MULTIMODAL VIDEO-BASED PERSONALITY RECOGNITION USING LONG SHORT-TERM MEMORY AND CONVOLUTIONAL NEURAL NETWORKS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Süleyman Aslan

July 2019

Multimodal Video-based Personality Recognition Using Long Short-Term Memory and Convolutional Neural Networks
By Süleyman Aslan
July 2019

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Uğur Güdükbay(Advisor)

Selim Aksoy

Ramazan Gökberk Cinbiş

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

MULTIMODAL VIDEO-BASED PERSONALITY RECOGNITION USING LONG SHORT-TERM MEMORY AND CONVOLUTIONAL NEURAL NETWORKS

Süleyman Aslan
M.S. in Computer Engineering
Advisor: Uğur Güdükbay
July 2019

Personality computing and affective computing, where recognition of personality traits is essential, have gained increasing interest and attention in many research areas recently. The personality traits are described by the Five-Factor Model along five dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism. We propose a novel approach to recognize these five personality traits of people from videos. Personality and emotion affect the speaking style, facial expressions, body movements, and linguistic factors in social contexts, and they are affected by environmental elements. For this reason, we develop a multimodal system to recognize apparent personality traits based on various modalities such as the face, environment, audio, and transcription features. In our method, we use modality-specific neural networks that learn to recognize the traits independently and we obtain a final prediction of apparent personality with a feature-level fusion of these networks. We employ pre-trained deep convolutional neural networks such as ResNet and VGGish networks to extract high-level features and Long Short-Term Memory networks to integrate temporal information. We train the large model consisting of modality-specific subnetworks using a two-stage training process. We first train the subnetworks separately and then fine-tune the overall model using these trained networks. We evaluate the proposed method using ChaLearn First Impressions V2 challenge dataset. Our approach obtains the best overall "mean accuracy" score, averaged over five personality traits, compared to the state-of-the-art.

Keywords: deep learning, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) network, personality traits, personality trait recognition, multimodal information.

ÖZET

ÇOK KİPLİ UZUN KISA-SÜRELİ BELLEK VE EVRİŞİMLİ SİNİR AĞLARI İLE VİDEODA KİŞİLİK TANIMA

Süleyman Aslan
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Uğur Güdükbay
Temmuz 2019

Kişilik özelliklerinin tanınmasının gerekli olduğu kişilik hesaplama ve duygusal hesaplama, son zamanlarda birçok araştırma alanında artan ilgi ve dikkate sahip olmuştur. Kişilik özellikleri, Beş Faktörlü Model tarafından beş boyutta tanımlanmaktadır: açıklık, sorumluluk, dışadönüklük, uyumluluk, ve duygusallık. Biz, insanların bu beş kişilik özelliklerini videolardan tanımak için yeni bir yaklaşım öneriyoruz. Kişilik ve duygu, konuşma tarzını, yüz ifadelerini, vücut hareketlerini ve sosyal bağlamdaki dilsel faktörleri etkiler ve ayrıca çevresel unsurlardan etkilenir. Bu nedenle, yüz, çevre, ses ve çevriyazı özellikleri gibi çeşitli kiplere dayanan belirgin kişilik özelliklerini tanımak için çok kipli bir sistem geliştiriyoruz. Yöntemimizde, özellikleri bağımsız olarak tanımayı öğrenen kipe özgü sinir ağları kullanıyoruz ve son bir belirgin kişilik tahminini bu ağların öznitelik düzeyinde bir kaynaşımı ile elde ediyoruz. Yüksek düzey öznitelikleri bulmak için ResNet ve VGGish ağları gibi önceden eğitilmiş derin evrişimli sinir ağlarını ve zamansal bilgiyi bütünleştirmek için uzun kısa-süreli bellek ağlarını kullanıyoruz. Kipe özgü alt ağlardan oluşan büyük modeli, iki-aşamalı bir eğitim yöntemi ile eğitiyoruz. İlk önce alt ağları ayrı olarak eğitiyoruz, ardından, genel modele bu eğitilmiş ağları kullanarak ince ayar yapıyoruz. Önerilen yöntemi "ChaLearn First Impressions V2 challenge" veri setini kullanarak değerlendiriyoruz. Yaklaşımımız, beş kişilik özelliklerinin "ortalama doğruluk" puanlarının ortalaması alındığında literatürdeki yöntemlere göre en iyi sonuçları elde etmektedir.

Anahtar sözcükler: derin öğrenme, Evrişimli Sinir Ağları, Tekrarlayan Sinir Ağları, Uzun Kısa-Süreli Bellek ağları, kişilik özellikleri, kişilik özellikleri tanıma, çok kipli bilgiler.

Acknowledgement

I would first like to express my gratitude to my thesis advisor Prof. Dr. Uğur Güdükbay for the continuous guidance and engagement through my MSc study and research. Throughout the writing of this thesis I have received a great deal of assistance. His expertise was very helpful in all stages of this work.

I would also like to thank to the members of the jury, Assoc. Prof. Dr. Selim Aksoy and Asst. Prof. Dr. Ramazan Gökberk Cinbiş, for their insightful comments and valuable questions.

Last but not the least, I owe more than thanks to my family members: my parents and my brother for their love, encouragement and support throughout my life. They are always so helpful to me in numerous ways and encouraging me in whatever I pursue. Without their guidance and support, it would not have been possible for me to successfully complete this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Personality and emotions have a strong influence on people's lives and they affect behaviors, cognitions, preferences, and decisions. Emotions have distinct roles in decision making, such as providing information about pleasure and pain, enabling rapid choices under time pressure, focusing attention on relevant aspects of a problem, and generating commitment concerning decisions [1]. Additionally, research suggests that human decision making process can be modeled as a two systems model, consisting of rational and emotional systems [2]. Accordingly, emotions are part of every decision making process instead of simply having an effect on these processes. Likewise, personality also has an important effect on decision making and it causes individual differences in people's thoughts, feelings, and motivations. It can be observed that there are significant relationships among attachment styles, decision making styles, and personality traits [3]. In addition, personality relates to individual differences in preferences, such as the use of music in everyday life [4, 5], and user preferences in multiple entertainment domains including books, movies, and TV shows [6]. Due to the fact that emotion and personality have an essential role in human cognition and perception, there has been a growing interest in recognizing the human personality and affect and integrating them into computing to develop artificial emotional intelligence, which is also known as "affective computing" [7], in combination with "personality computing" [8]. Hence, it becomes essential to recognize the personality and

emotion of humans precisely. Thereby, in this thesis we present a novel multimodal framework to recognize the personality traits of individuals from videos to address this problem.

1.1 Personality Traits

Personality can be defined as the psychological factors that influence an individual's patterns of behaving, thinking, and feeling that differentiate the individual from one another [9, 10]. The most mainstream and widely accepted framework for personality among psychology researchers is the Five-Factor Model (FFM) [11, 10]. FFM is a model based on descriptors of human personality along five dimensions as a complete description of personality. Various researchers have identified the same five factors within independent works in personality theory [11, 12, 13]. Therefore, it is considered reliable to define personality with FFM.

Based on the work by Costa, McCrae, and John [11, 10], the five factors are defined as follows.

- *Openness (O):* Appreciation of experience and curiosity of the unfamiliar.
- *Conscientiousness (C):* Level of organization and being dependable.
- *Extraversion (E):* Social activity and interpersonal interaction.
- *Agreeableness (A):* Tendency to work cooperatively with others and avoiding conflicts.
- *Neuroticism (N):* Emotional instability and being prone to psychological distress.

These five factors lead to bipolar characteristics that can be seen in individuals that score low and high on each trait, as seen in Table 1.1. The factors are often

Table 1.1: The characteristics of personality traits.

Low Scorer	*Personality Trait*	*High Scorer*
Calm, secure	Neuroticism	Nervous, sensitive
Quiet, reserved	Extraversion	Talkative, sociable
Cautious, conventional	Openness	Inventive, creative
Suspicious, uncooperative	Agreeableness	Helpful, friendly
Careless, negligent	Conscientiousness	Organized, reliable

represented by the acronym *OCEAN*. They are also known as "Big Five", as Goldberg states that "any model for structuring individual differences will have to encompass at some level something like these 'Big Five' dimensions" [14].

1.2 Background and Problem Definition

In recent years, there has been a growing interest in incorporating personality traits into multi-agent and artificial intelligence-based systems. Since FFM is commonly accepted as an accurate description of personality, it has been used for modeling of human behavior in agents as well [15]. Additionally, it has been shown that OCEAN factors, hence FFM, can be used as a basis for agent psychology to simulate the behavior of animated virtual crowds [16, 17] and FFM is suitable for agent-based simulations [18]. Therefore, FFM is influential in the simulation of autonomous agents.

One shortcoming of the mentioned works is that the personality traits in FFM still need to be provided to the system manually. Durupınar et al. [16] propose that parameters underlying crowd simulators can be mapped to OCEAN personality traits so that instead of the low-level parameter tuning process, higher-level concepts related to human psychology can be chosen for the simulation. However, in that work, they handpick traits to demonstrate various crowd behaviors, such as simulating people with low conscientiousness and agreeableness. Although the

effect of personality traits on the behavior of agents can be seen, selecting appropriate traits can be tedious and unscalable in the case of creating agents with unique personalities in large numbers.

In order to integrate the personality of a human into a virtual system, it would be needed to obtain the personality traits using commonly used measures such as NEO PI-R [19] or FFMRF [20] and then apply the obtained traits with existing approaches. The applications of behavior simulations are rapidly growing so this field of computer science seeks the development of automatic assessment of personality. Hence, this thesis proposes a novel two-stage multimodal system to obtain the personality traits automatically in a data-driven manner. Our proposed model predicts the traits based on various modalities including facial, ambient, audio, and transcription features and the model makes use of convolutional neural networks (CNNs) in combination with Long Short-Term Memory networks (LSTMs) in a two-stage training phase as explained in detail in Chapter 3. The proposed approach obtains state-of-the-art results using ChaLearn First Impressions V2 (CVPR'17) challenge dataset [21]. In the following chapters, an overview of previous work done in personality recognition, an outline of the limitations and problems of existing approaches, and the details of the proposed approach to recognition of personality traits are provided.

1.3 Contributions

This thesis contributes to the area of automatic recognition of people's personality from videos. The main contributions are:

i) A multimodal neural network architecture to recognize apparent personality traits from various modalities such as the face, environment, audio, and transcription features. The system consists of modality-specific deep neural networks that aim to predict apparent traits independently where the overall prediction is obtained with a fusion.

ii) Integrating the temporal information of the videos learned by LSTM networks to the extracted spatial features with CNNs such as facial expressions and ambient features.

iii) A two-stage training method that trains the modality-specific networks separately in the first stage and fine-tunes the overall model to recognize the traits accurately in the second stage.

1.4 Outline of the Thesis

Chapter 2 scopes the focus of this thesis and reviews the related work. Chapter 3 presents the proposed method which effectively learns a mapping from multimodal data to personality trait vectors. Chapter 4 shows the results of the approach and evaluates it with different quality aspects. Chapter 5 concludes the thesis and presents some future research directions.

Chapter 2

Related Work

Personality computing benefits from methods aimed towards understanding, predicting, and synthesizing human behavior [8]. The effectiveness in analyzing such important aspects of individuals is the main reason behind the growing interest in this topic. Automatic recognition of personality is a part of many applications such as human-computer interaction, computer-based learning, automatic job interviews, and autonomous agents [22, 23, 21, 16, 17]. Similarly, emotion is incorporated into adaptive systems in order to improve the effectiveness of personalized content and bring the systems closer to the users [24]. As a result, personality and emotion-based user information is used in many systems, such as affective e-learning [25], conversational agents [26], and recommender systems [27, 28, 29]. Overall, personality is usually relevant in any system involving human behavior. Rapid advances in personality computing and affective computing led to the releases of novel datasets for personality traits and emotional states of people from various sources of information such as physiological responses or video blogs [30, 21]. One of the latest problems is recognizing apparent five personality traits automatically from videos of people speaking in front of a camera.

2.1 Personality Recognition

Recently, there have been many approaches to recognizing personality traits. By analyzing the audio from spoken conversations [31] and based on the tune and rhythm aspects of speech [32] it is possible to annotate and recognize the personality traits or predict the speaker attitudes automatically. These approaches demonstrate that audio information is important for personality. Moreover, nonverbal aspects of verbal communication such as linguistic cues in conversations can be used to predict speaker's personality traits [33] and for this task, it is shown that models trained on observed personality have better performance than models trained using self-reports [33].

One other usage of nonverbal aspects is predicting personality in the context of human-human spoken conversations independently from the speakers [34]. These approaches provide automatic analysis of personality traits which is quite complex in nature. Similarly, using users' status text on social networks can be a way of recognition of personality traits [35] and it is also possible to explore the projection of personality, especially extraversion, through specific linguistic factors across different social contexts using transcribed video blogs and dialogues [36]. Therefore, it is indicated that there is a strong correlation between users' behavior on social networks and their personality [37]. Additionally, it is observed that the effective verbal content of video transcripts and gender have a predictive effect on personality impressions [38].

There are other methods for recognition based on combinations of speaking style and body movements. Personality traits can be automatically detected in social interactions from acoustic features encoding specific aspects of the interaction and visual features such as head, body, and hands fidgeting [39]. Likewise, five-factor personality traits can be automatically detected in short self-presentations based on the effectiveness of acoustic and visual non-verbal features such as pitch, acoustic intensity, hand movement, head orientation, posture, mouth fidgeting, and eye-gaze [40].

The impact of body movements and speaking style is examined further in other studies. For example, automatic social behavior analysis for subjects involved in the experimental sessions is performed using audio-visual cues including pitch and energy, hand and body fidgeting, speech rate, and head orientation [41]. In addition to this, meeting behaviors, namely, speaking time and social attention are shown to be effective for the detection and classification of the extraversion personality trait [42]. An exploratory study involving a professional speaker producing speech using different personality traits demonstrates that there is a high consistency between the acted personalities, human raters' assessments, and automatic classification outcomes [43]. As a result, it can be seen that body gestures, head movements, facial expressions, and speech based on naturally occurring human affective behaviour leads to effective assessment of personality and emotion [44]. Additionally, by using speaking activity, prosody, visual activity, and estimates of facial expressions of emotion as features, it is possible to perform automatic analysis of natural mood and impressions in conversational social videos [45, 46].

There are other means of predicting personality and emotion. For example, an analysis of human affective states and emotion using physiological signals indicates that there is a correlation between the signals, personality, emotional behavior, and participant's ratings for music videos [47, 48, 49, 50, 51]. Likewise, in a human-computer interaction (HCI) scenario, analyzing spontaneous emotional responses of participants to effective videos can be a method for inferring the five-factor personality traits [52]. In a similar manner, users' personality in human-computer interaction can be automatically recognized from videos in which there are different levels of human-computer collaborative settings [53]. These show that personality plays an important role in human-computer interaction. In human behavior analysis, behavioral indicators based on communicative cues that are present in the conversations coming from the field of psychology are effective for the analysis of non-verbal communication in order to predict satisfaction, agreement, and receptivity in the conversations [54, 55]. On top of these, it has been demonstrated that the interaction between human beings and computers becomes more natural when human emotions are recognized by

computers, which can be done by an analysis of facial expressions and acoustic information [56].

Apart from emotions, facial physical attributes from ambient face photographs can be an important factor in modeling trait factor dimensions underlying social traits [57]. It can be seen that valid inferences for personality traits can be made from the facial attributes. This is supported by experiments that are carried out in order to evaluate personality traits and intelligence from facial morphological features [58], to predict the personality impressions for a given video depicting a face [59], and to identify the personality traits from a face image [60].

Some studies support the idea that it is human behavior to evaluate individuals by their faces with respect to their personality traits and intelligence since self-reported personality traits can be predicted reliably from a facial image [58], and impressions that influence people's behavior towards other individuals can be accurately predicted from videos [59]. Additionally, it has been shown that prediction of impressions can be done by obtaining visual-only and audio-only annotations continuously in time to learn the temporal relationships by combining these visual and audio cues [61], whereas in some other work, predicting personality factors for personality-based sentiment classification is shown to be beneficial in the analysis of public sentiment implied in user-generated content [62]. Accordingly, it can be seen that the personality has an effect on various different modalities, therefore in many studies, automatic recognition of personality traits is accomplished by combining multiple features to present a multimodal approach [39, 41, 53, 54, 56, 59, 61].

According to [63], attributes and features such as audio-visual, text, demographic and sentiment features are essential parts of a personality recognition system. Likewise, measuring personality traits depends on different behavioral cues in daily communication including linguistic, psycholinguistic and emotional features, based on research in behavioral signal processing [64]. Another multimodal approach uses automatic visual and vocal analysis of personality traits

and social dimensions [65]. Finally, a multivariate regression approach for predicting how the personality of YouTube video bloggers is perceived by their viewers performs better when compared to a single target approach [66]. Although multimodal approaches are commonly used to recognize personality traits, relatively limited work has been done to present a comprehensive method utilizing a considerable amount of informative features. In this work, we propose such a comprehensive method to recognize personality trait factors.

2.2 Deep Learning in Personality Computing

According to Wright's commentary on Personality Science [67], there are issues that are neglected in Personality Computing, such as the hierarchical structure of personality traits and "person-situation integration". Vinciarelli et al. state that deep learning approaches might be able to address these issues by capturing the hierarchical structure underlying personality traits [68]. Moreover, modeling context becomes necessity because of person-situation problem. One other problem is that computing approaches require simplifications because low-level information extracted from sensor data is not adequate to capture the complexity of high-level information like personality traits and this is the main reason behind issues for current Automatic Personality Recognition (APR) approaches [68].

Deep learning architectures, being able to automatically discover and represent multiple levels of abstraction [69], might be able to contribute towards addressing the problems. These methods can extract and organize the discriminative information from the data, learn representations of the data, and extract more abstract features at higher layers of representations [70]. Therefore, in order to automatically infer people's personality, deep learning methods can be used to capture complex non-linear features in multimodal data for the personality recognition task.

Recently, there have been automatic emotion recognition systems that predict high-level information from low-level signal cues, however, these methods

capture only linear relationships between features [71]. On the other hand, deep learning methods automatically learn suitable data representations, nevertheless, there are relatively few deep learning based approaches for automatic personality recognition [72, 73]. However, it can be observed that there have been successful deep learning based approaches for personality computing related tasks [74, 75, 76, 77, 78]. Consequently, we utilize the deep neural networks' capability to learn complex representations to effectively perform APR in this work.

Chapter 3

The Proposed Framework

In our framework, we take a video clip of a single person as input and predict the personality traits associated with that person. The proposed framework is based on learning personality features separately using different modality-specific neural networks, then combining those learned high-level features to obtain a final prediction of personality traits. For this purpose, four neural networks are trained independently to extract high-level features, namely, *ambient features*, *facial features*, *audio features*, and *transcription features*.

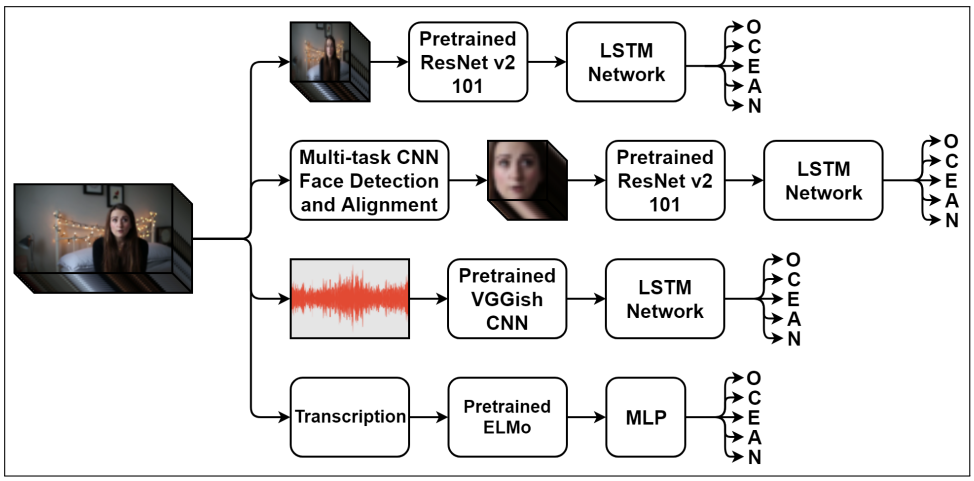

Figure 3.1: The first stage of the proposed model. Subnetworks learn to recognize personality traits based on the corresponding input features.

In this approach, there are two stages. In the first stage, each subnetwork is trained to obtain personality traits according to the various input features. The flowchart illustrating this first stage is given in Figure 3.1. Modality-specific networks are trained separately because first, to make sure that each network is able to learn corresponding features and improves the final prediction, and second, to prevent the model to focus on only one dominant feature in training phase. In the second stage, trained neural networks are used as feature extractors and the final trait scores are obtained through a fusion. The flowchart of second stage is given in Figure 3.2. We elaborate on each subnetwork in the following sections.

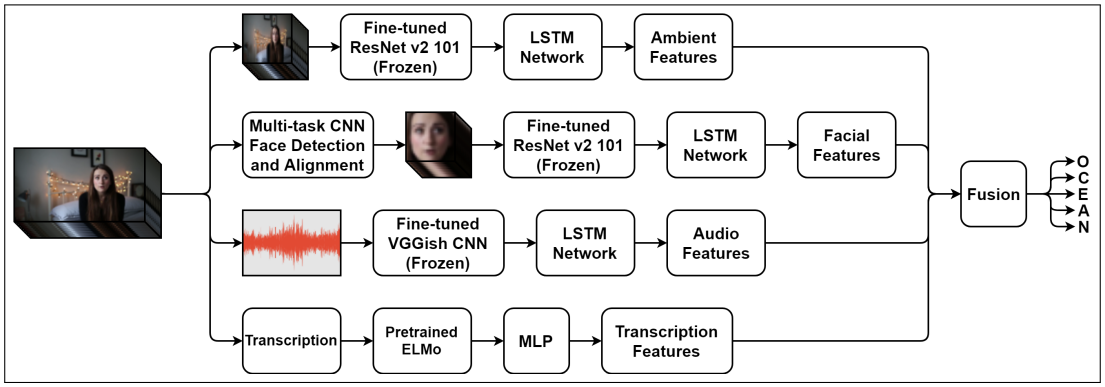

Figure 3.2: The second stage of the proposed model. We use trained subnetworks as feature extractors and fuse the results of them to obtain the final score for traits.

3.1 Ambient Feature-based Recognition

One of the approaches used in the proposed framework is to recognize personality traits based on the ambient features related to the person such as surrounding objects, lighting, and clothing. The intuition behind this approach is that those features can influence the apparent personality of the person. It has been demonstrated that environmental elements such as surroundings, colors, and lighting have an effect on the mood and perception [79]. Additionally, these features provide more information about the video clip, which makes a deep neural network to learn more effectively.

We first sample the frames at equal intervals of one second because video clips can have varying frame rates and taking consecutive frames would be inconsistent in terms of the temporal relation of frames. Besides, for a video with a high frame rate and long duration, the total number of frames become quite large so training the neural networks would be unnecessarily slow and memory intensive. With uniform sampling, the learning process is efficient without losing significant information. Another preprocessing operation applied to the frames is resizing. Currently, images with high resolution such as frames of 720p or 1080p videos make training a convolutional neural network infeasible. Because of this reason, all frames are resized to 224×224 pixels. Color information is retained and all of the frames have an RGB color space.

To recognize apparent personality trait factors from preprocessed frames, first, we use a convolutional neural network. Deep Convolutional Neural Networks (DCNN) achieve superior recognition results on a wide range of computer vision problems [80], [81] and they are most suitable for this task as well. We evaluate various deep neural networks and provide comparisons. We use a "warm start" of pretrained ResNet-v2-101 [82], trained on ILSVRC-2012-CLS image classification dataset [83], with fine-tuning to fit the model to the problem. ResNet is a part of a larger model, where it corresponds to the lower layers of the model and more layers are added on top of ResNet.

We apply the CNN to the video per-frame basis so that high-level spatial features are learned. We then exploit the temporal information between video frames. Recurrent Neural Networks (RNNs) allow information from previous events to persist and can connect that information to the present event, however, it has been shown that RNNs are unable to learn dependencies that are long-term [84], [85]. Long Short-Term Memory (LSTM) networks are designed to address this problem and have been demonstrated to be successful [86]. In order to integrate the temporal information, we experiment with RNNs and various types of LSTM networks. We use LSTM units based on Gers et al. style of LSTM networks [87]. In this architecture, the LSTM network corresponds to higher layers of the model. We add this LSTM network on top of ResNet. Figure 3.3 shows the architecture of this subnetwork.

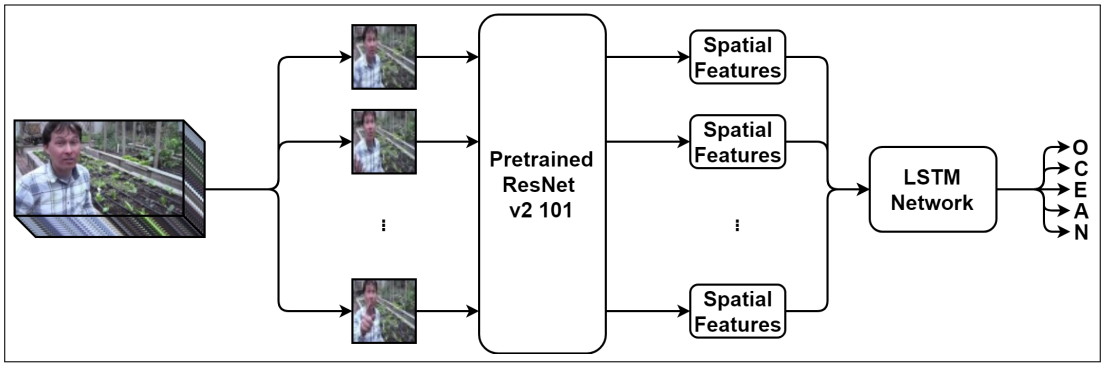

Figure 3.3: The ambient feature-based neural network.

We train this neural network to recognize personality traits based on ambient features only. Afterward, we use all of the trained layers in a larger model where the subnetwork will be a component of the model.

3.2 Facial Feature-based Recognition

One other approach in the proposed method is to recognize the traits based on facial features. It has been shown that personality can be accurately assessed from faces [88] and facial symmetry is associated with five-factor personality factors [89]. Hence, it is crucial to make use of this information in order to assess personality. Although faces are included in the images used by the previously mentioned subnetwork, they become too small after scaling and in order to analyze faces properly, other parts of the images should be removed. Therefore, we use faces as the sole input of the neural network in this approach.

In order to obtain facial features, we first detect faces and align them. One face detector that has shown to work well is Multi-task CNN (MTCNN) [90] and we make use of MTCNN in our method. However, we test other methods such as OpenFace face alignment method [91] and provide comparisons. We apply face alignment to the frames used in the ambient feature-based subnetwork, to ensure that the time step is consistent across different modalities. Likewise, we scale frames after face alignment, resulting in 224×224 pixel face images.

Because both facial feature-based recognition and ambient feature-based recognition are computer vision tasks, the rest of this process is similar. First, there is a convolutional neural network that learns high-level spatial features per-frame basis, then a recurrent neural network, specifically an LSTM network, integrates the temporal information. We use ResNet-v2-101 [82] as the CNN because it is the most effective one according to the experiments.

The LSTM network built on top of the CNN is similar to the one for the ambient feature-based subnetwork. We train this neural network consisting of ResNet and LSTM units to learn features from aligned faces and to assess the five personality factors. Figure 3.4 depicts the architecture of this subnetwork.

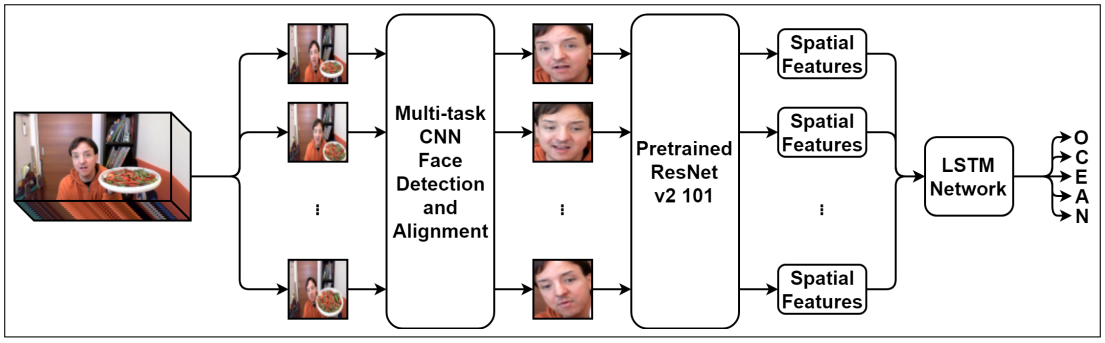

Figure 3.4: The facial feature-based neural network.

3.3 Audio Feature-based Recognition

The third modality used in the proposed model is the audio. We extract input features from the audio waveforms for the model before using a neural network. This process is the same as the preprocessing method used to train a VGG-like audio classification model, called VGGish [92], on a large YouTube dataset that is a preliminary version of YouTube-8M [93]. As a result of this process, we compute a log mel-scale spectrogram and convert these features into a sequence of successive non-overlapping patches of approximately one second for each audio waveform [94].

After we obtain the audio feature patches, we use a convolutional neural network to convert these features into high-level embeddings. The input of CNN is 2D log mel-scale spectrogram patches where the two dimensions represent frequency bands and frames in the input patch. Although we tested different architectures for the neural network, we used the pretrained VGGish model [92] as a "warm start" and fine-tuned that model in our framework. This model outputs 128-dimensional embeddings for each log mel-scale spectrogram patch.

Because there is a patch for each one-second interval and embeddings are obtained from these patches, we make use of temporal correlation before predicting the apparent personality. This part is the same as integrating the time information in the video for ambient feature-based and facial feature-based subnetworks; so we use an LSTM network again. Consequently, this composition of neural networks learns to recognize personality from audio. Figure 3.5 shows the architecture of this subnetwork.

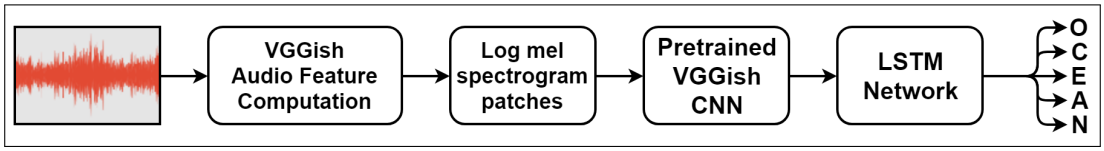

Figure 3.5: The audio feature-based neural network.

3.4 Transcription Feature-based Recognition

The last modality used in the proposed method is the transcription of the speech of people in the videos. Psychological research has shown that personality influences the way a person writes or talks and word use and expressions are associated with personality [95]. For example, individuals that score high in extraversion prefer complex, long writings and conscientious people tend to talk more about achievements and work [96]. These studies indicate that people with similar personality factors are likely to use the same words and choose similar sentiment expressions. Therefore, it is essential that this information is analyzed to make an accurate prediction of personality traits.

In this approach, we apply a language module to the text features in order to compute contextualized word representations and to encode the text into high dimensional vectors before the learning phase. For this purpose, there are several language models that can be applied. One particular approach that is suitable for this subnetwork is a language module that computes contextualized word representations using deep bidirectional LSTM units, which is trained on one billion word benchmark [97], called Embeddings from Language Models (ELMo) [98]. This model outputs 1024-dimensional vector containing a fixed mean-pooling of all contextualized word representations. Although the model has four trainable scalar weights, in this setting, we fix all parameters so there is no additional training for this language module.

After obtaining the embeddings, the next step is to directly learn to recognize personality traits from these features. At this stage, unlike all other subnetworks, there is no LSTM network or any other variation of RNNs because the information related to the sequences of words is already encoded into the embeddings through the bidirectional LSTM units in ELMo. As a result, a few additional layers on top of this language module are added to train a regressor neural network which performs recognition of personality factors from transcription features. Figure 3.6 depicts the architecture of this subnetwork.

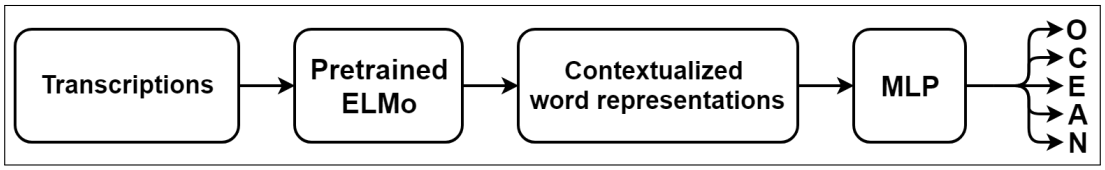

Figure 3.6: The transcription feature-based neural network.

Chapter 4

Experimental Results and Evaluation

In this chapter, we present the experiments that are carried out for the proposed method, the dataset which the model is trained on, and the experimental results. The proposed approach and various other alternatives are experimented with and compared to each other, and the best performing method is compared to the state-of-the-art. The results demonstrate that the proposed method outperforms the current state-of-the-art. In the following sections, the dataset, evaluation method, and the experiments are explained in detail.

4.1 Dataset

The dataset used to evaluate the proposed approach is the ChaLearn First Impressions V2 (CVPR'17) challenge dataset [21]. The aim of this challenge is to automatically recognize apparent personality traits according to the five-factor model. The dataset for this challenge consists of 10000 videos of people facing and speaking to a camera. Videos are extracted from YouTube, they are mostly in high-definition (1280×720 pixels), and in general, they have an average duration

of 15 seconds with 30 frames per second. In the videos, people talk to the camera in a self-presentation context and there is a diversity in terms of age, ethnicity, gender, and nationality. The videos are labeled with personality factors using Amazon Mechanical Turk (AMT), so the ground truth values are obtained by using human judgment. For the challenge, videos are split into training, validation and test sets with a 3:1:1 ratio. The dataset is publicly available [1]. Figure 4.1 shows some examples of videos.

Figure 4.1: Sample videos from the training set depicting various cases of how personality traits are perceived by human judgment.

[1]The dataset is available at: http://chalearnlap.cvc.uab.es/dataset/24/description

In the data collection process, AMT workers compare pairs of videos and evaluate the personality factors of people in the videos by choosing which person is likely to have more of an attribute than the other person for each personality factor [21]. Multiple votes per video, pairwise comparisons, and labelling small batches of videos are used to address the problem of bias for the labels. Final scores are obtained from the pairwise scores by using a Bradley-Terry-Luce (BTL) model [99], while addressing the problem of calibration of workers and worker bias [100].

The evaluation metric is also defined by the challenge. From the trained models, it is expected that the models output continuous values for the target five personality traits in the range of [0, 1]. These values are produced separately for each trait, therefore there are 5 predicted values to be evaluated. For this purpose, the "mean accuracy" over all predicted personality trait values is computed as the evaluation metric [21]. Accordingly, it is defined as:

$$A = 1 - \frac{1}{N} \sum_{i=1}^{N} |t_i - p_i| \tag{4.1}$$

where t_i are the ground truth scores and p_i are the estimated values for traits with the sum running over N videos.

In the dataset, there are in total over 2.5 million frames that can be processed during training, and although encoded data size (size of videos) is about 27 gigabytes, decoded size (size of tensors) is over 10 terabytes. Because of this large-scale data, training CNNs and RNNs, which are computationally intensive, with the full usage of this dataset is infeasible for this task given the hardware used in experiments and time limitations. Therefore sampling and resizing are applied to this dataset prior to training the neural networks.

4.2 First Stage Training

The first stage consists of training each subnetwork independently and does not give a final prediction for personality traits. In this stage, we train the subnetworks as explained in the proposed framework as well as some alternatives to compare the results. We report the validation set performances of the subnetworks with different architectures in Table 4.1. We finetune the hyperparameters for each system. We use Adam optimizer, which is a stochastic gradient descent method for parameter optimization [101]. Table 4.2 provides the performances of the best-performing subnetworks for each personality trait.

Table 4.1: The validation set performances of the subnetworks with different architectures. Best-performing ones are shown in bold.

Subnetwork	Mean Accuracy
Ambient: CNN	0.9012
Ambient: 3D-CNN	0.8962
Ambient: Inception-v2	0.9089
Ambient: ResNet-v2-101	**0.9116**
Face: MTCNN + Inception-v2	0.9067
Face: MTCNN + ResNet-v2-101	**0.9136**
Face: Dlib + Inception-v2	0.9058
Face: Dlib + ResNet-v2-101	0.9107
Audio: VGGish	**0.9049**
Transcription: USE_T	0.8869
Transcription: ELMo	**0.8872**
Transcription: Skip-gram	0.8870

In order to obtain a baseline network, we train an ambient feature-based subnetwork initially. For this purpose, we implement and train a network consisting of a simple convolutional neural network and an LSTM network, without using a pre-trained network in the architecture. The reason of this is to demonstrate that using such type of a network is suitable for our problem even though it

Table 4.2: The performances of the subnetworks for individual personality traits.

Subnetwork	Open.	Cons.	Extr.	Agre.	Neur.	Mean
Ambient	0.9101	0.9151	0.9120	0.9136	0.9073	0.9116
Facial	0.9103	0.9160	0.9165	0.9148	0.9104	0.9136
Audio	0.9061	0.9012	0.9025	0.9096	0.9050	0.9049
Transcription	0.8877	0.8789	0.8845	0.8993	0.8858	0.8872

may not produce the best results necessarily. In this network, the architecture of the CNN consists of 12 convolution layers with rectified linear unit (ReLU) activations, max pooling layers, batch normalization layers [102], and one dropout layer [103]. The dimensionality of the output space (number of filters) and the height and width of the 2D convolution windows vary between layers, which can be seen in Figure 4.2. The stride value of the convolution along the height and width is 1 for all convolutions. Padding is applied to input so that the input image gets fully covered by the filter. Because the stride value is 1, the output image size is the same as input after the convolutional layers. Downsampling happens at max pooling layers. Size of the pooling window is 2 and stride of the pooling operation is 1 for all spatial dimensions in every max pooling layer.

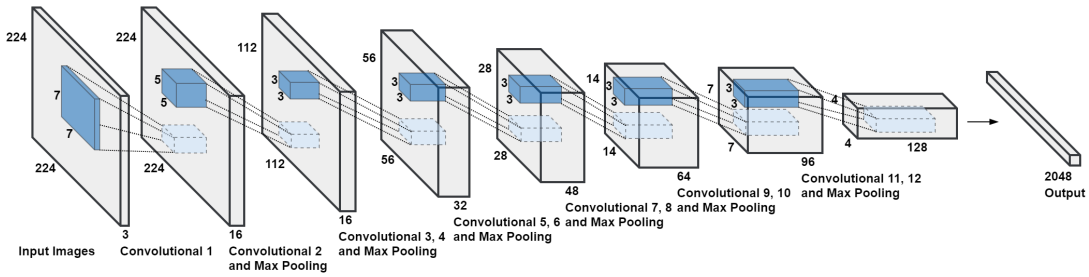

Figure 4.2: Architecture of CNN.

We do not use strided convolutions or average pooling as downsampling methods based on empirical results. In order to initialize this convolutional neural network, "MSRA" initialization [104] is used because it has been shown that it is useful to keep the scale of the input variance constant, so that it does not diminish or explode in the subsequent layers. As the result of the final layer, this CNN maps the input 224×224×3 px video frames into 2048 dimensional high-level features.

The LSTM network consists of 3 layers, where LSTM units are based on [105], and the LSTM architecture contains "peephole connections" from internal cells to the gates to learn precise timing of the outputs [106]. In these LSTM cells, the cell state is clipped by a fixed value before the activation of cell output. Additionally, attention is added to the cells, resulting in a long short-term memory cell with attention (LSTMA) based on the implementation given in [107]. The number of units in the LSTM cells varies across layers. Similar to the CNN, dropout is added to the outputs of the cells, however, input or state dropouts are not applied.

For the training of the neural network, the loss is defined as:

$$A = \frac{1}{N} \sum_{i=1}^{N} |t_i - p_i| \tag{4.2}$$

which is based on "mean accuracy" and similar to the metric, t_i denotes ground truth scores and p_i denotes estimated values for traits with the sum running over N videos. In order to compute the loss, the predicted scores are computed using a sigmoid activation after the last layer of LSTM network. The optimizer for this loss is the implementation of the Adam algorithm [101], which uses the formulation where the order of computation is changed as explained before Section 2.1 of the Kingma and Ba paper instead of the formulation in Algorithm 1. During the training, data augmentation is also performed to the inputs, including randomly flipping the contents of images horizontally and adjusting the brightness, saturation, hue, and contrast of RGB images by random factors. We apply feature scaling to the input images as well because gradient descent converges much faster with feature scaling than without it [108]. Because this is a starting point, we compare the results against a "dummy regressor" that always predicts the mean of the training set as a simple baseline. As expected, this type of neural network works well in this problem, obtaining a mean accuracy score of 0.9012 on validation data (see Figure 4.3).

The next step is to compare this CNN and LSTM network against other

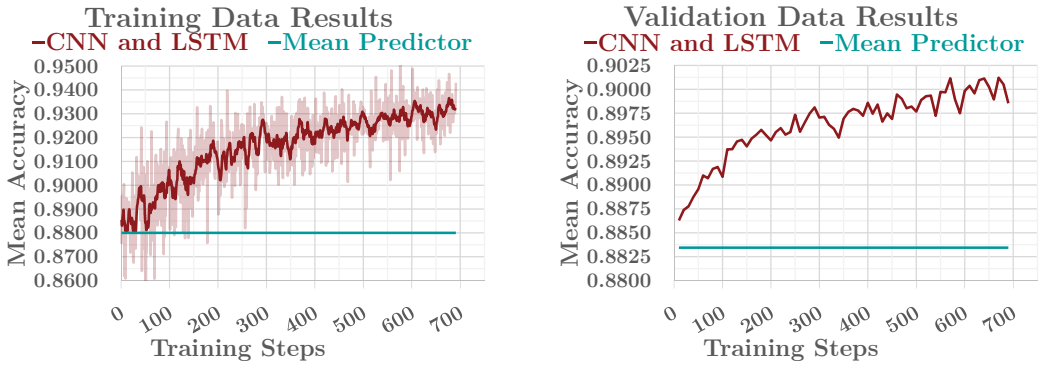

Figure 4.3: The results of simple CNN and LSTM network.

possible network types. One potentially suitable approach is using a 3D convolutional neural network instead of LSTM network. 3D convolutions apply a three-dimensional filter to the dataset where the filter moves in all dimensions to calculate the features. In this way, temporal information can be captured through convolutions and there is no need for the LSTM network. It has been demonstrated that 3D convolutional networks are effective for spatio-temporal feature learning on a large scale supervised video dataset [109]. Hence, we trained a 3D-CNN that is similar in terms of the network architecture on this dataset to compare the performances.

The results show that our approach using LSTM network outperforms the 3D-CNN approach, which obtains a score of 0.8962 (see Figure 4.4). One observation is that 3D-CNN fits to training data better, which indicates that 3D convolutions

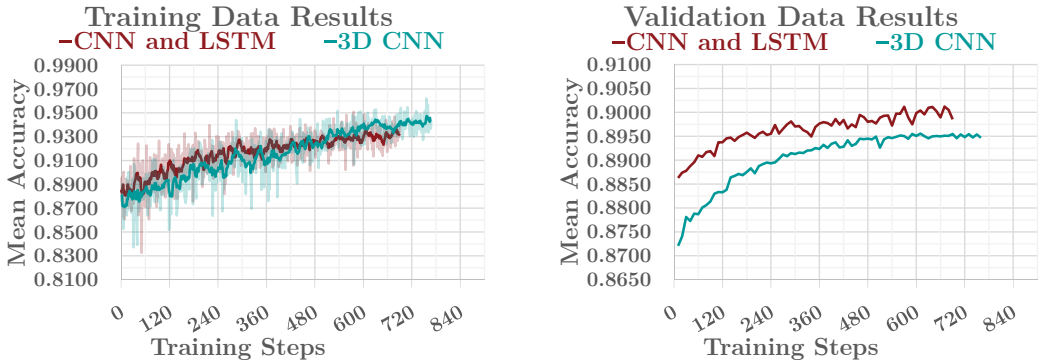

Figure 4.4: The results for 3D-CNN and CNN-LSTM networks.

have a problem of overfitting for this dataset.

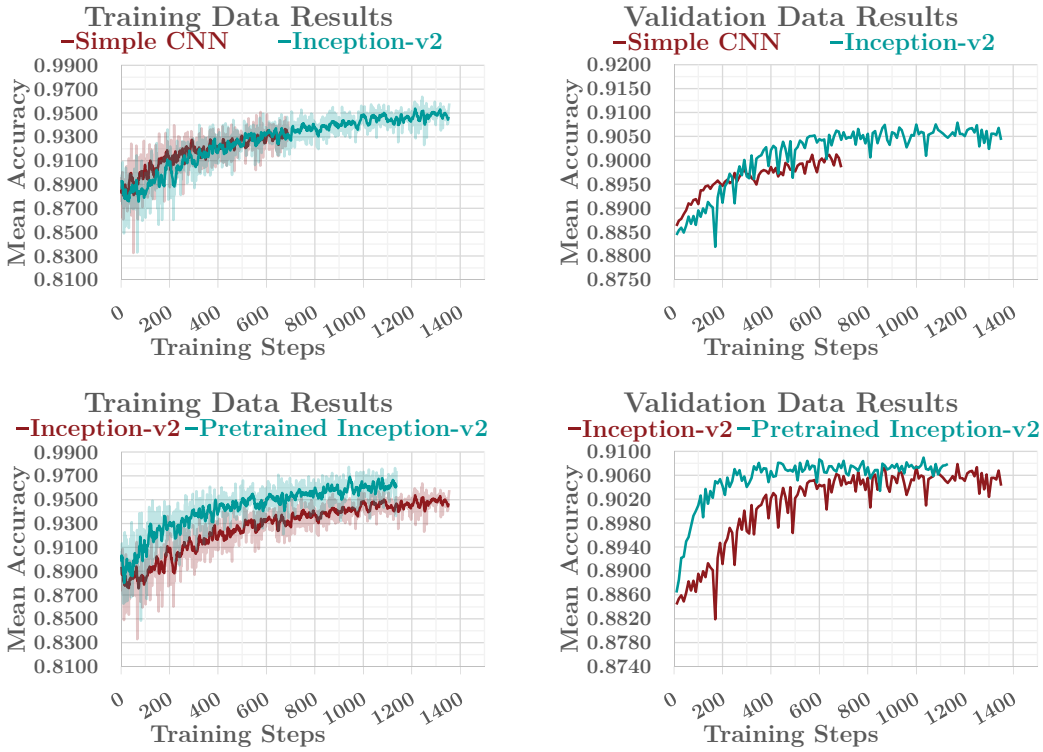

Figure 4.5: The results of training Inception-v2 from scratch compared to simple CNN (top), and fine-tuning the pretrained Inception-v2 model compared to the previous version (bottom).

After establishing that the proposed approach is effective, we replaced the simple convolutional neural network architecture with a complex network that is known to perform well in various domains. For this purpose, we have initially trained the Inception-v2 model [102] without using the pre-trained network and following this, we have also trained this model by fine-tuning the pre-trained network which is trained on ILSVRC-2012-CLS image classification dataset [83]. Experiments showed that in both cases the results were better than the simple CNN, and as expectedly, fine-tuning the pre-trained Inception-v2 network resulted in improved performance. Additionally, we observed that pre-trained Inception-v2 fits both training and validation data quicker and better. Accordingly, the best score obtained is 0.9089, as seen in Figure 4.5.

We used pre-trained networks in all of the following experiments. Although

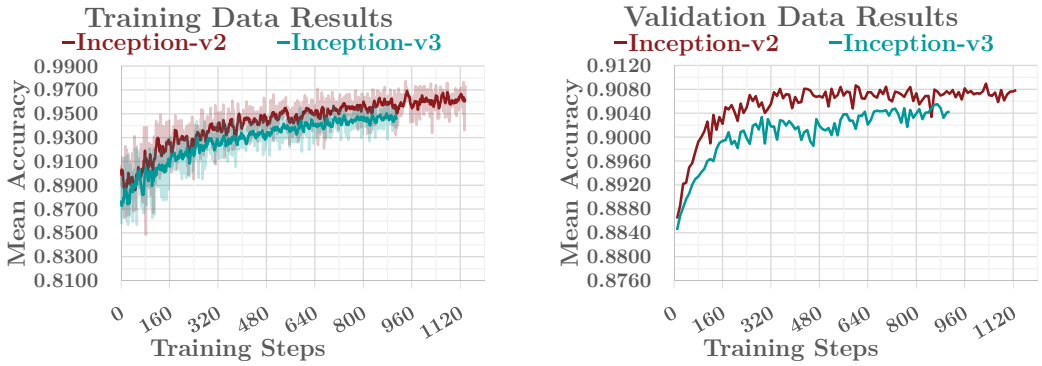

Figure 4.6: The results for Inception-v2 vs. Inception-v3 networks.

Inception-v2 resulted in better results, there are different versions of this model. Therefore, we trained Inception-v3 [110], Inception-v4 [111], and Inception-ResNet-v2 [111] versions as well. However, despite being later generations and outperforming previous Inception networks on the test set of the ImageNet classification (CLS) challenge [110, 111, 83], in our experiments all of these networks failed to outperform Inception-v2 version for the personality recognition task (see Figures 4.6, 4.7, and 4.8)

After getting the results of various Inception networks, experiments were carried out to compare Inception-v2 to other network architectures. Starting with the ResNet models [112, 82], we have fine-tuned the ResNet-v2-101 [82] network on our dataset which has been previously trained on ImageNet Large Scale Visual

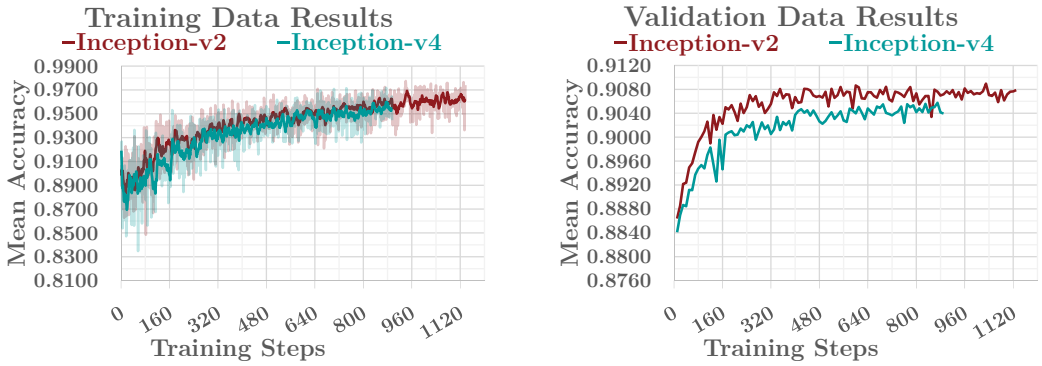

Figure 4.7: The results for Inception-v2 vs. Inception-v4 networks.

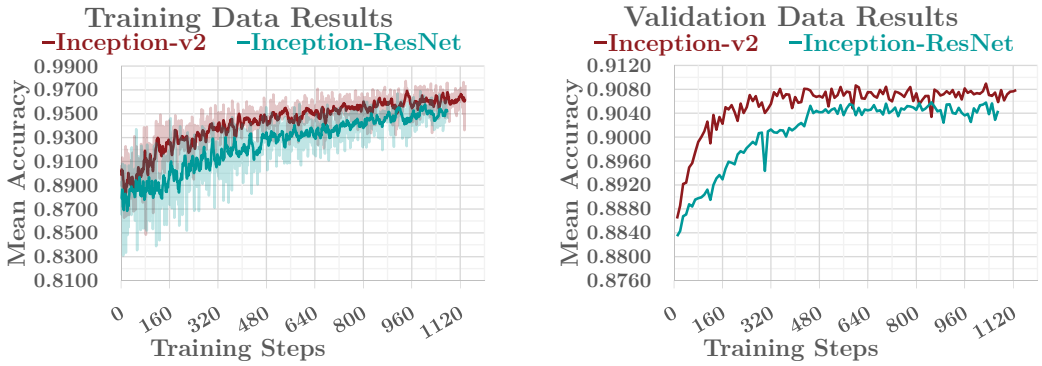

Figure 4.8: The results for Inception-v2 vs. Inception-ResNet-v2 networks.

Recognition Challenge dataset [83]. We have found out that the ResNet architecture outperforms Inception networks significantly in both training and validation sets, obtaining a "mean accuracy" score of 0.9100 (see Figure 4.9).

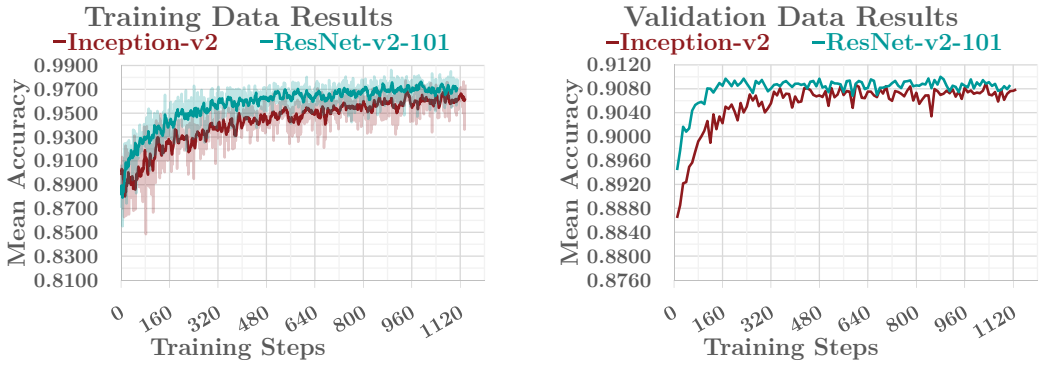

Figure 4.9: The results for Inception-v2 and ResNet-v2-101 network architectures.

Similar to the Inception networks, ResNet architecture also has various versions. Therefore, instead of the 101-layer full preactivation "v2" variant ResNet, it is possible to use the 50-layer or 152-layer versions or the "v1" variant [112, 82]. For this reason, we have trained the 50-layer and 152-layer ResNet-v2 networks as well as the 101-layer ResNet-v1 network to compare their results to the 101-layer ResNet-v2 network. According to our experiments, all versions of the ResNet architecture have shown similar performances in terms of "mean accuracy" for the recognition of personality traits, which can be seen in Figures 4.10, 4.11, and Figure 4.12. As a result, we made no changes to the architecture and continued using the 101-layer ResNet-v2 network in the following experiments.

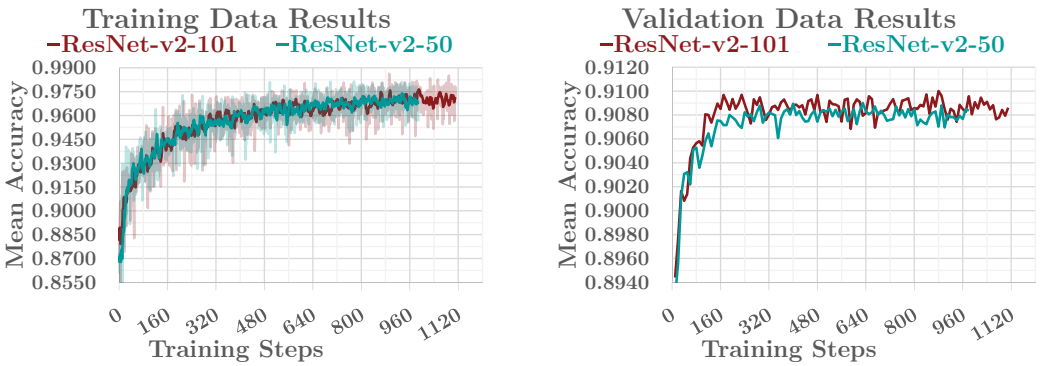

Figure 4.10: Comparison of ResNet-v2-101 and ResNet-v2-50 networks.

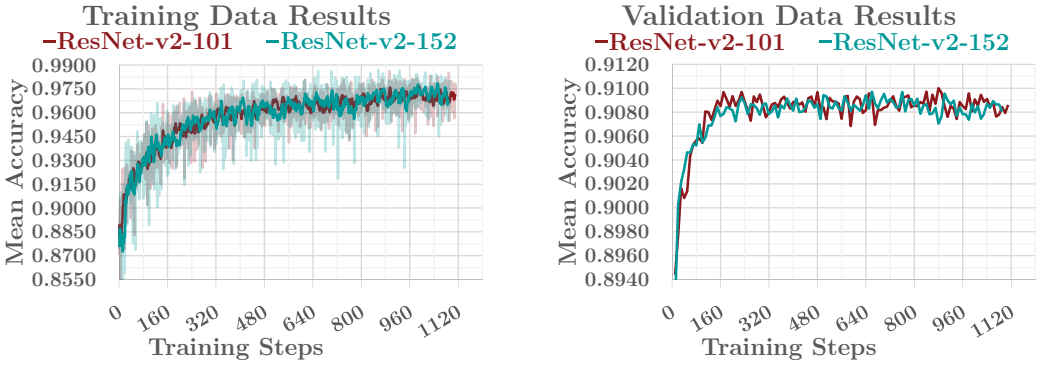

Figure 4.11: Comparison of ResNet-v2-101 and ResNet-v2-152 networks.

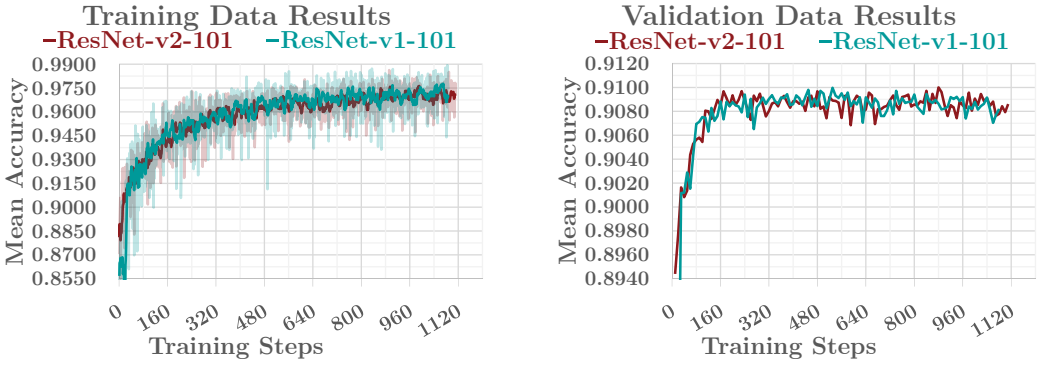

Figure 4.12: Comparison of ResNet-v2-101 and ResNet-v1-101 networks.

In order to finalize the ambient feature-based personality traits recognition convolutional neural network architecture, we have tried two more network types, MobileNet networks [113, 114] and NASNet networks [115]. For these architectures, we used MobileNetV2 (1.4) version and NASNet-A model. However, neither of these networks were able to outperform the ResNet architecture. The results can be seen in Figures 4.13 and 4.14 for MobileNet and NASNet comparisons, respectively.

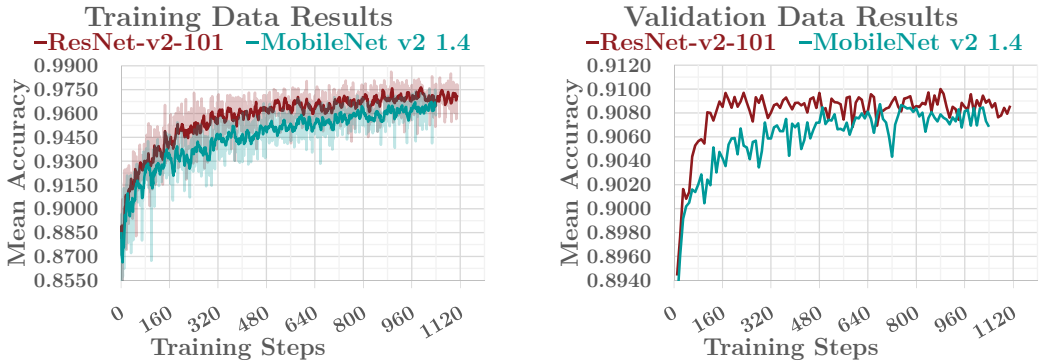

Figure 4.13: The results for ResNet-v2-101 and MobileNetV2 (1.4) networks.

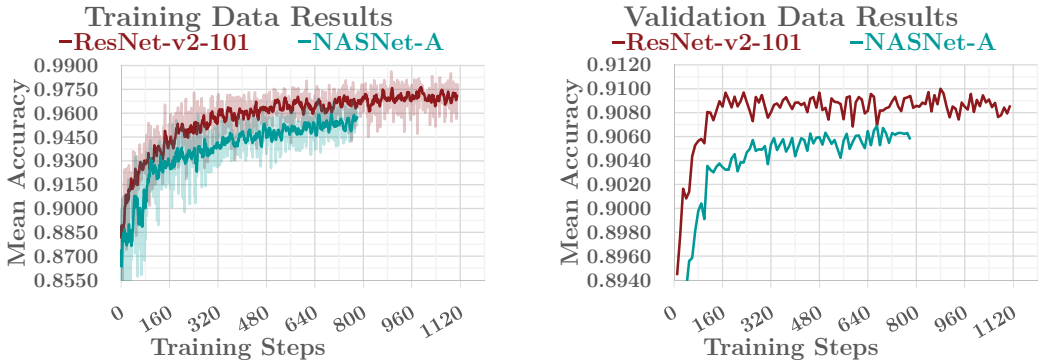

Figure 4.14: The results for ResNet-v2-101 and NASNet-A networks.

The last step after finalizing the CNN architecture is to optimize the hyperparameters and the LSTM network. For this purpose, we have conducted several experiments to obtain the network with the best performance results. Consequently, we have found out that removing the attention from LSTM cells, using 2 layers with 1000 and 5 units in the cells, applying the implementation based on [87] instead of augmenting the network by "peephole connections" [106], and

not clipping the cell states have resulted in improved performance. We have also dropped the sigmoid function at the end of the network in order to obtain the scores directly from the LSTM network (which uses the hyperbolic tangent "tanh" function). For the training, we have used learning rate of 10^{-5}, and a batch size of 8 videos with 6 frames for each video resulting in 48 images for the convolutional neural network and 8 high-dimensional features from CNN for the LSTM networks. The dropout probability is 0.5. For the data augmentation, we have removed randomly flipping the contents of images horizontally but kept other augmentation methods. In this configuration, this CNN and LSTM network for ambient feature-based features obtained a "mean accuracy" score of 0.9116 (see Figure 4.15).

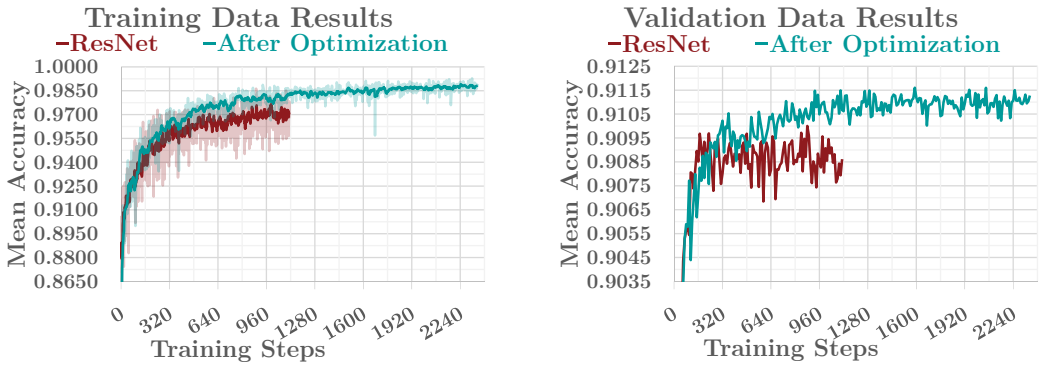

Figure 4.15: The results of ResNet-v2-101 network after hyperparameter and LSTM network optimization.

Similar to the ambient feature-based personality recognition subnetwork, a convolutional neural network with an LSTM network was trained for facial features as well. In order to get the input images for this network, we have performed face recognition and alignment. For this task, we have applied two different algorithms, Dlib face detector [116] and Multi-task CNN face detection and alignment [90]. Based on the qualitative and quantitative evaluation, Multi-task CNN has demonstrated to work better in our setting because Dlib face detector fails to detect the face in some videos due to partial occlusion, lighting or some other factors, whereas Multi-task CNN does not miss any video and produces visually better-aligned faces. One example of a video where Dlib does not detect the face is given in Figure 4.16.

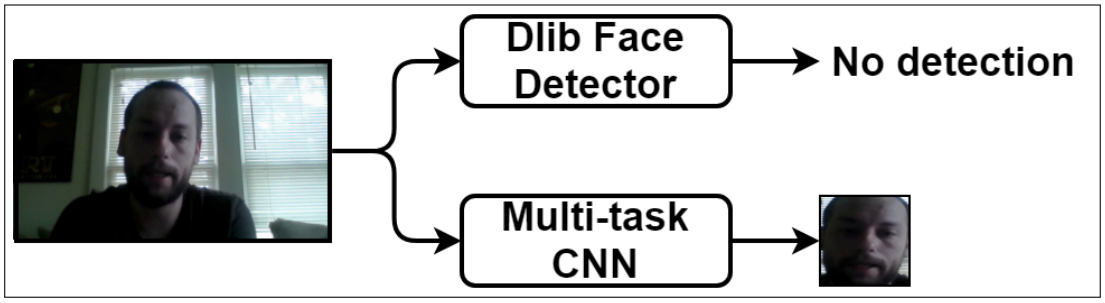

Figure 4.16: Dlib and Multi-task CNN face alignment methods on an example video.

After obtaining the face aligned frames for videos, we use the same CNN and LSTM network approach to recognize the personality traits. Based on the previous experiments, we do not train networks with comparatively worse performances such as 3D-CNN, MobileNet, and NASNet networks and also networks that are trained from scratch and not from a pre-trained version. Initially, we trained the same network that gave the best results for ambient feature-based features, which is ResNet-v2-101 for facial features and compared the performance to the ambient feature-based subnetwork. The results show that, although both neural networks fit training data similarly, on the validation data, facial feature-based subnetwork outperforms significantly by obtaining a "mean accuracy" score of 0.9136. The results are given in Figure 4.17. This demonstrates that, for personality traits, focusing on facial features is more relevant and informative than looking at the full frame including surroundings. However, in our approach, using ambient feature-based subnetwork is still beneficial for the final prediction of personality traits, which is explained in the following section.

We have then trained other network architectures for face aligned input images to see if it is possible to obtain better results with these networks given that the input is different. For this purpose, we only considered networks that were similar in terms of performance based on previous experiments and excluded other architectures. However, none of these networks were able to outperform ResNet-v2-101 model. Additionally, we have observed considerably worse results for some networks such as Inception-ResNet-v2 and ResNet-v1 so the training phases of these networks were terminated early. Figure 4.18 shows the results of

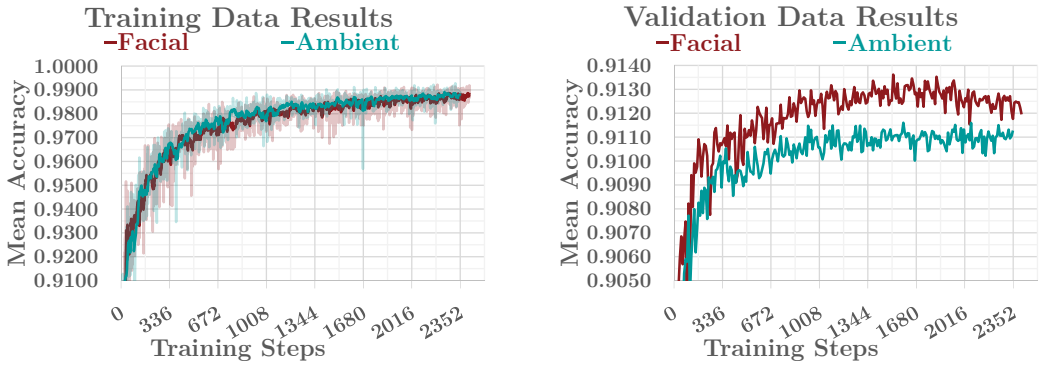

Figure 4.17: Comparison of facial feature-based subnetwork and ambient feature-based subnetwork.

Inception-v2 as an example. Finally, we conducted experiments to optimize the hyperparameters and the LSTM network similar to the ambient feature-based subnetwork, however, no changes were made to this neural network.

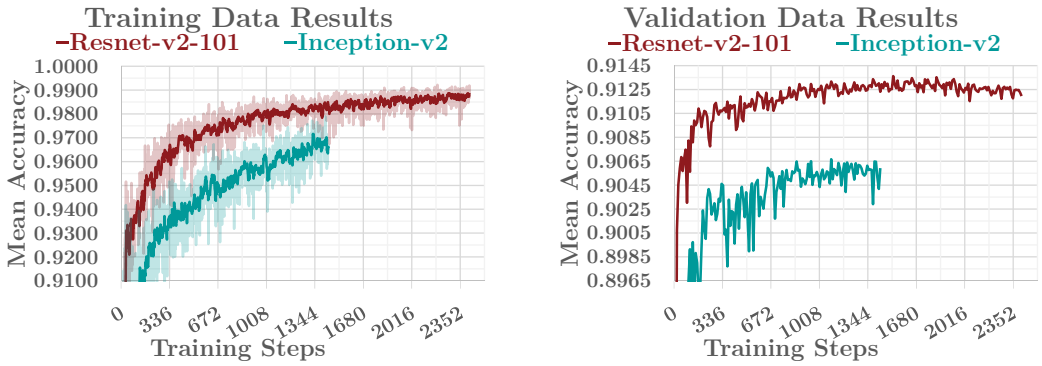

Figure 4.18: The results for ResNet-v2-101 and Inception-v2 networks using face aligned images.

Another neural network was trained for the third modality, which is the recognition of personality traits based on audio features. For this neural network, the architecture is based on VGGish [92], which is trained on a large YouTube dataset that is a preliminary version of YouTube-8M [93]. We also apply the same preprocessing methods to compute audio features that are used to train VGGish network [92]. In this procedure, all audio input is resampled to 16 kHz monaural sound and using magnitudes of the Short-Time Fourier Transform (STFT) [117, 118, 119] and a periodic Hann window [120] a spectrogram

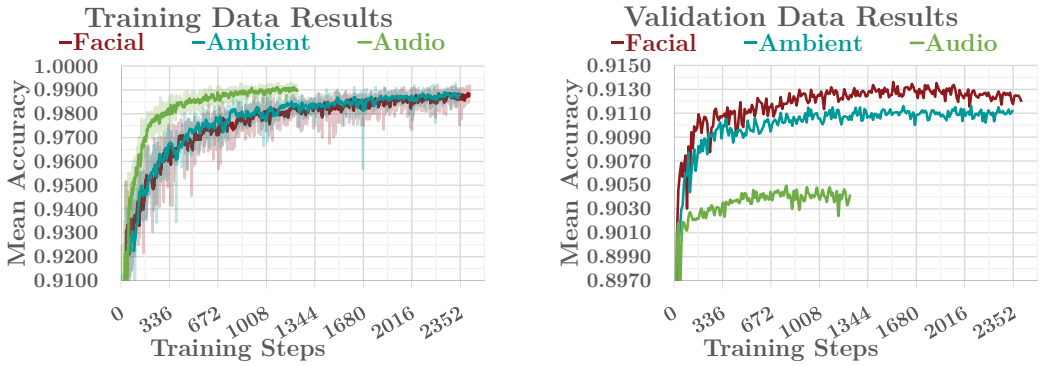

Figure 4.19: The comparison of audio feature-based, facial feature-based, and ambient feature-based subnetworks.

is computed. Then, log mel spectrogram features are computed by using mel bins and applying logarithmic function. Then, these features are converted into a sequence of successive non-overlapping frames where each example covers 64 mel bands and 96 frames of 10 ms each. Temporal correlation is integrated using an LSTM network and the LSTM network is same as ambient feature-based and facial feature-based LSTM networks. The only hyperparameter that differs from previous subnetworks is the learning rate, which is 10^{-4} for this neural network.

Based on the experiments, we found out that other network architectures are not suitable. As a result, we have obtained a score of 0.9049 for trait recognition using audio features (see Figure 4.19). The results show that audio feature-based personality recognition is not as effective as visual feature-based recognition. One important problem is that this network is unable to generalize well and in fact, it overfits to training data, by obtaining lower error rates during training compared to visual feature-based networks. Although one possible option is to modify the network to a complexity just large enough to provide an adequate fit for both training and validation data, we do not change the network structure in order to preserve the architecture of pre-trained VGGish model and instead of this, we reduce the number of units in the LSTM cells during the second stage of training. In this way, we utilize audio feature-based subnetwork to improve the overall performance in the second stage without causing overfitting to training data.

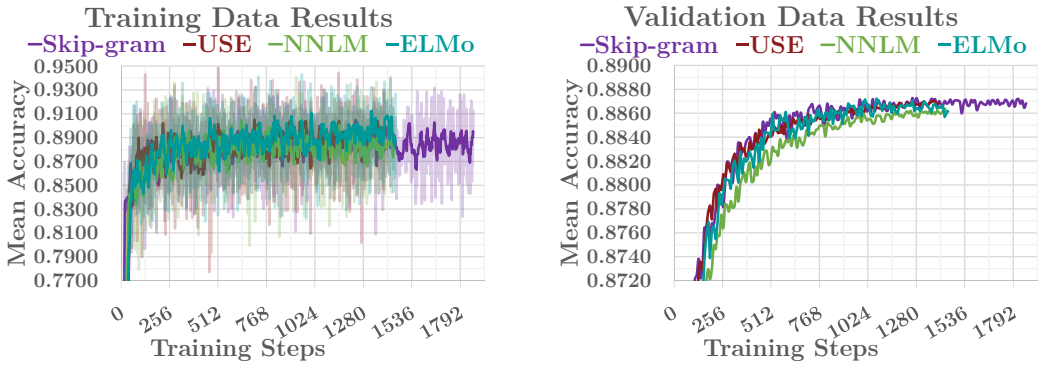

Figure 4.20: The results of various models used for transcription input.

As the final step of the first stage, we train a neural network for the last modality which is based on transcription features. To this end, we apply a language model to encode the transcription text into high dimensional vectors that can be used for personality recognition. For this purpose, we have experimented with various models. Universal Sentence Encoder is a model trained for natural language tasks such as semantic similarity, text classification, and clustering on a variety of data sources [121]. It also differs from word level embedding models since the input can be variable length English text including phrases and sentences. Therefore, it is suitable to use this model in our approach. Similarly, ELMo is a model trained on one billion word benchmark to compute contextualized word representations using deep bidirectional LSTMs [98], so we have trained another model using ELMo. In addition to these, other possible models are the feed-forward Neural-Net Language Model based text embedding [122] trained on English Google News 200B corpus, and text embedding based on the skip-gram version of word2vec [123, 124] trained on English Wikipedia corpus. Accordingly, we have experimented with all these models and compared the performances. To obtain the prediction for personality traits, we have added several fully-connected layers on top of the language models that take the embeddings as input and output the values for personality traits. According to the results, we have observed that Neural-Net Language Model had slightly worse performance while all other models had very similar results (see Figure 4.20). Consequently, we used ELMo embeddings in the final model, which obtains a score of 0.8872.

We finalize the first stage of training the model by obtaining the separately trained subnetwork for all modalities. Comparison of the subnetworks is given in Figure 4.21. It can be seen that visual feature-based recognition gives the best results while audio feature-based network does not generalize well and the transcription feature-based network is unable to fit as effective as the others.

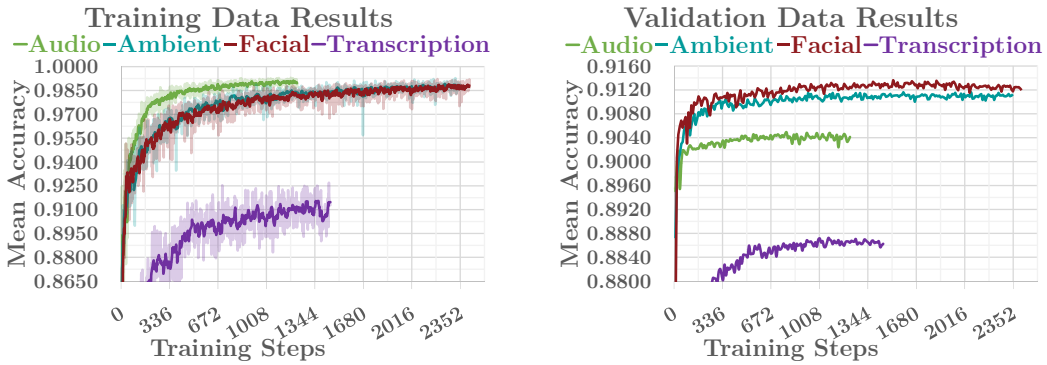

Figure 4.21: The comparison of facial feature-based, ambient feature-based, audio feature-based, and transcription feature-based subnetworks.

4.3 Second Stage Training

The second stage of training the model consists of combining the separately trained modality-specific neural networks to obtain a final prediction of personality traits. For this purpose, we modify subnetworks in order to apply early feature-level fusion and train a larger model. We keep higher level features such as the outputs of ResNet-v2-101 and VGGish CNNs and the ELMo embeddings fixed while changing the outputs of each subnetwork by applying modifications to LSTM networks and dropping the last layer of the transcription feature-based subnetwork. The reasoning is that instead of getting five-dimensional outputs as the predicted personality traits from each subnetwork, we obtain higher-dimensional features so that the larger model can learn correlations between various modalities. Additionally, audio-based subnetwork has the tendency to overfit to training data (see Figure 4.21). We reduce the capability of this subnetwork by changing network structures and increase the complexity of better performing subnetworks such as

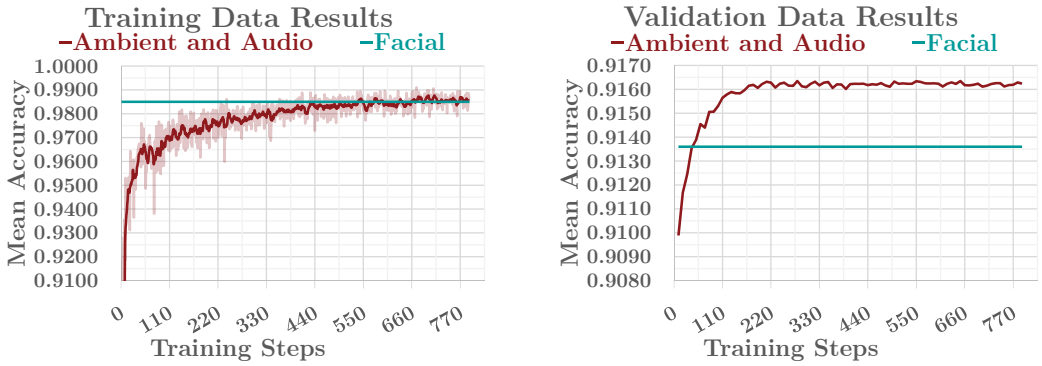

Figure 4.22: The results of the simple multimodal network consisting of ambient feature-based and audio feature-based subnetworks with early fusion.

visual feature-based networks. For the evaluation, we combine modalities one by one in order to observe the effect of each subnetwork to the final model. To begin with, we use the ambient feature-based subnetwork and the audio feature-based subnetwork in order to utilize both visual and audio input. These subnetworks perform worse than facial feature-based subnetwork separately, so we compare the combined network to the facial feature-based subnetwork.

According to the results, the score of the combined network is 0.9163, outperforming the score of the facial feature-based subnetwork, which is 0.9136 (see Figure 4.22). Therefore, it can be observed that the multimodal network, even though it consists of relatively underperformer networks, is able to give better results when compared to a model with only one modality.

Next, we modify the model by including the facial features in order to utilize three features. The score for the three-feature network is 0.9185, which is better than the two-feature version as expected because facial feature-based subnetwork is best performing one (see Figure 4.23).

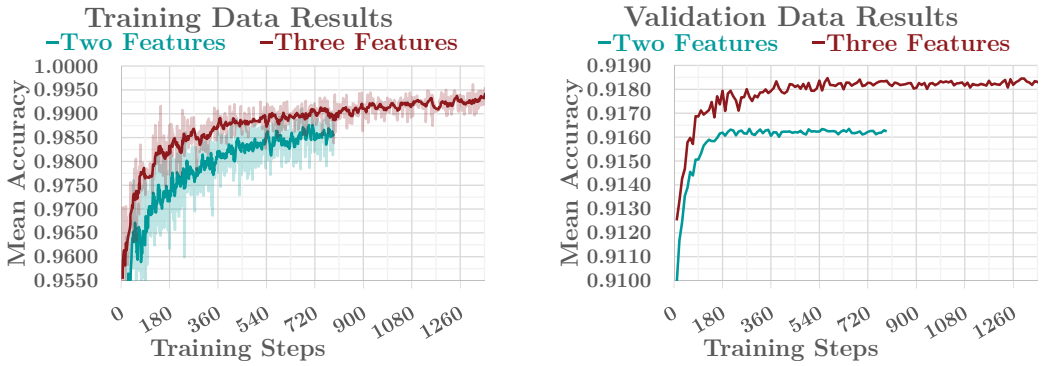

Figure 4.23: The comparison of the two and three feature-based networks. The two feature-based network uses ambient and audio features and the three-feature-based network uses ambient, audio, and facial features.

Finally, we use all features by adding the transcription input. For this four-feature model, we have experimented with Gated Recurrent Unit (GRU) cells [125] instead of LSTMs. They gave similar but not better results, so we decided to keep using LSTM network. However, we made changes to the network by adding the inputs of LSTM cells to the outputs, creating a residual network. In the finalized model, ambient feature-based, facial feature-based and audio feature-based subnetworks have six LSTM layers, whereas transcription feature-based subnetwork has three fully connected layers. For the feature-level fusion, there are 80-dimensional features from ambient feature-based and facial feature-based subnetworks and 20-dimensional features from audio feature-based and transcription feature-based subnetworks. The final score obtained from this model is 0.9188, which is the best score that we have obtained (see Figure 4.24). As a result, our experiments demonstrate that all modality-specific subnetworks contribute to the larger model by improving the performance.

We compare the performance of the finalized model to the state-of-the-art methods (see Table 4.3). We observe that the mean accuracy scores of the methods are between 0.91 and 0.92. In comparison to others, our approach obtains the best performance in terms of the mean accuracy metric. For the individual personality traits, our method performs better than the other methods in three traits and three-feature late fusion method obtains the best performance for one trait (agreeableness). From the results, it can be seen that our approaches give

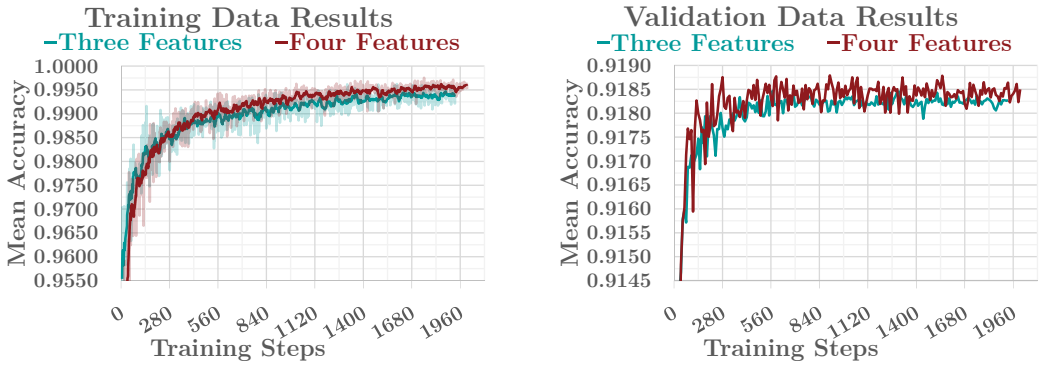

Figure 4.24: The comparison of the three and four feature-based networks. The four-feature network uses ambient, audio, facial, and transcription features.

better performance specifically for the conscientiousness and agreeableness traits. Apart from the validation set performances given in Table 4.3, NJU-LAMDA [126] yields a mean accuracy of 0.9130 and the scores of 0.9123 for openness, 0.9166 for conscientiousness, 0.9133 for extraversion, 0.9126 for agreeableness, and 0.9100 for neuroticism on the test set.

Table 4.3: The comparison of the validation set performances of various approaches.

Method	Mean	Open.	Cons.	Extr.	Agre.	Neur.
DCC [77]	0.9122	0.9117	0.9133	0.9110	0.9158	0.9091
evolgen [76]	0.9134	0.9130	0.9136	0.9145	0.9157	0.9098
Gürpınar et al. [127]	0.9147	0.9141	0.9141	0.9186	0.9143	0.9123
PML [128]	0.9155	0.9138	0.9166	0.9175	0.9166	0.9130
BU-NKU [129]	0.9170	**0.9169**	0.9166	0.9206	0.9161	0.9149
Proposed method	**0.9188**	0.9166	**0.9214**	**0.9208**	**0.9189**	**0.9162**

In addition to the proposed method, we calculate the final scores for each trait by taking the mean of the predicted scores of each subnetwork. With this approach, the best score we obtain is that of the late fusion of three modalities, which are ambient, facial, and audio. The scores obtained with this approach are 0.9162 for openness, 0.9201 for conscientiousness, 0.9195 for extraversion, 0.9196 for agreeableness, and 0.9156 for neuroticism, with a mean accuracy of 0.9182.

Chapter 5

Conclusions

We propose a novel approach for the recognition of personality traits from videos. In our method, we use a multimodal neural network that consists of modality-specific CNNs to extract spatial features such as ambient features and facial expressions, and LSTMs to integrate the temporal information of the videos. The modalities for the neural network include face, environment, audio, and transcription features. We train the network with a two-stage training method where the modality-specific neural networks are trained to predict apparent traits independently in the first stage, and the model is fine-tuned to recognize traits accurately with a feature-level fusion of modality-specific networks in the second stage.

First, we demonstrate that each modality-specific network is effective for the personality recognition and contributes to the final model. To this end, we trained these networks separately and as a result, facial feature-based network gave the best results, followed by ambient feature-based and audio feature-based networks, and lastly, transcription feature-based network. Then, we combined the ambient feature-based and audio feature-based networks to create a simple multimodal network and compared this network to the facial feature-based network. The results showed that the multimodal approach outperforms the single modality approach. Finally, we used all modalities to obtain the finalized neural network

while the improvements of each subnetwork were verified by the experiments. For the final network, we obtained best results by fine-tuning pretrained ResNet-v2-101 network with six LSTM layers for ambient feature-based and facial feature-based subnetworks, using an architecture including a pretrained VGGish network with six LSTM layers for the audio feature-based subnetwork, and using ELMo embeddings with three fully connected layers for the transcription feature-based network.

Quantitative results on personality recognition using ChaLearn First Impressions V2 dataset demonstrate that our method outperforms the state-of-the-art methods. Our model advances the state-of-the art mean accuracy score to 0.9188 (previously 0.9170), while also giving best results for four individual personality traits out of five traits.

As future research directions, we envision that correlation between personality, body movements, posture, eye-gaze and emotion can be investigated to improve the performance given by the usage of ambient, face, audio, and transcription features.

Bibliography

[1] H.-R. Pfister and G. Böhm, "The multiplicity of emotions: A framework of emotional functions in decision making," *Judgment and Decision Making*, vol. 3, no. 1, p. 5, 2008.

[2] D. Kahneman and P. Egan, *Thinking, Fast and Slow*, vol. 1. Farrar, Straus and Giroux New York, 2011.

[3] M. Deniz, "An investigation of decision making styles and the five-factor personality traits with respect to attachment styles," *Educational Sciences: Theory and Practice*, vol. 11, no. 1, pp. 105–113, 2011.

[4] T. Chamorro-Premuzic and A. Furnham, "Personality and music: Can traits explain how people use music in everyday life?," *British Journal of Psychology*, vol. 98, no. 2, pp. 175–185, 2007.

[5] P. J. Rentfrow and S. D. Gosling, "The do re mi's of everyday life: the structure and personality correlates of music preferences.," *Journal of Personality and Social Psychology*, vol. 84, no. 6, p. 1236, 2003.

[6] I. Cantador, I. Fernández-Tobías, and A. Bellogín, "Relating personality types with user preferences in multiple entertainment domains," in *CEUR Workshop Proceedings*, Shlomo Berkovsky, 2013.

[7] R. W. Picard, *Affective Computing*. MIT press, 2000.

[8] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[9] D. Cervone and L. Pervin, *Personality: Theory and Research, 12th Edition*. Wiley Global Education, 2013.

[10] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of Personality*, vol. 60, no. 2, pp. 175–215, 1992.

[11] R. R. McCrae and P. T. Costa, "Validation of the five-factor model of personality across instruments and observers," *Journal of Personality and Social Psychology*, vol. 52, pp. 81–90, 1 1987.

[12] T. E. C. and C. R. E., "Recurrent personality factors based on trait ratings," *Journal of Personality*, vol. 60, no. 2, pp. 225–251, 1992.

[13] L. Goldberg, "The structure of phenotypic personality traits," *American Psychologiest*, no. 48, pp. 26–34, 1993.

[14] L. R. Goldberg, "Language and individual differences: The search for universals in personality lexicons," *Review of Personality and Social Psychology*, vol. 2, no. 1, pp. 141–165, 1981.

[15] S. Ahrndt, J. Fähndrich, M. Lützenberger, and S. Albayrak, "Modelling of personality in agents: From psychology to logical formalisation and implementation," in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, (Richland, SC), pp. 1691–1692, International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[16] F. Durupinar, J. Allbeck, N. Pelechano, and N. Badler, "Creating crowd variation with the ocean personality model," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '08, (Richland, SC), pp. 1217–1220, International Foundation for Autonomous Agents and Multiagent Systems, 2008.

[17] F. Durupınar, N. Pelechano, J. M. Allbeck, U. Güdükbay, and N. I. Badler, "How the Ocean personality model affects the perception of crowds," *IEEE Computer Graphics and Applications*, vol. 31, no. 3, pp. 22–31, 2011.

[18] S. Ahrndt, A. Aria, J. Fähndrich, and S. Albayrak, "Ants in the OCEAN: Modulating agents with personality for planning with humans," in *Proceedings of the European Conference on Multi-Agent Systems*, EUMAS '14, pp. 3–18, Springer, 2014.

[19] P. T. Costa and R. R. McCrae, "Normal personality assessment in clinical practice: The neo personality inventory.," *Psychological Assessment*, vol. 4, no. 1, p. 5, 1992.

[20] D. R. Lynam and T. A. Widiger, "Using the five-factor model to represent the DSM-IV personality disorders: An expert consensus approach.," *Journal of Abnormal Psychology*, vol. 110, no. 3, p. 401, 2001.

[21] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "ChaLearn Lap 2016: First round challenge on first impressions-dataset and results," in *Proceedings of the European Conference on Computer Vision*, ECCV '16, pp. 400–418, Springer, 2016.

[22] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction.," *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, p. 171, 2001.

[23] A. Tlili, F. Essalmi, M. Jemni, N.-S. Chen, *et al.*, "Role of personality in computer based learning," *Computers in Human Behavior*, vol. 64, pp. 805–813, 2016.

[24] M. Tkalčič, B. De Carolis, M. de Gemmis, A. Odić, and A. Košir, "Introduction to emotions and personality in personalized systems," in *Emotions and Personality in Personalized Services*, pp. 3–11, Springer, 2016.

[25] L. Shen, M. Wang, and R. Shen, "Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment," *Journal of Educational Technology & Society*, vol. 12, no. 2, pp. 176–189, 2009.

[26] G. Ball and J. Breese, "Emotion and personality in a conversational agent," *Embodied Conversational Agents*, pp. 189–219, 2000.

[27] J. A. Recio-Garcia, G. Jimenez-Diaz, A. A. Sanchez-Ruiz, and B. Diaz-Agudo, "Personality aware recommendations to groups," in *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pp. 325–328, ACM, 2009.

[28] R. Hu and P. Pu, "A study on user perception of personality-based recommender systems," in *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*, UMAP 2010, pp. 291–302, Springer, 2010.

[29] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah, and J. M. Jose, "Integrating facial expressions into user profiling for the improvement of a multimodal recommender system," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, ICME '09, pp. 1440–1443, IEEE, 2009.

[30] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, pp. 147–160, April 2018.

[31] F. Valente, S. Kim, and P. Motlicek, "Annotation and recognition of personality traits in spoken conversations from the AMI Meetings Corpus," in *Thirteenth Annual Conference of the International Speech Communication Association*, InterSpeech '12, 2012.

[32] N. A. Madzlan, J. Han, F. Bonin, and N. Campbell, "Automatic recognition of attitudes in video blogs—prosodic and visual feature analysis," in *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*, InterSpeech '14, 2014.

[33] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[34] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of personality traits from human spoken conversations," in *Proceedings of the*

Twelfth Annual Conference of the International Speech Communication Association, 2011.

[35] F. Alam, E. A. Stepanov, and G. Riccardi, "Personality traits recognition on social network - facebook," in *Proceedings of the International AAAI Conference on Web and Social Media*, WCPR (ICWSM - 13), (Cambridge, MA), 2013.

[36] S. Nowson and A. J. Gill, "Look! who's talking?: Projection of extraversion across different social contexts," in *Proceedings of the ACM Multi Media on Workshop on Computational Personality Recognition*, WCPR '14, pp. 23–26, ACM, 2014.

[37] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, and M. De Cock, "Computational personality recognition in social media," *User Modeling and User-Adapted Interaction*, vol. 26, 02 2016.

[38] S. Gievska and K. Koroveshovski, "The impact of affective verbal content on predicting personality impressions in YouTube videos," in *Proceedings of the ACM Multi Media on Workshop on Computational Personality Recognition*, WCPR '14, pp. 19–22, ACM, 2014.

[39] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proceedings of the 10th International Conference on Multimodal interfaces*, ICMI '08, pp. 53–60, ACM, 2008.

[40] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself: automatic personality assessment using short self-presentations," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI'11, pp. 255–262, ACM, 2011.

[41] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro, "Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection," in *Proceedings of the Workshop on Tagging, Mining and Retrieval of Human Related Activity Information*, pp. 9–14, ACM, 2007.

[42] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion—a systematic study," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 443–455, 2012.

[43] T. Polzehl, S. Moller, and F. Metze, "Automatically assessing personality from speech," in *Proceedings of the IEEE Fourth International Conference on Semantic Computing*, ICSC '10, pp. 134–140, IEEE, 2010.

[44] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[45] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "FaceTube: predicting personality from facial expressions of emotion in online conversational video," in *Proceedings of the 14th ACM International Conference on Multimodal interaction*, ICMI '12, pp. 53–56, ACM, 2012.

[46] D. Sanchez-Cortes, J.-I. Biel, S. Kumano, J. Yamato, K. Otsuka, and D. Gatica-Perez, "Inferring mood in ubiquitous conversational video," in *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, MUM '13, p. 22, ACM, 2013.

[47] S. Zhao, G. Ding, J. Han, and Y. Gao, "Personality-aware personalized emotion recognition from physiological signals," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 1660–1667, International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[48] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM Transactions on Multimedia Computing Communication Applications*, vol. 15, pp. 14:1–14:18, Jan. 2019.

[49] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[50] V. Kollia and N. Tayebi, "A controlled set-up experiment to establish personalized baselines for real-life emotion recognition," *CoRR*, vol. abs/1703.06537, 2017.

[51] V. Kollia, "Personalization effect on emotion recognition from physiological data: An investigation of performance on different setups and classifiers," *CoRR*, vol. abs/1607.05832, 2016.

[52] M. K. Abadi, J. A. M. Correa, J. Wache, H. Yang, I. Patras, and N. Sebe, "Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos," in *Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, vol. 1 of *FG '15*, pp. 1–8, IEEE, 2015.

[53] L. Batrinca, B. Lepri, N. Mana, and F. Pianesi, "Multimodal recognition of personality traits in human-computer collaborative tasks," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pp. 39–46, ACM, 2012.

[54] V. Ponce-López, S. Escalera, and X. Baró, "Multi-modal social signal analysis for predicting agreement in conversation settings," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pp. 495–502, ACM, 2013.

[55] V. Ponce-López, S. Escalera, M. Pérez, O. Janés, and X. Baró, "Non-verbal communication analysis in Victim–Offender Mediations," *Pattern Recognition Letters*, vol. 67, pp. 19–27, 2015.

[56] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pp. 205–211, ACM, 2004.

[57] R. J. Vernon, C. A. Sutherland, A. W. Young, and T. Hartley, "Modeling first impressions from highly variable facial images," *Proceedings of the National Academy of Sciences*, vol. 111, no. 32, pp. E3353–E3361, 2014.

[58] R. Qin, W. Gao, H. Xu, and Z. Hu, "Modern physiognomy: an investigation on predicting personality traits and intelligence from the human face," *Science China Information Sciences*, vol. 61, no. 5, p. 058105, 2018.

[59] F. Gürpınar, H. Kaya, and A. A. Salah, "Combining deep facial and ambient features for first impression estimation," in *Proceedings of the European Conference on Computer Vision*, ECCV '16, pp. 372–385, Springer, 2016.

[60] K. Ilmini and T. Fernando, "Persons' personality traits recognition using machine learning algorithms and image processing techniques," *Advances in Computer Science: An International Journal*, vol. 5, no. 1, pp. 40–44, 2016.

[61] O. Celiktutan and H. Gunes, "Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 29–42, 2017.

[62] J. Lin, W. Mao, and D. D. Zeng, "Personality-based refinement for sentiment classification in microblog," *Knowledge-Based Systems*, vol. 132, pp. 204–214, 2017.

[63] C. Sarkar, S. Bhatia, A. Agarwal, and J. Li, "Feature analysis for computational personality recognition using youtube personality data set," in *Proceedings of the ACM Multi Media on Workshop on Computational Personality Recognition*, WCPR '14, pp. 11–14, ACM, 2014.

[64] F. Alam and G. Riccardi, "Predicting personality traits using multimodal information," in *Proceedings of the ACM Multi Media on Workshop on Computational Personality Recognition*, WCPR '14, pp. 15–18, ACM, 2014.

[65] M. Sidorov, S. Ultes, and A. Schmitt, "Automatic recognition of personality traits: A multimodal approach," in *Proceedings of the Mapping Personality Traits Challenge and Workshop*, pp. 11–15, ACM, 2014.

[66] G. Farnadi, S. Sushmita, G. Sitaraman, N. Ton, M. De Cock, and S. Davalos, "A multivariate regression approach to personality impression recognition of Vloggers," in *Proceedings of the ACM Multi Media on Workshop on Computational Personality Recognition*, WCPR '14, pp. 1–6, ACM, 2014.

[67] A. G. Wright, "Current directions in personality science and the potential for advances through computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 292–296, 2014.

[68] A. Vinciarelli and G. Mohammadi, "More personality in personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 297–300, 2014.

[69] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[70] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[71] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '13, pp. 3687–3691, IEEE, 2013.

[72] F. Liu, J. Perez, and S. Nowson, "A language-independent and compositional model for personality trait recognition from short texts," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 754–764, 2017.

[73] G. An and R. Levitan, "Lexical and acoustic deep learning model for personality recognition," *Proceedings of Interspeech 2018*, pp. 1761–1765, 2018.

[74] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[75] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep bimodal regression for apparent personality analysis," in *Proceedings of the European Conference on Computer Vision*, ECCV '16, pp. 311–324, Springer, 2016.

[76] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *Proceedings of the European Conference on Computer Vision*, ECCV '16, pp. 337–348, Springer, 2016.

[77] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition," in *Proceedings of the European Conference on Computer Vision*, ECCV '16, pp. 349–358, Springer, 2016.

[78] J. Yu and K. Markov, "Deep learning based personality recognition from facebook status updates," in *Proceedings of the IEEE 8th International Conference on Awareness Science and Technology*, iCAST '17, pp. 383–387, IEEE, 2017.

[79] J. V. Kasmar, W. V. Griffin, and J. H. Mauritzen, "Effect of environmental surroundings on outpatients' mood and perception of psychiatrists," *Journal of Consulting and Clinical Psychology*, vol. 32, no. 2, p. 223, 1968.

[80] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[81] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proceedings of the European Conference on Computer Vision*, ECCV '16, pp. 630–645, Springer, 2016.

[83] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[84] S. Hochreiter, "Untersuchungen zu Dynamischen Neuronalen Netzen (Studies on Dynamic Neural Networks) (in German)," 1991.

[85] Y. Bengio, P. Simard, P. Frasconi, *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[86] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[87] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proceedings of the Ninth International Conference on Artificial Neural Networks*, vol. 2 of *ICANN '99*, pp. 850–855, IEEE, 1999.

[88] A. C. Little and D. I. Perrett, "Using composite images to assess accuracy in personality attribution to faces," *British Journal of Psychology*, vol. 98, no. 1, pp. 111–126, 2007.

[89] B. Fink, N. Neave, J. T. Manning, and K. Grammer, "Facial symmetry and the 'big-five' personality factors," *Personality and Individual Differences*, vol. 39, no. 3, pp. 523–529, 2005.

[90] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[91] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," Tech. Rep.

CMU-CS-16-118, Carnegie Mellon University, School of Computer Science, 2016.

[92] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '17, pp. 131–135, IEEE, 2017.

[93] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *CoRR*, vol. abs/1609.08675, 2016.

[94] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian, "Hmm-based audio keyword generation," in *Advances in Multimedia Information Processing - PCM 2004* (K. Aizawa, Y. Nakamura, and S. Satoh, eds.), (Berlin, Heidelberg), pp. 566–574, Springer Berlin Heidelberg, 2005.

[95] G. Stemmler and J. Wacker, "Personality, emotion, and individual differences in physiological responses," *Biological Psychology*, vol. 84, no. 3, pp. 541–551, 2010.

[96] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker, "Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life," *Journal of Personality and Social Psychology*, vol. 90, no. 5, p. 862, 2006.

[97] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One billion word benchmark for measuring progress in statistical language modeling," *CoRR*, vol. abs/1312.3005, 2013.

[98] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '2018, pp. 2227–2237, June 2018.

[99] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[100] B. Chen, S. Escalera, I. Guyon, V. Ponce-López, N. Shah, and M. O. Simón, "Overcoming calibration problems in pattern labeling with pairwise ratings: application to personality traits," in *Proceedings of the European Conference on Computer Vision*, ECCV '16, pp. 419–432, Springer, 2016.

[101] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[102] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, vol. Volume 37 of *ICML'15*, pp. 448–456, 2015.

[103] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[104] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, ICCV '15, pp. 1026–1034, 2015.

[105] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*, Interspeech 2014, 2014.

[106] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *Journal of Machine Learning Research*, vol. 3, no. Aug, pp. 115–143, 2002.

[107] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *CoRR*, vol. abs/1601.06733, 2016.

[108] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 563–582, 2001.

[109] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.

[110] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pp. 2818–2826, 2016.

[111] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 4 of *AAAI '17*, p. 12, 2017.

[112] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pp. 770–778, 2016.

[113] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[114] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '18, pp. 4510–4520, IEEE, 2018.

[115] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '18, pp. 8697–8710, IEEE Computer Society, 2018.

[116] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[117] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[118] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[119] J. Allen, "Applications of the short time Fourier transform to speech processing and spectral analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7 of *ICASSP '82*, pp. 1012–1015, IEEE, 1982.

[120] A. V. Oppenheim, J. R. Buck, and R. W. Schafer, *Discrete-time Signal Processing. Vol. 2.* Upper Saddle River, NJ: Prentice Hall, 2001.

[121] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil, "Universal sentence encoder," *CoRR*, vol. abs/1803.11175, 2018.

[122] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[123] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

[124] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Workshop Track Proceedings of the 1st International Conference on Learning Representations*, ICLR 2013, (Scottsdale, Arizona, USA), 2013.

[125] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pp. 1724–1734, 2014.

[126] X. Wei, C. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, pp. 303–315, July 2018.

[127] F. Gürpinar, H. Kaya, and A. A. Salah, "Multimodal fusion of audio, scene, and face features for first impression estimation," in *Proceedings of the 23rd International Conference on Pattern Recognition*, ICPR '16, pp. 43–48, IEEE, 2016.

[128] S. Eddine Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed, "Personality traits and job candidate screening via analyzing facial videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPR '17, pp. 10–13, 2017.

[129] H. Kaya, F. Gurpinar, and A. Ali Salah, "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*, CVPR '17, pp. 1–9, 2017.