *Frontiers in psychology* 4 (2013), 222.

[14] Colin G DeYoung, Jacob B Hirsh, Matthew S Shane, Xenophon Papademetris, Nallakkandi Rajeevan, and Jeremy R Gray. 2010. Testing predictions from personality neuroscience: Brain structure and the big five. *Psychological science* 21, 6 (2010), 820–828.

[15] Sefik Emre Eskimez, Ross K Maddox, Chenliang Xu, and Zhiyao Duan. 2018. Generating talking face landmarks from speech. In *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 372–381.

[16] Michael W Eysenck and Mark T Keane. 2005. *Cognitive psychology: A student's handbook*. Taylor & Francis.

[17] Sheng Fang, Catherine Achard, and Séverine Dubuisson. 2016. Personality classification and behaviour interpretation: An approach based on feature categories. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 225–232.

[18] Yağmur Güçlütürk, Umut Güçlü, Marcel AJ van Gerven, and Rob van Lier. 2016. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *European Conference on Computer Vision*. Springer, 349–358.

[19] H. Gunes, O. Çeliktutan, and E. Sariyanidi. 2019. Live human-robot interactive public demonstrations with automatic emotion and personality prediction. *Philosophical Transactions of the Royal Society B* 374, 1771 (2019), 1–8.

[20] Mark Andrew Hall. 1999. Correlation-based feature selection for machine learning. (1999).

[21] Jonathan Jackson, David A Balota, and Denise Head. 2011. Exploring the relationship between personality and regional brain volume in healthy aging. *Neurobiology of aging* 32, 12 (2011), 2162–2171.

[22] V Kaasinen, RP Maguire, T Kurki, A Brück, and JO Rinne. 2005. Mapping brain structure and personality in late adulthood. *Neuroimage* 24, 2 (2005), 315–322.

[23] Saul M Kassin. 2003. *Essentials of psychology*. Prentice Hall.

[24] Heysem Kaya, Furkan Gurpinar, and Albert Ali Salah. 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*. 1–9.

[25] Otto F Kernberg. 2016. What is personality? *Journal of Personality Disorders* 30, 2 (2016), 145–156.

[26] Nathan Kogan and Michael A Wallach. 1964. Risk taking: A study in cognition and personality. (1964).

[27] Veena Kumari, Steven CR Williams, Jeffrey A Gray, et al. 2004. Personality predicts brain responses to cognitive demands. *Journal of Neuroscience* 24, 47 (2004), 10636–10641.

[28] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. 2016. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710* (2016).

[29] Yunan Li, Jun Wan, Qiguang Miao, Sergio Escalera, Huijuan Fang, Huizhou Chen, Xiangda Qi, and Guodong Guo. 2020. CR-Net: A Deep Classification-Regression Network for Multimodal Apparent Personality Analysis. *International Journal of Computer Vision* (2020), 1–18.

[30] Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2018. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055* (2018).

[31] S. Mariooryad and C. Busso. 2013. Exploring Cross-Modality Affective Reactions for Audiovisual Emotion Recognition. *IEEE Transactions on Affective Computing* 4, 2 (2013), 183–196. https://doi.org/10.1109/T-AFFC.2013.11

[32] Robert R McCrae. 1993. Openness to experience as a basic dimension of personality. *Imagination, Cognition and Personality* 13, 1 (1993), 39–55.

[33] W. Mou, H. Gunes, and I. Patras. 2019. Your Fellows Matter: Affect Analysis across Subjects in Group Videos. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*. 1–5. https://doi.org/10.1109/FG.2019.8756514

[34] Laurent Son Nguyen, Alvaro Marcos-Ramiro, Martha Marrón Romera, and Daniel Gatica-Perez. 2013. Multimodal analysis of body communication cues in employment interviews. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 437–444.

[35] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. 2015. Personality Trait Classification via Co-Occurrent Multiparty Multimodal Event Discovery. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 15–22.

[36] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. 2015. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 15–22.

[37] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. 2018. Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning*. PMLR, 4095–4104.

[38] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). 400–418.

[39] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European Conference on Computer Vision*. Springer, 400–418.

[40] Ricardo Darío Pérez Principi, Cristina Palmero, Julio C Junior, and Sergio Escalera. 2019. On the Effect of Observed Subject Biases in Apparent Personality Analysis from Audio-visual Signals. *IEEE Transactions on Affective Computing* (2019).

[41] Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2016. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access* 5 (2016), 705–721.

[42] Gerard Saucier. 1994. Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of personality assessment* 63, 3 (1994), 506–516.

[43] K Warner Schaie, Sherry L Willis, and Grace IL Caskie. 2004. The Seattle longitudinal study: Relationship between personality and cognition. *Aging Neuropsychology and Cognition* 11, 2-3 (2004), 304–324.

[44] Siyang Song, Shashank Jaiswal, Enrique Sanchez, Georgios Tzimiropoulos, Linlin Shen, and Michel Valstar. 2021. Self-supervised Learning of Person-specific Facial Dynamics for Automatic Personality Recognition. *IEEE Transactions on Affective Computing* (2021).

[45] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. 2020. Spectral Representation of Behaviour Primitives for Depression Analysis. *IEEE Transactions on Affective Computing* (2020).

[46] Siyang Song, Enrique Sánchez-Lozano, Mani Kumar Tellamekala, Linlin Shen, Alan Johnston, and Michel Valstar. 2019. Dynamic Facial Models for Video-based Dimensional Affect Estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

[47] Siyang Song, Linlin Shen, and Michel Valstar. 2018. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 158–165.

[48] Lucía Teijeiro-Mosquera, Joan-Isaac Biel, José Luis Alba-Castro, and Daniel Gatica-Perez. 2015. What your face vlogs about: expressions of emotion and big-five traits impressions in YouTube. *IEEE Transactions on Affective Computing* 6, 2 (2015), 193–205.

[49] Carles Ventura, David Masip, and Agata Lapedriza. 2017. Interpreting cnn models for apparent personality trait regression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 1705–1713.

[50] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing* 5, 3 (2014), 273–291.
[51] Xiu-Shen Wei, Chen-Lin Zhang, Hao Zhang, and Jianxin Wu. 2018. Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing* 9, 3 (2018), 303–315.
[52] Lee Willerman, Robert Schultz, J. N. A Rutledge, and Erin Bigler. 1994. *Brain Structure and Cognitive Function.* Soc Neuroscience.
[53] Le Zhang, Songyou Peng, and Stefan Winkler. 2019. PersEmoN: A deep network for joint analysis of apparent personality, emotion and their relationship. *IEEE Transactions on Affective Computing* (2019).