MOHAMED KHIDER UNIVERSITY - BISKRA
FACULTY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF ELECTRICAL ENGINEERING
DIVISION OF ELECTRONICS
Ref: .....

جامعة محمد خيضر - بسكرة
كلية العلوم والتكنولوجيا
قسم الهندسة الكهربائية
شعبة الإلكترونيك
المرجع: .....

Face image analysis in dynamic scenarios

By

TELLI HICHEM

Thesis submitted to the department of electrical engineering in candidacy for the Degree of **Doctorate (3rd Cycle)** in **Biometrics**.

Defended on:

Members of the jury:

President:	OUAFI Abdelkrim	Prof	University of Biskra
Supervisor:	SBAA Salim	Prof	University of Biskra
Co-supervisor:	DORNAIKA Fadi	Prof	University of the Basque Country UPV/EHU, Spain
Examiner:	BAARIR Zine-Eddine	Prof	University of Biskra
Examiner:	BENLAMOUDI Azeddine	MCA	University of Ouargla

FEBRUARY 2023

Dedication

Alhamdulillah, I praise and thank Allah Subḥānahu wa Ta'āla for His greatness and for giving me the strength and courage to complete this thesis.

I dedicate this modest work to:

- My dear parents, I want you to know that I hold you in the highest esteem and care, and I hope that you feel that from reading this.
- My brothers for being there for me whenever I needed a boost in spirit.
- My wife, for her continuous support, as well as her family.
- My daughter, who made me stronger.
- All my teachers during my academic career.
- All my friends and colleagues, especially these two: Dr. Bekhouche Salah Eddine, Dr. Bougourzi Fare.

In closing, I want to thank everyone who has helped me along the way.

Acknowledgements

Pr. Salim Sbaa, my co-supervisor, Pr. Fadi Dornaika and my mentors Dr. Salah Eddine Bekhouche, and Pr. Abdelmalik Taleb-Ahmed, have provided me with essential guidance and support throughout my Ph.D. studies. Without their insightful guidance, the thesis would have never been completed.

I would also want to thank the jury members for agreeing to read and evaluate this manuscript.

In addition, I send my profound gratitude and thanks to all the instructors who have supported and directed me in my work with a critical and discerning eye.

Abstract

Abstract

Automatic personality analysis using computer vision is a relatively new research topic. It investigates how a machine could automatically identify or synthesize human personality. Utilizing time-based sequence information, numerous attempts have been made to tackle this problem. Various applications can benefit from such a system, including pre-screening interviews and personalized agents.

In this thesis, we address the challenge of estimating the Big-Five personality traits along with the job candidate screening variable from facial videos. We proposed a novel framework to assist in solving this challenge. This framework is based on two main components: (1) the use of Pyramid Multi-level (PML) to extract raw facial textures at different scales and levels; and (2) the extension of the Covariance Descriptor (COV) to combine several local texture features of the face image, such as Local Binary Patterns (LBP), Local Directional Pattern (LDP), Binarized Statistical Image Features (BSIF), and Local Phase Quantization (LPQ). The video stream features are then represented by merging the face feature vectors, where each face feature vector is formed by concatenating all

the PML-COV feature blocks. These rich low-level feature blocks are obtained by feeding the textures of PML face parts into the COV descriptor.

The state-of-the-art approaches are even hand-crafted or based on deep learning. The Deep Learning methods perform better than the hand-crafted descriptors, but they are computationally and experimentally expensive. In this study, we compared five hand-crafted methods against five methods based on deep learning in order to determine the optimal balance between accuracy and computational cost. The obtained results of our PML-COV framework on the ChaLearn LAP APA2016 dataset compared favourably with the state-of-the-art approaches, including deep learning-based ones. Our future aim is to apply this framework to other similar computer vision problems.

Keywords: Computer vision, ChaLearn, APA2016 dataset, First impression, Big-Five personality traits, job candidate screening, Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience, multi-media CV, PML-COV descriptor, framework, PML, LDP, LPQ, BSIF, LBP, COV, VGG, Resnet, SE-Resnet, Arcface, MobileFaceNets, feature selection, Relief, MRMR, NCA, SVM, regression, SVR, GPR, MAE.

الملخص

يعد تحليل الشخصية التلقائي باستخدام الرؤية الحاسوبية (computer vision) موضوعًا بحثيًا جديدًا نسبيًا. إنه يبحث على امكانية تزويد الآلة بالقدرة على تحديد أو تجميع الشخصية البشرية تلقائيًا. لقد تم إجراء العديد من المحاولات لمعالجة هذه المشكلة باستخدام المعلومات التسلسلية المستندة إلى الوقت. يمكن أن تستفيد مختلف التطبيقات من مثل هذا النظام، بما في ذلك تطبيق الفرز المسبق للمقابلات، وتطبيق الوكلاء الشخصيين.

في هذه الأطروحة ، نتناول التحدي المتمثل في تقدير سمات الشخصية الخمس الكبرى جنبًا إلى جنب مع متغير فرز مرشح الوظيفة من مقاطع فيديو للوجه. لقد اقترحنا إطار عمل (framework) جديدًا للمساعدة في حل هذا التحدي. يعتمد هذا الإطار على مكونين رئيسيين: (1) استخدام الهرم متعدد المستويات (PML) لاستخراج انماط الوجه الدفينة بمقاييس ومستويات مختلفة ؛ و (2) الاضافة لواصف التباين المشترك (COV) للجمع بين العديد من الميزات النمطية المحلية لصورة الوجه، مثل الأنماط الثنائية المحلية (LBP) ، ونمط الاتجاه المحلي (LDP) ، وميزات الصورة الإحصائية الثنائية (BSIF) ، واصف المرحلة التكميمية المحلية .(LPQ) يتم بعد ذلك تمثيل ميزات تسلسل الفيديو من خلال دمج مصفوفة الأشعة لميزات الوجه، حيث يتم تشكيل كل شعاع لميزة الوجه من خلال ربط تسلسلي لجميع كتل الميزات PML-COV. يتم الحصول على كتل الميزات الغنية بأنماط منخفضة المستوى هذه عن طريق تغذية واصف التباين المشترك COV بجميع الأنسجة المتحصلة من أجزاء الوجه الممثلة بهرم متعدد المستويات PML.

الأساليب الحديثة تكون مصنوعة يدويًا او بالاعتماد على التعلم العميق.يعد أداء أساليب التعلم العميق أفضل من الواصفات اليدوية، لكنها باهظة الثمن من الناحية الحسابية والتجريبية. في هذه الدراسة ، قمنا بمقارنة خمس طرق مصنوعة يدويًا بخمس طرق تعتمد على التعلم العميق لإيجاد المفاضلة الصحيحة بين الدقة وتكلفة الحساب. تمت مقارنة النتائج التي تم الحصول عليها من إطار عملنا PML-COV على قاعدة بيانات ChaLearn LAP APA2016 بشكل إيجابي مع أحدث الأساليب ، بما في ذلك الأساليب القائمة على التعلم العميق. هدفنا المستقبلي هو تطبيق هذا الإطار على مشاكل أخرى مماثلة في الرؤية الحاسوبية.

الكلمات الدلالية: رؤية الحاسوب، ChaLearn، مجموعة بيانات APA2016 ، الانطباع الأول، السمات الشخصية الخمس الكبرى ، فحص مرشح الوظيفة، الانبساط، الوفاق ، الضمير ، العصابية ، الانفتاح على التجارب، السيرة الذاتية متعددة الوسائط، ، واصف PML-COV، PML، LDP، LPQ، BSIF، LBP، COV، VGG ،Resnet ، SE-Resnet، Arcface، MobileFaceNets، اختيار الميزات، Relief، MRMR، NCA، SVM، الانحدار، SVR، GPR، MAE.

Scientific Productions

Publications in journals

- **Telli, H.**, Sbaa, S., Bekhouche, S.E., Dornaika, F., Taleb-Ahmed, A., López, M.B. (2021). A novel multi-level Pyramid Co-Variance operators for estimation of personality traits and job screening scores. Traitement du Signal, Vol. 38, No. 3, pp. 539-546. https://doi.org/10.18280/ts.380301

Publications in international conferences

- **Djelfa on 29-30 April, Algeria [1]:** Chergui, A., Ouchtati, S., **Telli, H.**, Bougourzi, F., & Bekhouche, S. E. (2018, April). LPQ and LDP descriptors with ml representation for kinship verification. In The second edition of the International Workshop on Signal Processing Applied to Rotating Machinery Diagnostics (SIGPROMD'2018) (pp. 1-10).

- **Hammamet on 19-22 March, Tunisia [2]:** Chergui, A., Ouchtati, S., Sequeira, J., Bekhouche, S. E., Bougourzi, F., & **Telli, H.** (2019, March). Deep features for kinship verification from facial images. In 2019 International Conference on Advanced Systems and Emergent Technologies (IC_ASET) (pp. 64-67). IEEE. https://doi.org/10.1109/ASET.2019.8871011

- **Skikda on 10-11 December, Algeria [3]:** Chergui, A., Ouchtati, S., Sequeira, J., Bekhouche, S. E., & **Telli, H.** (2018, December). Robust kinship verification using local descriptors. In Proc. Third Int. Conf. Adv. Technol. and Electr. Eng.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AMT Amazon Mechanical Truk.

ANN Artificial Neural Networks.

ArcFace Additive Angular Margin Loss.

ARD Automatic Relevance Determination.

ASR Automatic Speech Recognition.

BSIF Binarized Statistical Image Features.

CNN Convolutional Neural Network.

COV Co-Variance Operator descriptor.

CV curriculum vitae.

CVs curricula vitae.

DAN+ Descriptor Aggregation Network.

DBR Deep Bimodal Regression.

DESC Desired hand-crafted descriptor.

DFT Discrete Fourier Transform.

GP Gaussian process.

GPR Gaussian Process Regression.

ICA Independent Component Analysis.

ILSVRC The ImageNet Large Scale Visual Recognition Challenge.

LBP Local Binary Pattern.

LDP Local Directional Pattern.

LGBP-TOP Gabor Binary Patterns from Three Orthogonal Planes.

LPQ Local Phase Quantization.

LSTM Long short-term memory.

MAE mean absolute error.

MB Multi-block.

MIQ Mutual Information Quotient.

ML Multi-level.

MRMR Minimum redundancy maximum relevance.

NCA Neighborhood component analysis.

PC Pearson correlation coefficient.

PML Pyramid multi-level.

PML-COV Pyramid Multi-Level Co-Variance Operator descriptor.

ResNet residual neural network.

ResNet-50 residual neural network consisting of 50 trainable layers.

RGB Red, Green and Blue.

RMS Root Mean Square.

SE-ResNet-50 squeeze-and-excitation based ResNet-50.

SMO Sequential minimal optimisation.

SVM Support Vector Machine.

SVR Support Vector regression.

SVRs Support Vector regressors.

VGG16 Visual Geometry Group.

GENERAL INTRODUCTION

Introduction

A machine that can recognize or synthesize human personalities through automatic perception is the subject of automated personality perception and synthesis. As humans, we assess the personalities of others at first glance, even without having interacted with them. We are able to make this assessment in a fraction of a second due to our rapid reaction time [10].

Personality traits and their classification have been the subject of many studies over the past several decades. In this context, several models have been proposed, such as the Big-Five [11], BigTwo [12], or 16PF [13], among many others. The Big-Five (or Five-Factor Model) is a personality model widely used in the field of psychology. It characterizes an individual's personality based on five independent dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience.

Automatic personality analysis using computer vision is a relatively new research topic, where various applications can use personality analysis systems such as prescreening interviews, personalized agents, criminal activities, and political ideology. The first competition in this field was the *ChaLearn Looking at People 2016 First impression challenge*. It targeted researchers around the world to try to solve the problem of identifying these Big-Five personality traits from facial videos. Due to its success, another competition was proposed. The *ChaLearn Looking at People CVPR 2017 Challenge* came with an extension of the problem, namely adding a score for the screening attribute of the applicant to be estimated along with the Big-Five personality traits.

Motivation and problem statement

Due to the lack of systems that can synthesize a human personality from a video, the *ChaLearn LAP 2016 APA dataset* has been introduced to help solve the problem of identifying the personality traits and job interview screening variable from facial videos. Governments and big companies will benefit from such a system because they will be able to sort through the curriculum vitae (CV) of their candidates using automatic recommendations based on multi-media CV. This system will predict whether the candidates are promising enough that the recruiter wants to invite him/her to an interview. The approaches to solving this problem are either hand-crafted or based on deep learning. Deep learning approaches are the best way to solve this problem, but they are computationally very expensive and time-consuming methods. Also, CNN have

several hyper-parameters to be tuned, which makes finding these hyper-parameters a tedious task, and they depend on large amounts of labeled data in order to successfully train a reliable model. The primary objective is to find the optimal balance between accuracy and computational efficiency. The hand-crafted approach typically finds the optimal balance between accuracy and computational power, and it depends on less labeled data.

Contribution

In this manuscript, we propose the use of a computationally efficient hand-crafted descriptor that can extract low-level facial features from video sequences. This descriptor naturally merges multiple local texture features using a Pyramid Multi-Level (PML) representation [14] and a Co-Variance Operator (COV). It extracts and fuses information from multiple scales and face regions.

Inspired by our previous work in descriptors applied to the discrimination among classes [15] [16], we improve it by making the following modifications: (i) we improve the selection of the low-level image descriptors that feed the Co-Variance Operator (COV), (ii) we modify the feature selection scheme to produce a real score value, (iii) we apply the descriptor to a regression task from facial videos.

The contributions of the paper can be summarized as follows:

- A novel Pyramid Multi-Level Co-Variance Operator descriptor (PML-COV); a low computational cost descriptor that extends Co-Variance Operator to solve regression problems from videos.
- The application of the novel descriptor to obtain state-of-the-art results in estimating personality traits and job screening scores, using benchmark datasets.

Thesis structure

The remaining of the thesis are structured as follows:

In the first chapter, we provide a brief summary of the historical approaches that utilize the ChaLearn LAP 2016 APA dataset to estimate personality traits and job screening scores. These approaches are even hand-crafted or based on deep learning. In the second chapter, we begin with a basic psychological overview of personality and the Big Five traits that explain it, including their primary distinctions. Then, we describe

the used database in detail, including its statistics, the methodologies used to gather it, the strategies utilized to label it, and a sample screenshot of labeled videos from it. Finally, we describe the evaluation protocols. In the third chapter, we described the general structure of the employed techniques: hand-crafted and deep, including the proposed PML-COV framework [17] and the seven steps that have been used to obtain the interview variable, which are: face preprocessing, feature representation, feature extraction, video descriptor computation, feature selection, personality traits estimation, and interview variable estimation. In the fourth chapter, we describe our proposed framework. Then, hierarchical experiments are conducted to determine the optimal configuration suited for our framework, which involves analyzing the effects of the statistical descriptor, the effects of PML level, and the impact of hyper-parameter optimization on the final results. We compared the obtained findings with the state-of-art methods, including those based on deep learning. In conclusion, we summarized our research and talked about what we plan to do next.

CHAPTER 1

LITERATURE REVIEW

1.1 Introduction

In recent years, many experts have made some advances in the field of personality assessment by evaluating personality traits using visual information. The developed algorithms aim to estimate the personality traits and job screening scores from two types of data, either from only the visual modalities [4] [18] or from the combination of multiple modalities [5] [6] [7] [8]. The first survey on automatic personality detection, perception, and synthesis was presented by Vinciarelli and Mohammadi [19]. It summarizes the models based on features that most effectively predict measurable aspects in people's lives. In 2018, Escalante et al. [18] reviewed and investigated the mechanisms related to first impression analysis, and summarized the results of the CVPR 2017 Challenge, while the most recent review of previous image-based approaches to overt personality trait detection is presented by Jacques Junior et al. [20]. In this chapter, we will describe the recent hand-crafted and deep learning based methods that were developed to estimate the personality traits and job screening scores from either visual modalities or audio-visual modalities, which we will use them to rank our approaches performance.

1.2 Hand-crafted methods

Hand-crafted based techniques have been known for decades, and still serve as a powerful tool when combined with machine learning classifiers, or regressors. The hand-crafted methods are manually engineered using a numerical model that is developed with previous knowledge of specific attributes to overcome certain obstacles. In general the used hand-crafted methods follows these three main steps which starts by data preprocessing, then feature extraction and selection, finally these features are fed to a machine learning classifiers or regressors to estimate the a given class or score.

The FDMB team [18] used frame differences and Local Phase Quantization (LPQ)) [21] descriptors at several fixed image regions with the support vector regression (SVR) [22] technique to predict the interview variable and the Big-Five traits. After face detection and normalization, differences between successive frames were computed. Then, they extract and concatenate LPQ features from each region of those frame-differences. The video representation is then obtained by adding these feature vectors. Finally, they employ an SVR machine to estimate the Big-five traits and the interview variable.

The ROCHCI team [18] extracted a set of multimodal features; firstly using the

SHORE library [23] to obtain visual information, and secondly using the audio signal to obtain pitch and intensity features, they also hand picked some terms from the Automatic Speech Recognition (ASR) transcriptions. Finally, these features were combined, and a gradient boosting regression algorithm [24] was applied to predict the personality traits and jobs screening score.

The (PML) team [4] uses only visual features. They first detect and normalise each detected face in each video. Then, they use two different texture descriptors, Local Phase Quantization (LPQ) [21] and Binarized Statistical Image Features (BSIF) [25], to extract face features, which are represented by Pyramid Multi-Level (PML). PML concatenates features from each region and each resolution of the image. When Local Phase Quantization (LPQ) is used, Figure 1.1 shows how the PML principle works at level four.

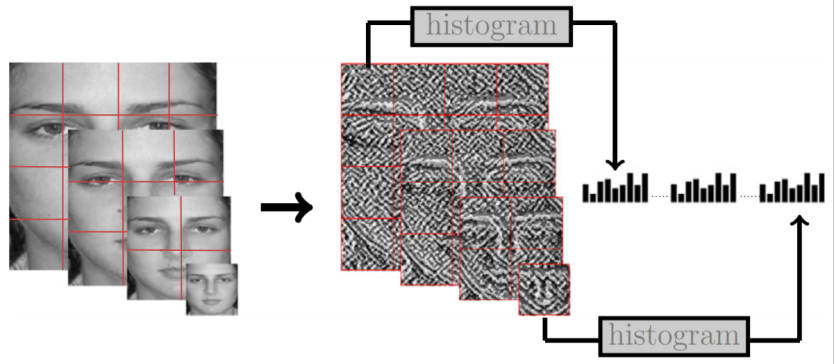


Figure 1.1: Pyramid Multi-Level: Local Phase Quantization at level 4 [4].

The temporal feature representation of the whole video is obtained by computing the mean over this sequence of feature vectors. For prediction (see Figure 1.2), they used five Support Vector Regressors (SVRs) [22] to estimate the big-five traits. The resulted estimation is used as an input feature for the final decision model, which is by using Gaussian Process Regression (GPR) [26] to estimate the invite for interview variable.

1.3 Deep Learning methods

Deep Learning methods are widely used nowadays in all aspects of life due to their capabilities to learn by example, which is inspired by the structure and function of the brain's artificial neural networks (ANN). Since [27] published their first deep learning architecture, "AlexNet", convolutional neural network (CNN) methods have become dominant in almost every field of computer vision.

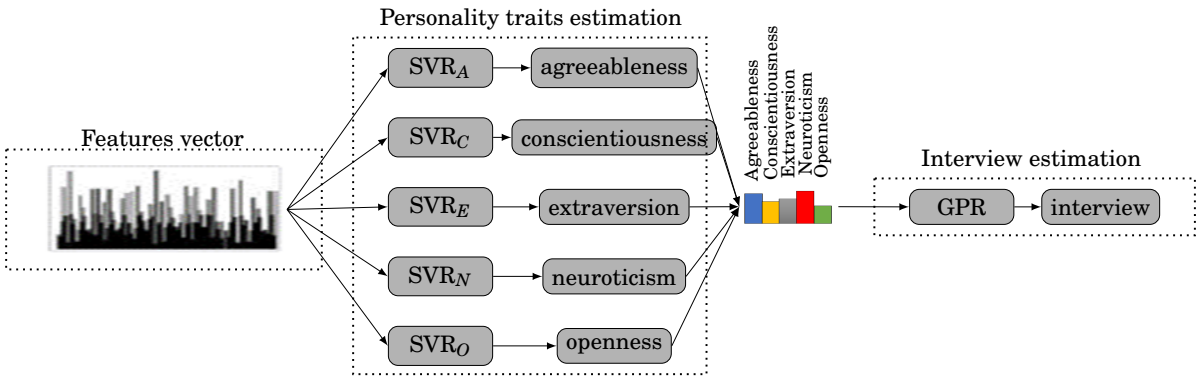


Figure 1.2: PML prediction methodology [4].

In [5] they introduce a Deep Bimodal Regression (DBR) frameworks which uses two type of modalities, In visual modality regression they first down-sample the video images extraction to sex frame per second and labeled with the same Big-Five traits values as the values of their corresponding video. Then fed them to a CNN model they named Descriptor Aggregation Network DAN+. Figure 1.3 show the architecture design of the proposed model (DAN+). Finally, they use the average score of images from a video as the predicted scores of that video. In audio modality, they used handcrafted spectral audio features as audio representations, and also employ deep learning based audio models. The final prediction of this framework is by fusing the results of these two modalities.

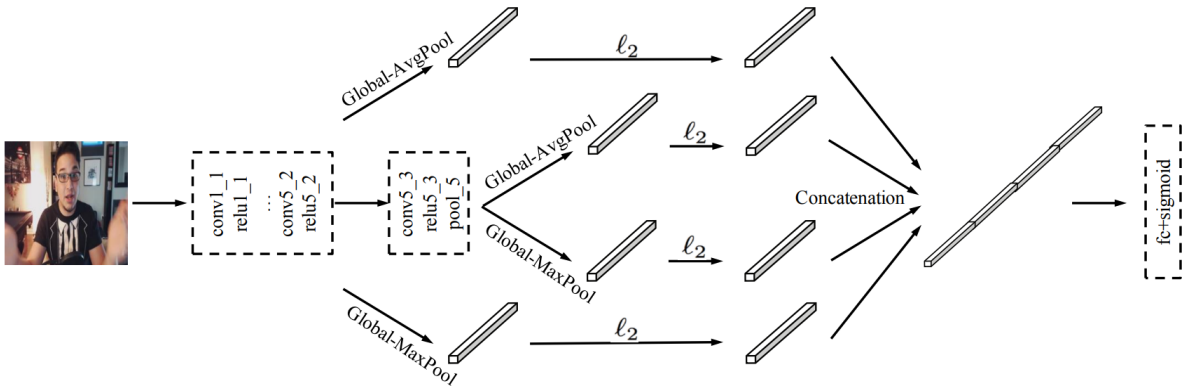


Figure 1.3: The DAN+ Model's Architecture [5].

Evolgen [6] proposed two end-to-end bi-model deep learning models using visual and audio features. These features are merged later on. They trained their network with temporally ordered audio and novel stochastic visual features from a few frames, taking care to avoid overfitting. They divided each video into N non-overlapping segments. Then, from each segment, the visual and audio features are extrated and fed to the proposed

models accordingly, as shown in figures 1.4, 1.5. The audio features are extracted using the mean and standard deviation. And for visual information, to prevent the background from influencing the perdictions, they used Dlib [28] 68 facial landmarks points to detect, segment, and align the face image.

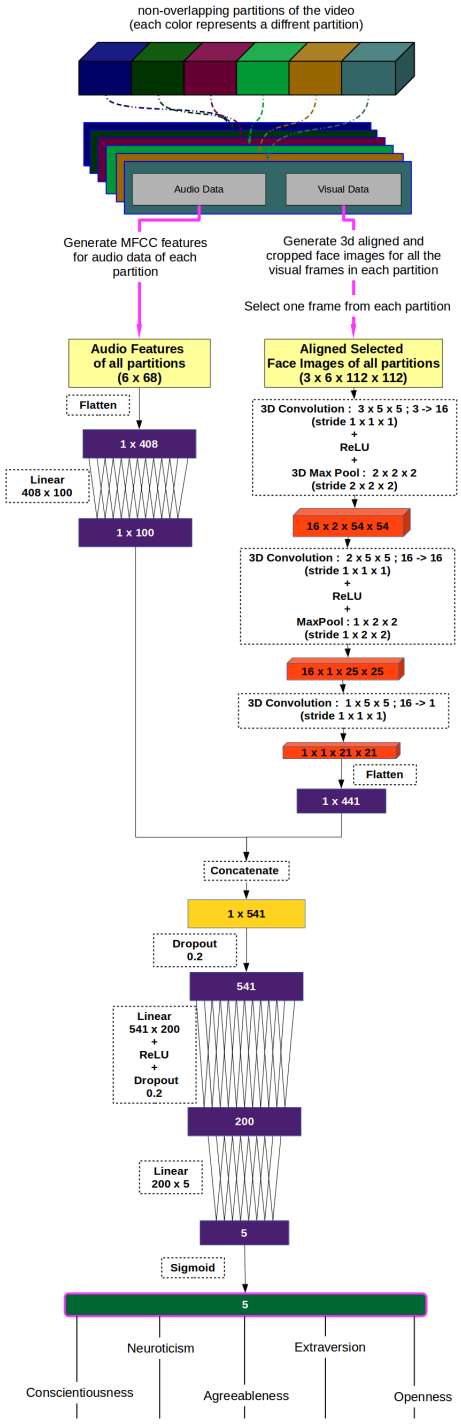


Figure 1.4: Architectures of the Evolgen bi-model volumetric CNN [6].

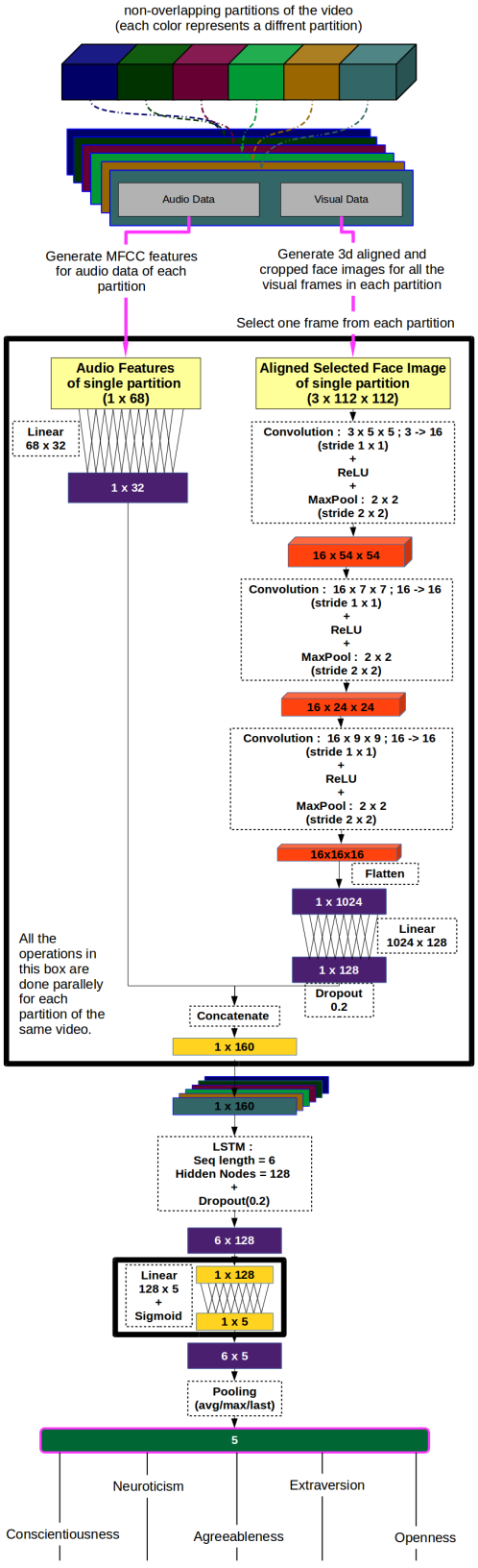


Figure 1.5: Architectures of the Evolgen bi-model LSTM neural network [6].

Baseline [7] proposed an audiovisual architecture based on deep learning. They used two similar residual streams of 17-layer followed by one fully connected layer, which outputs the score of the Big-Five personality traits. Each stream consists of one convolutional layer, followed by eight residual blocks containing two convolutional layers each (see figure 1.6). A random temporal cropping of the audio data taken from each video is used to feed the auditory stream. And the visual stream is also fed by a random $224 * 224$ spatial crop of random frames of visual data. Finally, they used a linear regression model to explain interview decision based on these traits predictions.

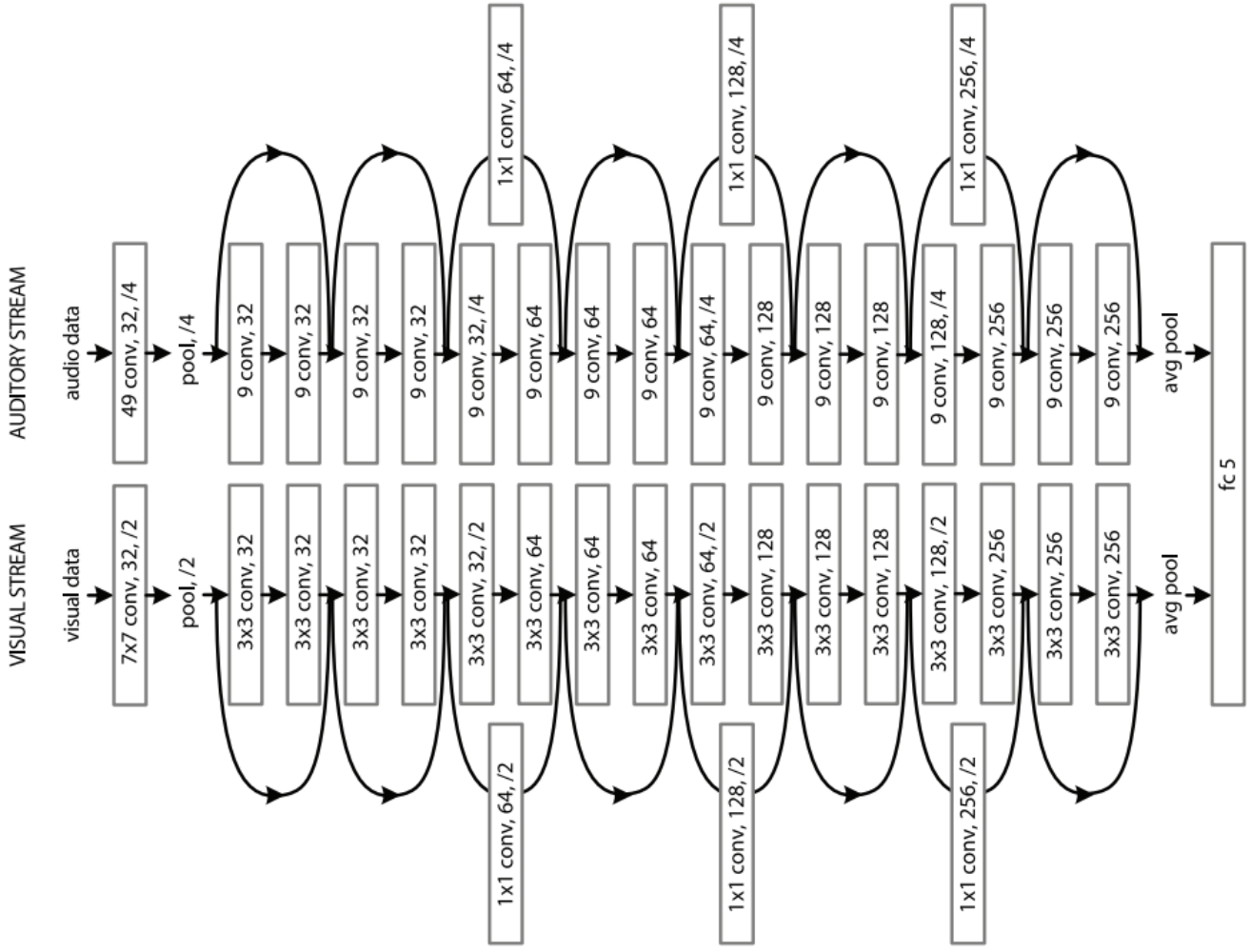


Figure 1.6: Baseline audiovisual architecture [7].

BU-NKU [8] developed a multi-modal system using face, scene, and audio features. A modality-specific regressors are used to predict apparent personality traits with an ensemble of decision trees. And a single decision tree, combined with a rule-based algorithm to predict the interview decision. The face image is detected and aligned

using the eyes' pose. The facial image is then cropped with a 20 percent margin and scaled to $64 * 64$ pixels. The visual features are computed by a combination of two techniques. The first is to feed the aligned faces to a modified version of the VGG-Face network [29] trained on facial emotion recognition. The second is using the Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [30] descriptor with 18 Gabor filters on the aligned faces. Finally, they summarize these video features using functional statistics descriptors, which include mean, standard deviation, offset, slope, and curvature. The scene features are extracted using the first frame only. The VGG-VD-19 network [31] pretrained on object recognition is used to get these features. The audio features are obtained by using an open source openSMILE tool [32]. Figure 1.7 illustrates the general structure of this system.

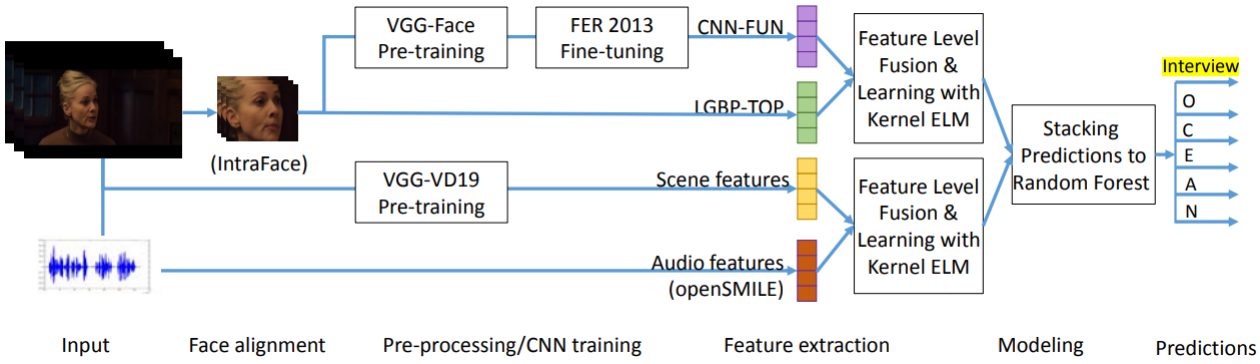


Figure 1.7: BU-NKU General structure [8].

1.4 Conclusion

In this chapter, we discussed different techniques used in the estimation of personality traits and interview variable. In fact, the state-of-the-art techniques are hand-crafted or based on deep learning. Hand-crafted descriptors are considered simple and suitable for real-time applications, as they can be easily deployed on low-cost hardware. However, they depend on perfect face alignment, so they are vulnerable to difficult face-pose scenarios. On the other hand, deep learning-based approaches are very good at solving highly complex problems and can be easily applied to similar problems. However, they rely on very expensive hardware, and their training is time-consuming. Also, they depend mostly on how much data there is and need careful choice of network design and hyper-parameters.

CHAPTER 2

FIRST IMPRESSIONS ANALYSIS

2.1 Background

Since personality is the most complex topic in psychology, scientists and researchers have paid great attention to it. Since then, it has become a touchstone and an area of discussion. It is the basis of the psychological formation of the personality [33]. According to the trait theory [34], a person's personality is just a collection of traits that overlap. Through these traits, behavior is explained and the person's psychological, behavioral, and professional path is determined. This shows the individual's subjectivity, as each behavior has its own meaning and significance in explaining how people act in different situations.

2.2 Big-Five personality traits

From this perspective, the Costa and McCrae model appeared [35], [36], [37], [38], which describes the traits, as it is one of the most important models that explain personality, and this model describes the personality through five factors so-named the Big Five, which are: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience [11]. They are the components that make up the personality of each individual. The five-factor model is used to help understand and even predict the relationships between personality traits and success in social, academic, and occupational challenges.

The Big Five is the most widely accepted personality trait in psychology, especially personality psychology, as it affects an individual's personality in various aspects of life. The individual lives a life of change at all levels of life accompanied by complications through which he lives a huge amount of pressure and tension that threaten the various aspects of his personal life. Whether it is family, social, or even professional including economic. Therefore, the individual has become vulnerable to facing many situations that may threaten his psychological strength, which increases his tension towards the drawbacks and challenges he will face at all levels, which cause stress and psychological problems, the most important of which is anxiety, which is one of the basic emotions and a natural part in the mechanisms of his behaviour [34].

2.2.1 Extraversion

Extraversion is characterized by a high level of emotional expressiveness, sociability, communicating with people, and excitability. Extraversion is a personality trait that

defines a person's friendliness, emotional expression, and social comfort. [39], [40].

- A high percentage indicates that a person prefers the presence of others in his life more, loves social events, and is full of energy and excitement.
- A low percentage indicates an introverted person with a conservative personality. His constant presence with others tires him. He prefers more activities alone, such as reading, his lifestyle is slower and tends to be calm and quiet.

2.2.2 Agreeableness

The trait of agreeableness indicates a person's overall affability, affection, trust, and altruism. These individuals are trustworthy because they are cooperative, care deeply about others, and are eager to provide assistance. [39], [40].

- The person who scores the highest for this trait is comfortable, kind, and friendly to others.
- A person who scores significantly lower on this trait is more manipulative, generally less friendly towards others, not concerned with the feelings or problems of others, and may also be seen as more competitive and less cooperative.

2.2.3 Conscientiousness

In general, people with this trait think deeply, have relatively good impulse control, and consistently strive to achieve their goals. People with this trait plan well for the future, think about how their behavior will affect others, and are aware of their responsibilities and duties [39], [40].

- People with the highest scores on this trait prefer to be organized with goal-oriented behaviors because they have an excellent ability to control their impulses and behaviors.
- People that are low on this trait are generally disorganized, cannot focus on accomplishing anything in particular, tend to be more chaotic, resistant to routines and routine-based tasks, and less concerned with the impact of their actions on the people around them.

2.2.4 Neuroticism

It is a personality trait that indicates a person's overall emotional instability. Neurotic people worry a lot and are prone to anxiety and depression even If everything goes well. They also tend to find things to worry about [39], [40].

- The person who scores the highest in this trait is moody, irritable, anxious, consistently pessimistic, and maybe depressed or have extreme mood swings.
- A person with a lower score on this trait is usually more emotionally stable, more resilient in interacting with people, and better at dealing with stress.

2.2.5 Openness to Experience

Openness is a personality trait that defines an individual's preference for intellectual, artistic, and creative pursuits [39], [40].

- High-scoring individuals on this trait tend to be highly intelligent, imaginative, or artistic. They usually have a strong sense of curiosity about the world around them and are always eager to expand their horizons.
- People who score poorly on this trait tend to be less interested in learning new things or coming up with fresh ideas. They dislike change and like to spend their time indoors whenever possible.

2.3 Job candidate screening variable

Job hiring interviews are important because they are the primary selection means for companies and professionals to find out who is qualified and who is not. Also, learn about various other personality traits of the candidate. This is due to the fact that personality influences workplace behavior and helps to better judge who will be able to be good in a given field. According to [41], an individual with high conscientiousness will outperform others in the workplace, although it is unclear if this is due to their perfectionist attitudes, punctuality, or systematic work approaches. [40] indicates that conscientiousness tends to have the strongest relationship with job success. In contrast to extraversion, neuroticism shows a significant negative correlation with job satisfaction.

2.4 Database

Personality trait estimation is a recent research topic. The number of databases available for this purpose is limited, particularly in the spatiotemporal domain. The ChaLearn LAP APA2016 dataset [9] is the only one we know of, which deals with the Big-Five personality traits and job candidate screening variable from a video sequence. To this end, the ChaLearn LAP APA2016 dataset is used in this research.

2.4.1 ChaLearn LAP APA2016 dataset

With more than 3,000 different YouTube high-definition (HD) videos of people facing the camera and talking in English in a self-presentation context. The ChaLearn LAP APA2016 dataset consists of 10,000 video clips gathered from publicly accessible YouTube videos, where certain criteria were required for the selection of these videos. These subjects are of different ages, genders, nationalities, and ethnicities. Additionally, the selection criteria are restricted to the following:

- Focused on Q & A videos, which are frequently linked to additional video contents including vlogs, HOWTOs, and beauty advice (mostly about makeup).
- Excellent audio and video quality.
- There should be no more than three videos per YouTube channel (author) to maintain the balance of unique subjects.
- In each of the 15-second videos, there should be only one face visible far from the camera.
- Subjects between the ages of 13 and 15 years old. Infants who cannot be identified but appear with their parents may be allowed.
- The camera should not move around excessively (changing the background is permitted, the foreground should not be constantly blurred).
- No inappropriate or violent content. Reject any defamatory, dubious, or troublesome content.
- No nudity, unless only the upper shoulders and neck are exposed.

– There may be people in the background, such as a crowd or an audience, who are not speaking and whose faces have a low resolution to prevent misunderstandings with the main subject.

– There will be no advertising (Information about products or business names in the form of video or audio).

– Cuts in the visual or audio stream should be avoided (abrupt changes).

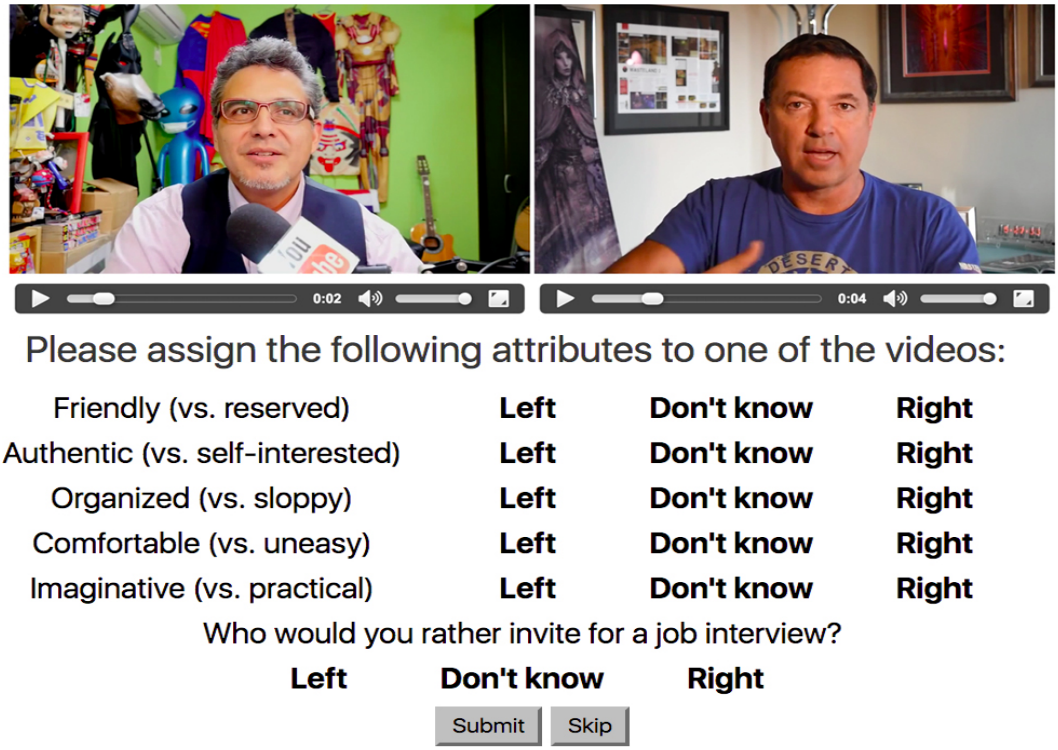


Figure 2.1: Custom-designed interface used for evaluating the Big-Five personality traits and job candidate screening variable on Amazon Mechanical Truk (AMT) [9].

A set of 3,060 unique originating videos with an average of 3.27 clips for each video are selected, which gives 10,000 video sequences of 15 seconds each and corresponds to $10000 * 15/3600 = 41.66$ hours of footage and 4.5 million frames. The selected video sequences are labeled using pairwise comparisons in a custom-designed interface (see Figure 2.1) using Amazon Mechanical Truk (AMT) to eliminate the problem of biased voters, such as, biases against racial, age, or gender, and cultural prejudices, which are extremely difficult to measure. Figure 2.2 shows a screenshot of sample videos in the dataset.

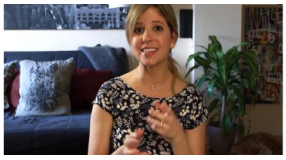
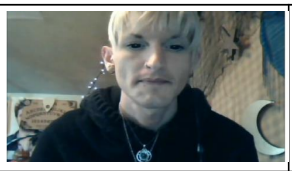
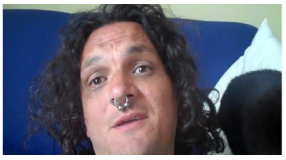
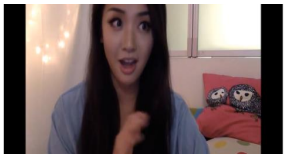
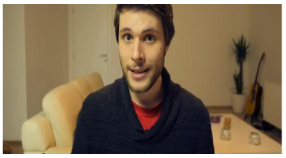
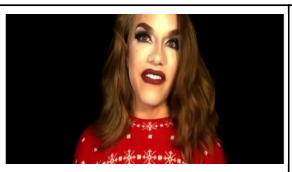
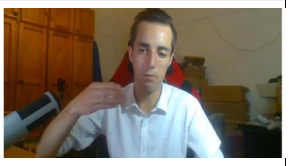
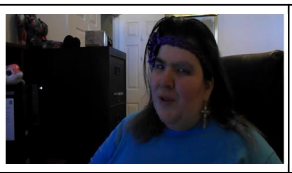
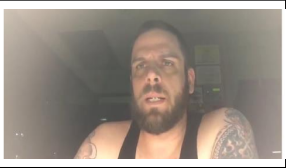
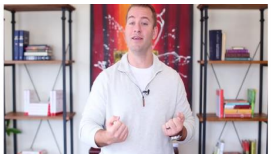
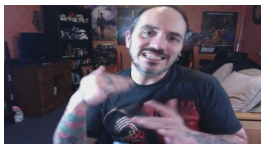
Extraversion			
Friendly		Reserved	
			
0.92523	0.8972	0.0093458	0.046729
Agreeableness			
Authentic		Self-interested	
			
0.93407	0.92308	0.087912	0.13187
Conscientiousness			
Organized		Sloppy	
			
0.97087	0.95146	0.1068	0.07767
Neuroticism			
Comfortable		Uneasy	
			
0.95833	0.96875	0.0625	0.072917
Openness			
Imaginative		Practical	
			
0.96667	0.97778	0.15555	0.16666

Figure 2.2: screenshot of sample videos in the dataset.

2.4.2 Evaluation metrics

For each video in the dataset, the ground truth labels for the Big-Five personality traits and interview variable were given by real values that fit the range $[0,1]$. We used mean accuracy to evaluate performance, which is given by Equation 2.1, for each personality trait and the interview variable. The performance P which is expressed in percentage, measures the accuracy of the model's predictions. This indicators were used in the previous challenge [9].

The Mean Absolute Error (MAE) represents the average absolute difference between the ground-truth scores and the estimated scores of the tested video sequences. For a paired data set $S = \{(x_i, y_i) \mid x_i \in \mathbb{R}, y_i \in \mathbb{R}; i = 1, ..., n\}$ with n observation, where x is the ground-truth scores and y is the estimated scores. MAE is given by Equation 2.2. Additionally, we also computed the linear correlation between ground-truth scores and estimated scores.

$$P = 100 \cdot (1 - MAE) \tag{2.1}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{2.2}$$

2.5 Conclusion

In this chapter, we explained what the job candidate screening variable is, and what the Big Five personality traits are and how they differ. Then we summarize each trait separately and explain the meaning of scoring high or low in each trait. Moreover, we described the used database, which is the ChaLearn LAP APA2016 dataset [9] as well as its collection methodology. Furthermore, we gave screenshots of example videos with strong and weak personality traits. Finally, we described the used evaluation metrics.

CHAPTER 3

PROPOSED APPROACH

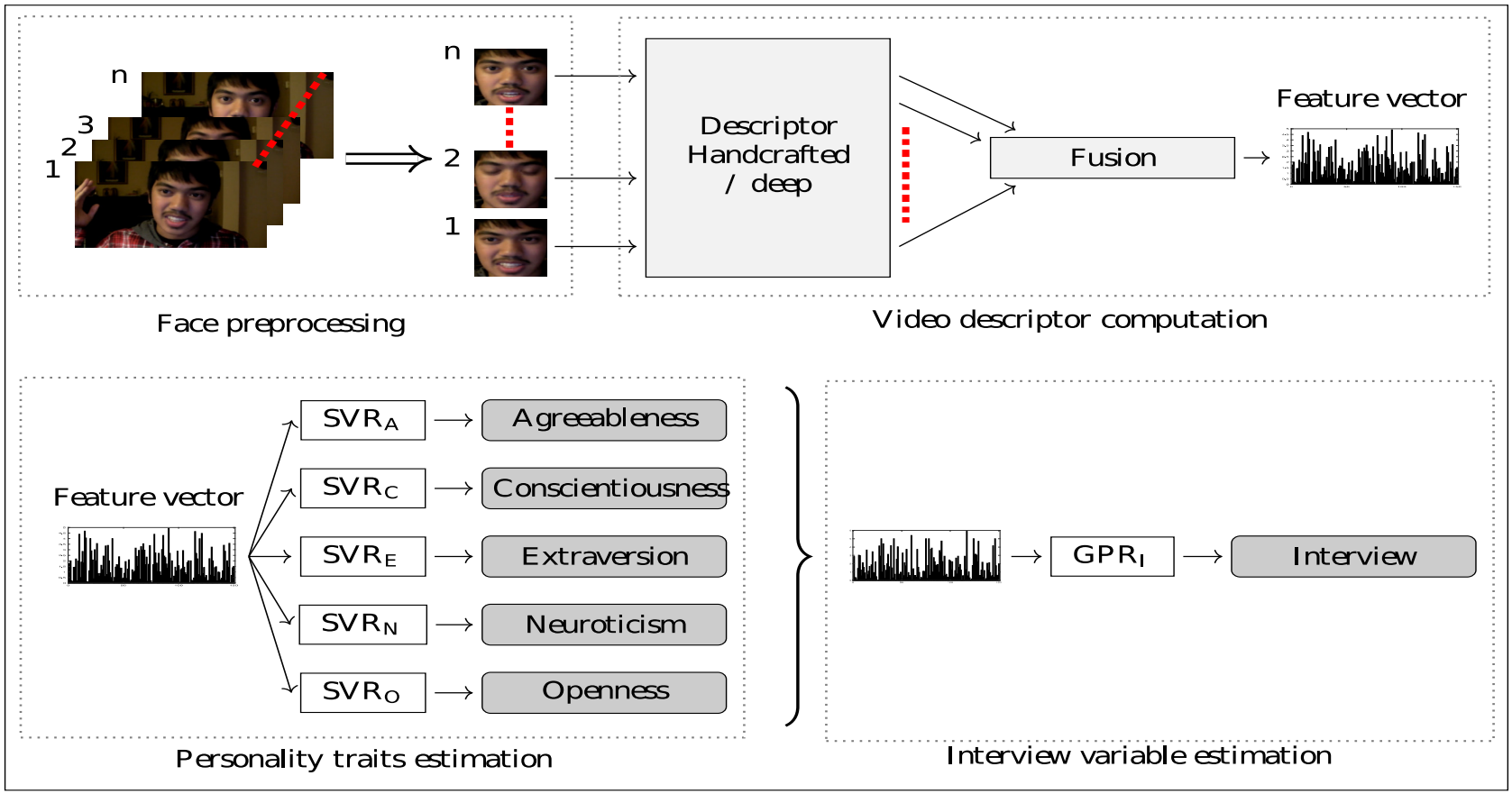


Figure 3.1: General structure of the proposed approach.

3.1 Introduction

In this chapter, we will describe the general structure of the used techniques, handcrafted and deep. In our work, seven steps are performed to obtain the interview score, which are: (i) face preprocessing, where the detected faces are aligned and cropped, (ii) feature representation, where the face image is divided into b block, (iii) feature extraction, where a set of features for each face is extracted, (iv) video descriptor computation, where we describe how the feature vector which represent the whole video is computed, (v) feature selection, where features are ranked in order to exclude possible irrelevant features, (vi) personality traits estimation, and (vii) interview variable estimation. Figure 3.1 illustrates the general structure of our approach.

3.2 Face preprocessing

In computer vision, face preprocessing is an important step. It help the model to improve by removing unnecessary information, not to forget that the original face can be affected by many factors, such as: pose variation, lighting variations, etc. To this end, face preprocessing plays an important role in the whole process. In this study, We first iterate through each video in the dataset. Then, in each frame, the face image is converted to grayscale level, so each pixel in the RGB color is converted using Eq. 3.1, after that, the grayscale face image is fed to a cascade object detector which uses the Viola-Jones [42] algorithm to detect the face bounding box (see Figure 3.2).

$$Y = 0.299 * R + 0.587 * G + 0.114 * B \tag{3.1}$$

Figure 3.2: Face detection.

Secondly, to estimate the eyes position for each face image, we used Dlib [28] to detect the face landmarks noted by ($I1...I68$) (see Figure 3.3). In our case we only used four of these points, which represents the feature points of each eye referenced by $I37, I40$ for

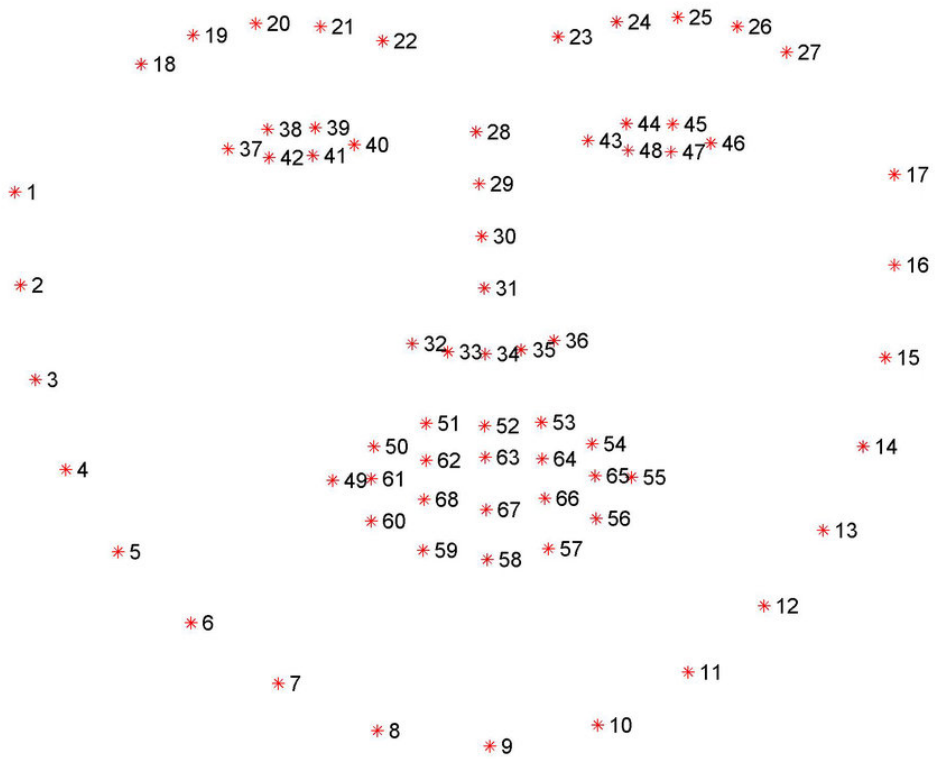


Figure 3.3: Face landmarks ($I1...I68$)

the left eye and $I43, I46$ for the right eye. Afterwards, these points are used to rectify the face pose.

In the face pose correction step, based on these four points for the left and right eyes, we calculated the centre of each using Eq. 3.2. Then we use Eq. 3.3 to calculate the transformation angle α, which will be used to rectify the face pose by clockwise rotation around the centre $C(x,y)$ of the image.

$$R_c(x,y) = \frac{I37 + I40}{2} \quad , \quad L_c(x,y) = \frac{I43 + I46}{2} \tag{3.2}$$

$$\alpha = tan^{-1}\left(\frac{R_c(y) \quad - \quad L_c(y)}{R_c(x) \quad - \quad L_c(x)}\right) \tag{3.3}$$

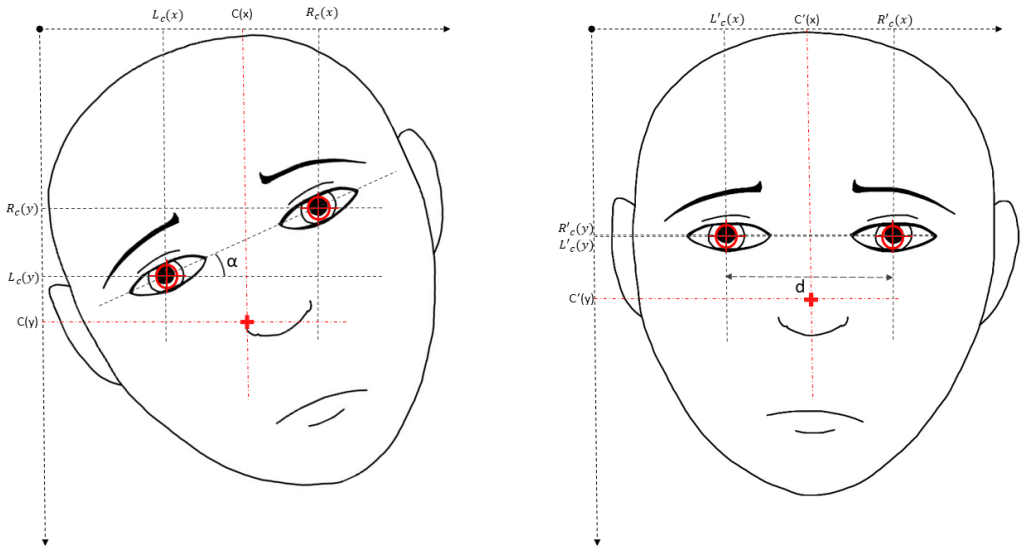
The new centre coordinates for the left and right eyes are calculated using the Eq. 3.4, Eq. 3.5, Eq. 3.6, Eq. 3.7, where $R'_c(x,y), L'_c(x,y)$ are the new right and left eyes centres respectively, and $C'(x,y)$ is the centre of the rotated image. Figure 3.4 illustrate this process.

$$R'_c(x) = C'(x) + (R_c(x) - C(x)).cos(\alpha) - (R_c(y) - C(y)).sin(\alpha)) \tag{3.4}$$

$$R'_c(y) = C'(y) + (R_c(x) - C(x)).sin(\alpha) - (R_c(y) - C(y)).cos(\alpha)) \tag{3.5}$$

$$L'_c(x) = C'(x) + (L_c(x) - C(x)).cos(\alpha) - (L_c(y) - C(y)).sin(\alpha)) \tag{3.6}$$

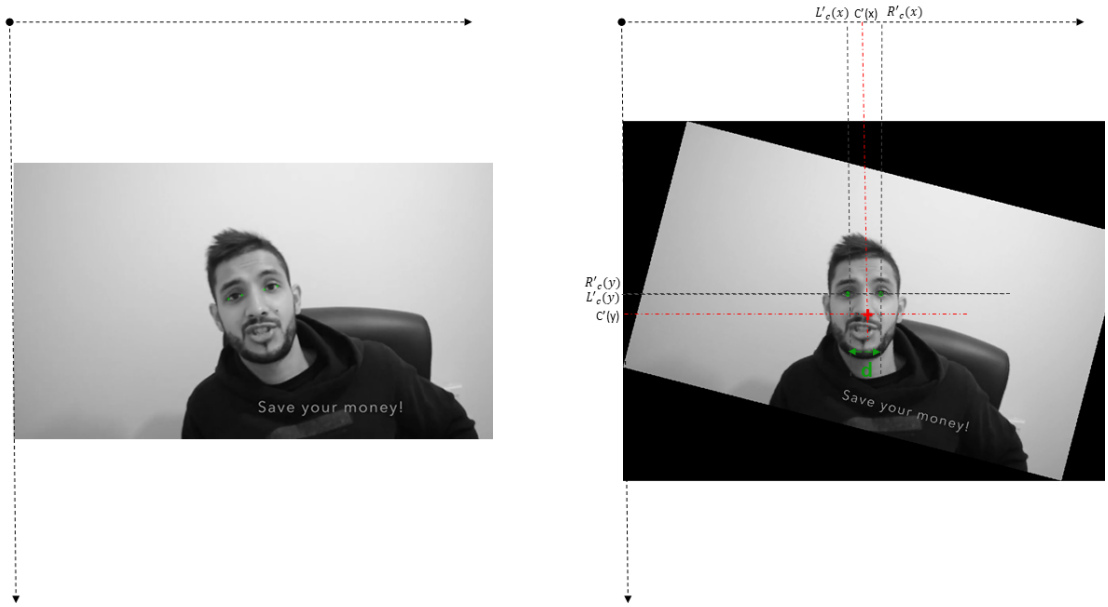
$$L'_c(y) = C'(y) + (L_c(x) - C(x)).sin(\alpha) - (L_c(y) - C(y)).cos(\alpha)) \tag{3.7}$$



(a) Original face (Before). **(b)** Corrected face (After).

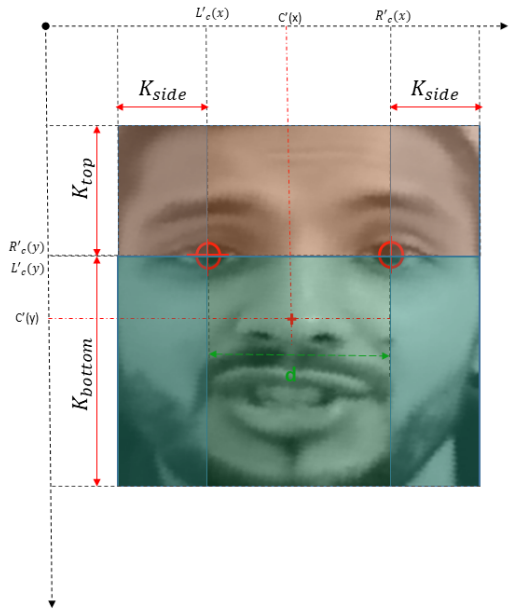
Figure 3.4: Eyes position transformation process.

The face region of interest is selected by measuring the absolute distance $d = |R'_c(x) - L'_c(x)|$ between the new centres of left and right eyes. After that we apply three factor on this distance to calculate the correct region of the face, this factors $k_{side} = 0.5$, $k_{top} = 1.0$, $k_{bottom} = 1.75$ has been used in previous works [14],[43]. Figure 3.5 shows the resulted cropped face of this process, which will be used in face extraction phase.



(a) Original face.

(b) Corrected face (Before).



(c) cropped face (After).

Figure 3.5: Face region selection process.

3.3 Feature representation

In computer vision, feature representation is an important step. It is connected to feature extraction by mapping raw image pixels into discriminant feature space. Effective feature representation can significantly impact the performance of machine learning methods.

3.3.1 Multi-block (MB)

Multi-block (MB) face representation is used in many biometrics tasks to represent face texture such as [44], [45]. It takes a face image and divides it by a regular grid of a fixed size window to obtain $b = l^2$ block, where l is the desired MB level. Figure 3.6 illustrates this MB representation.

Figure 3.6: (MB) face representation ($l = 3$).

3.3.2 Multi-level (ML)

The ML face representation used in various application such as [46], [47], [1]. It is a special pyramid that is formed by sorting a sequence of MB representations. The ML face representation level is obtained from the level of $1,..,l$ MB representation. So $b = \sum_{i=1}^{l} i^2$ block are obtained. This is depicted in Figure 3.7.

3.3.3 Pyramid multi-level (PML)

The Pyramid multi-level (PML) representation, first proposed by [14] allows for the extraction of multi-level multi-scale features from distinct divisions of each pyramid level, resulting in each face being divided into $b = \sum_{i=1}^{l} i^2 = l(l+1)(2l+1)/6$ regions. This is depicted in Figure 3.8.

Figure 3.7: (ML) face representation ($l = 3$).

Figure 3.8: Pyramid multi-level (PML) representation ($l = 3$).

3.4 Feature extraction

In this work, we compared multiple kinds of hand-crafted features and deep features to determine which one is more suitable for this task, where the details of each kind of features are mentioned below.

3.4.1 Hand-crafted features

Hand-crafted descriptors are either simple [48], [49], [50] or sophisticated algorithms [51], [16], [52] that extract the features through the information in the image itself. In our study, we used five kinds of hand-crafted descriptors.

3.4.1.1 Local Binary Patterns (LBP)

LBP is an efficient texture descriptor that seeks to summarize the local texture of an image in order to discriminate different images. The original LBP descriptor was first introduced by (Ojala et al. 1996) [53]. Except for the pixels in the border, it defines labels for each pixel in an image. The pixel value is used as a threshold for $3x3$ neighborhoods; the precisely negative values are encoded with zeros, while the others are encoded with ones. Concatenating all of these binary codes in a clockwise direction, starting from the

top-left one, yields a binary number. Figure 3.9 shows the overall steps of this process. As the neighborhood is made up of 8 pixels, there are a total of $2^8 = 256$ different labels. The obtained decimal numbers are referred to as LBP codes.

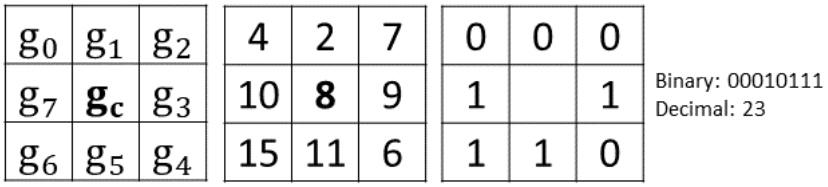


Figure 3.9: Basic LBP operator

LBP variants were developed to address the issue of capturing large-scale texture. Extending the neighborhood to P sample points symmetrically organized on a circle of radius R as depicted on Figure 3.10. The encoding $LBP_{P,R}$ of a point (x_c, y_c) is then calculated as follows:

$$LBP_{P,R} = \sum_{p=0}^{p-1} S.(g_p - g_c).2^p \tag{3.8}$$

where, $S(x)$ is defined by Equation 3.9 and g_p stands for the $p-th$ neighbor point's value and g_c represents the center point (x_c, y_c).

$$S(x) = \begin{cases} 0 & : x < 0 \\ 1 & : x \geq 0 \end{cases} \tag{3.9}$$

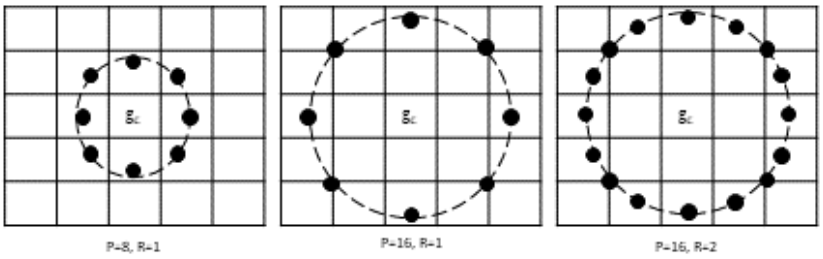


Figure 3.10: Examples of the extended LBP's with different (P, R)

Another LBP extension [54], [55] which aims to reduce the size of the original, so-called uniform patterns. It measures the transition from 0 to 1 or vice versa when the corresponding bit code is considered circular, then it is called uniform if the transactions

are less than two. for example, 00010000 (2 transitions) and 01111100 (2 transitions) are both uniforms whereas 01010000 (4 transitions) are not.

Other variants, such as Multi-Block Local Binary Pattern (MB-LBP) [56], Multi-quantized local binary patterns (MQLBP) [57], Median Local Binary Pattern (MBP) [58], and Divided Local Binary Pattern (DLBP) [59] have been developed to improve the performance of various applications by trying to capture more local features. In contrast, Doubled Local Binary Pattern (d-LBP), Reduced Divided Local Binary Pattern (RedDLBP) and Median Block Local Binary Pattern (MedBLBP) [60] use a group of pixels instead of a single pixel to reduce the noise effect and focus on capturing more global features by extending the neighborhood, enlarging the radius R, or adding another radius.

3.4.1.2 Local Phase Quantization (LPQ)

Local Phase Quantization LPQ [21] is one of the most successful LBP variants. LPQ is built on short-term Fourier transform and make use of the local phase information extracted by the 2D Discrete Fourier Transform (DFT) calculated over a rectangular $M-by-M$ local neighborhood Nx. The local frequency is determined using the short-term Fourier transform for each pixel $x = (x1, x2)^T$ from the input image $f(x)$ defined by:

$$F(u,x) = \sum_{y \in N_x} f(x-y) e^{-j2\pi u^T y} = w_u^T f_x \tag{3.10}$$

where w_u is the basis vector of the 2-D DFT at frequency u, and f_x is another vector containing all M^2 image samples from N_x. The transform Eq. 3.10 is efficiently evaluated for all image positions $x \in x_1, x_2, ..., x_N$ using simply 1-D convolutions for the rows and columns successively.

The local Fourier coefficients are computed at four frequency points $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$, and $u_4 = [a, -a]^T$, where a is selected as sufficiently small scalar. So, each pixel position results a vector.

$$F_x = [F(u_1,x), \quad F(u_2,x), \quad F(u_3,x), \quad F(u_4,x)] \tag{3.11}$$

Equation 3.12 describes how the signs of the real and imaginary parts of the Fourier coefficients components of $F(x)$ given by $G(x) = [Re\ \{F(x)\},\ Im\ \{F(x)\}]$ are employed to

generate LPQ binary codes which represent the phase information. Where Re and Im are the real and the imaginary parts of this Fourier transform.

$$q_j = \begin{cases} 1 & if \quad g_j \geq 0 \\ 0 & otherwise. \end{cases} \tag{3.12}$$

where g_j is the $j-th$ component of the vector $G(x)$. The resulting pattern of binary code, represented by the eight binary coefficients q_j, will be mapped to a decimal value in the range $0-255$ by $f_{LPQ(x)} = \sum_{j=1}^{8}(q_j 2^{(j-1)})$. Figure 3.11 illustrate this process.

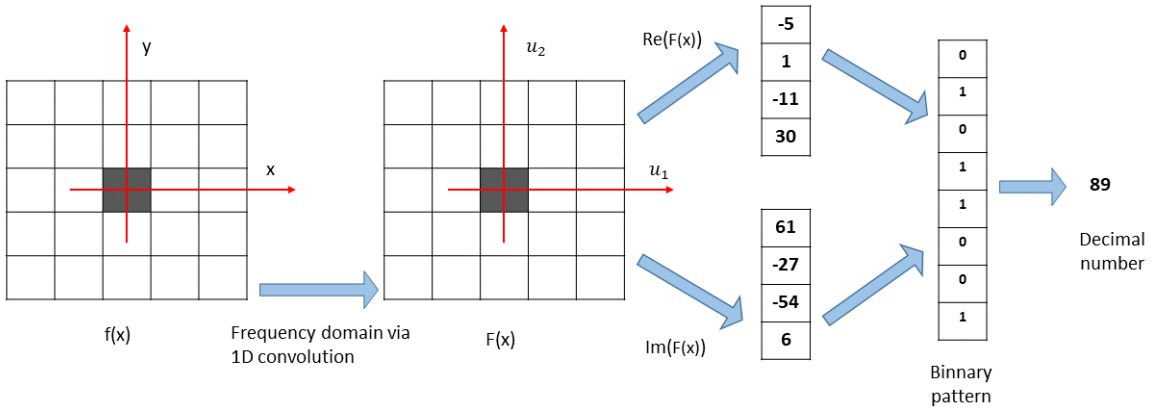


Figure 3.11: Example of LPQ calculation

3.4.1.3 Binarized Statistical Image Features (BSIF)

Binarized Statistical Image Features (BSIF) [25] is another variant of LBP that performs well in many computer vision tasks. Instead of hand-crafted filters as in LBP and LPQ, BSIF employs filters of a constant size that are dynamically learned from a minimal training set of natural images using Independent Component Analysis (ICA) by optimizing the statistical independence of the filter responses.

The filter response s_i is calculated using an image patch X of size $l \cdot l$ pixels and a linear filter W_i of the similar size by:

$$s_i = \sum_{i,v} W_i(i,v)X(u,v) = w_i^T x, \tag{3.13}$$

where vectors w and x contain the pixels of W_i and X. Equation 3.14 demonstrates the binarization of the s_i response for $i = 1,...,m$.

$$b_i = \begin{cases} 1 & if\ s_i > 0 \\ 0 & otherwise. \end{cases} \tag{3.14}$$

Therefore, b_i includes m binary digits, whereas the BSIF code is generated by $f_{BSIF}(x) = \sum_{i=1}^{m}(b_i \cdot 2^{(m-1)})$. Hence, the BSIF feature is represented by a histogram of $(0 : 2^{(m)} - 1)$ codes. In neighboring pixels, the code value of the pixel is interpreted as a local descriptor of the image intensity pattern. Moreover, histograms of pixel code values enable the characterization of texture features within sub-regions of an image.

3.4.1.4 Local Directional Pattern (LDP)

Local Directional Pattern (LDP) [61] is another LBP variant that analyzes different magnitude of edge responses in different directions of a particular pixel. LDP encodes each pixel to eight-bit binary code. This pattern is calculated by comparing the relative edge response value of a pixel in eight different directions. LDP uses Kirsch masks in eight different orientations $M_i, i = \{0,...,7\}$ centered on its position. These masks are shown in Figure 3.12.

$$\begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} M_0(\uparrow) \quad \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} M_1(\nwarrow) \quad \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} M_2(\leftarrow) \quad \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} M_3(\swarrow)$$

$$\begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} M_4(\downarrow) \quad \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} M_5(\searrow) \quad \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} M_6(\rightarrow) \quad \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix} M_7(\nearrow)$$

Figure 3.12: Kirsch edge response masks in eight directions.

By applying eight masks, $m_i, i = \{0,\ldots,7\}$ represents the edge significance in their respective direction. Since the presence of corner or edge show high response values in particular directions and LDP aims to promote them first by sorting them to show the k most prominent directions, a k value must be given. Then, the top k values of m_i are set to 1 and the other $(8-k)$ bits of the 8-bit LDP pattern are set to 0. Figure 3.13 shows an exemplary LDP code with $k = 3$. The formulas below are used to determine the LDP code for each pixel.

$$LDP_k = \sum_{i=0}^{7} b_i(m_i - m_k) \cdot 2^i \tag{3.15}$$

$$b_i(a) = \begin{cases} 1 & if\ a \geq 0 \\ 0 & otherwise. \end{cases} \tag{3.16}$$

where, m_k is the $k-th$ most significant directional response.

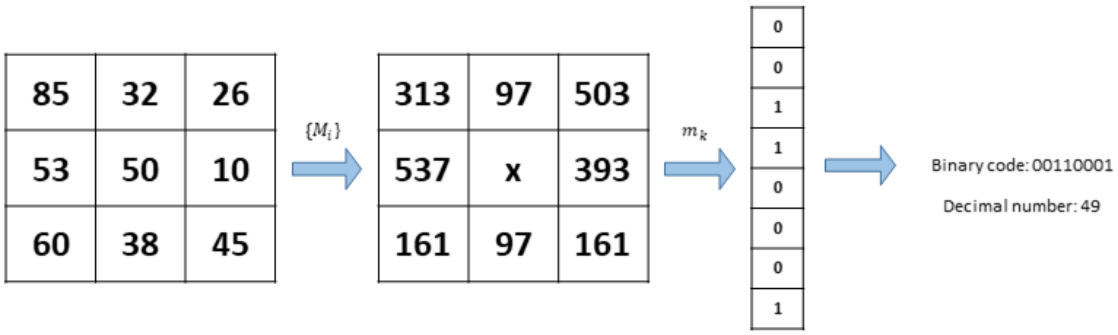


Figure 3.13: Example of LDP calculation

3.4.1.5 Co-Variance Operator descriptor (COV)

The **Covariance descriptor (COV)** (see Figure 3.14) was proposed in [62] as a region descriptor that could be used in object detection and texture classification problems. It takes advantage of the information provided by covariance matrices, that provides a natural way of fusing multiple features while keeping a low-dimensionality space due to its symmetry. Covariance matrices have only $d \cdot (d+1)/2$ different values.

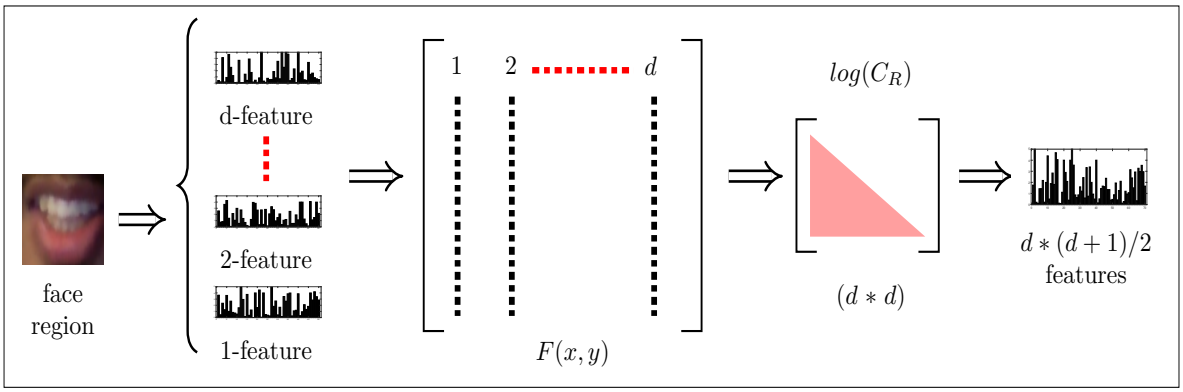


Figure 3.14: Covariance descriptor (COV).

The COV descriptor is computed as follows: Let I denote $M \cdot N$ intensity image, and F be the $M \cdot N \cdot d$ dimensional feature image extracted from I, which contain a collection

of image features such as horizontal coordinate, vertical coordinate, intensity, image gradient or any image feature array. This dimensional feature image can be written as $F(x,y) = \phi(I,x,y)$, where ϕ is the mapping function of each feature image in this collection. For a given region $R \subset F$ containing s points, let $\{v_i\}_{i=1..s}$ be the d-dimensional feature points inside R. The region R is described by $d \cdot d$ covariance matrix of the feature points (See Eq. (3.17)). This region R can be characterized by $log(C_R)$, where $log(C_R)$ is the matrix logarithm of the square matrix C_R.

$$C_R = \frac{1}{s-1}\sum_{i=1}^{s}(v_i - \mu)(v_i - \mu)^T, \tag{3.17}$$

where μ is the mean of the points.

3.4.2 Deep features

In the recent years, Deep learning methods successfully achieved state-of-the-art performance in many computer vision tasks [63], [64], [65]. In particular, the CNNs are used as end-to-end training [66], [67], [68] or as feature extractor [69], [70], [71] where the deep features are generally extracted from one of the last layers of a convolutional neural network (CNN). In this study, we used five pretrained deep learning architectures.

3.4.2.1 VGG16 architecture

The VGG16 model, which supports 16 learnable layers is a CNN model developed and introduced by [72] of the Visual Geometry Group Lab at Oxford University in 2014. VGG16 won the 2014 ILSVRC challenge [73] and achieves 92.7% top-5 test accuracy on the ImageNet dataset which contains 14 million images belonging to nearly 1000 classes. Figure 3.15 illustrates this architecture. VGG16 takes in an RGB image with an input size of $224 \cdot 224$ and it consists of 13 convolutional layers, five Max-Pooling layers, three fully connected layers, and a SoftMax layer for the output. All the hidden layers use ReLU as its activation function. VGG16 is a massive network with around 138 million parameters. In this work, we used the VGG16 architecture trained on VGGFace dataset [29], then we extract the deep features from the $FC7$ linear layer which produce 4096 features.

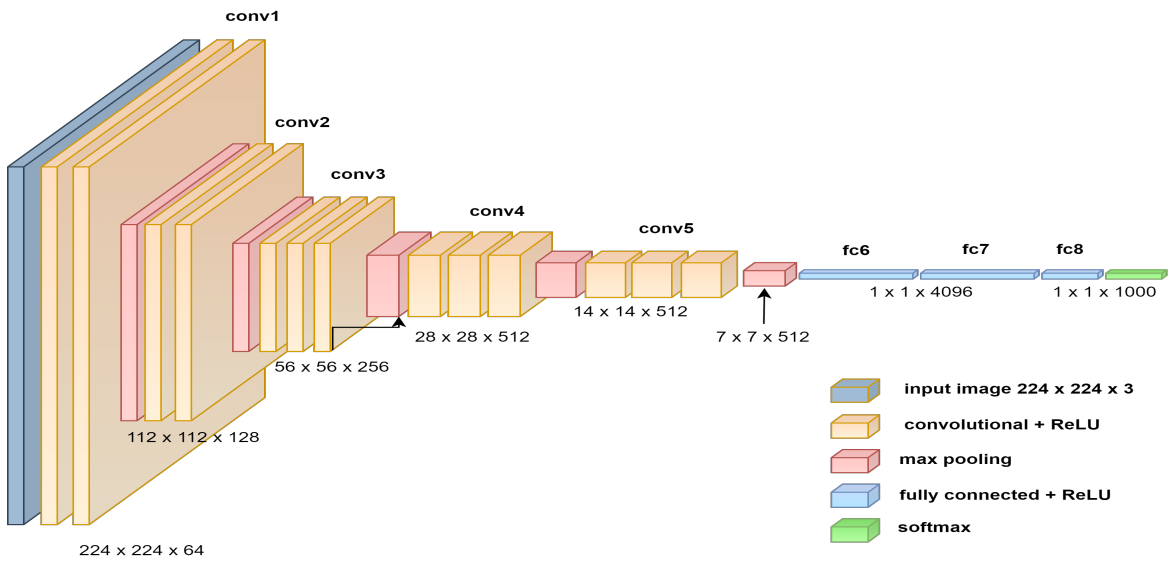


Figure 3.15: VGG16 architecture.

3.4.2.2 ResNet-50

ResNet-50 is a convolutional neural network consisting of 50 trainable layers. A residual neural network (ResNet) first introduced in [74] is a CNN network made up of residual blocks. These residual blocks solved the CNN depth restriction problem, which suffers from vanishing gradients and accuracy getting saturated and degrading rapidly. ResNet introduced a new concept of shortcut connections, shortcut connections (also referred to as skip connections) allow the network to learn the residual mapping rather than the desired unknown mapping between the inputs and outputs. In this work, ResNet-50 is trained on VGGFace2 dataset [75] and generates 2048 features from the global average pooling layer. Figure 3.16 shows the ResNet-50 architecture.

3.4.2.3 SE-ResNet-50

The SE-ResNet-50 [76] architecture is based on ResNet-50. However, instead of residual blocks, it employs squeeze-and-excitation blocks (see Figure 3.17) to allow the network to perform adaptive channel-wise feature recalibration by explicitly modeling channel interdependencies. In this work, SE-ResNet-50 is also trained on the VGGFace2 dataset and provides the same number of features as ResNet-50.

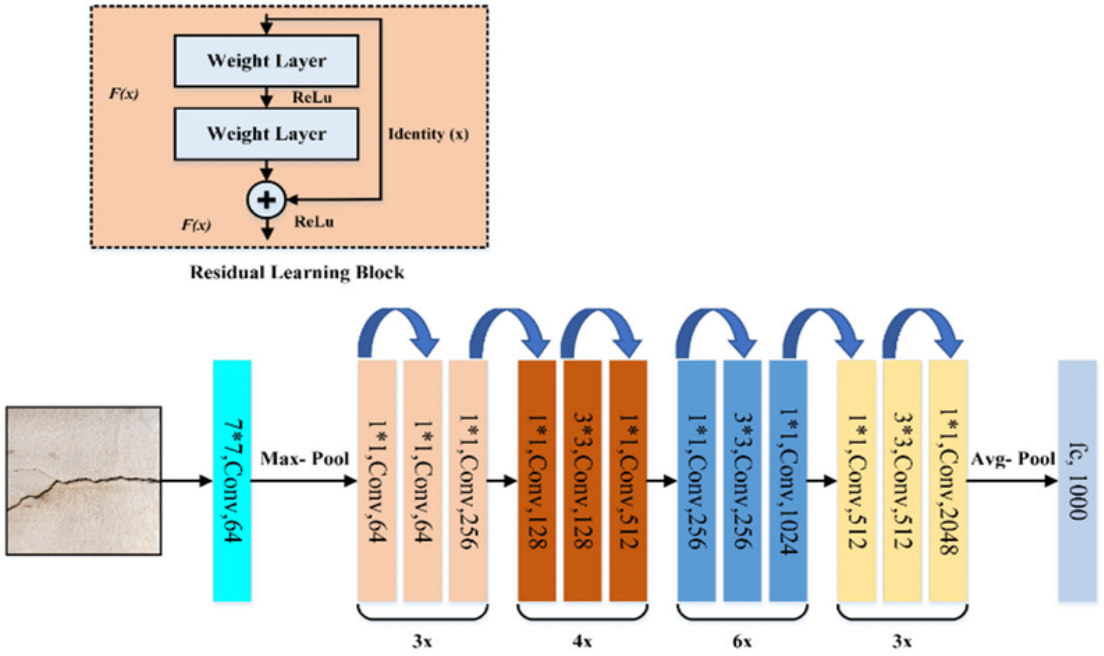


Figure 3.16: ResNet-50 architecture.

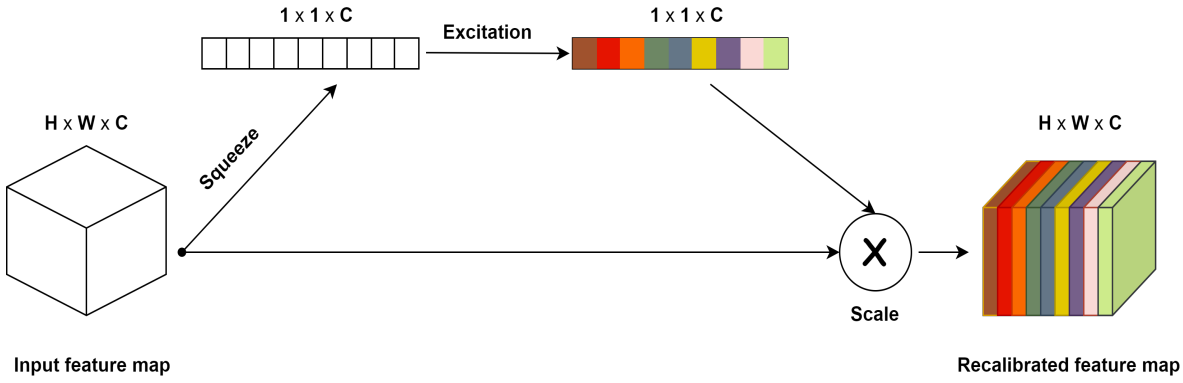


Figure 3.17: Squeeze-and-Excitation block.

3.4.2.4 ArcFace architecture

Additive Angular Margin Loss (ArcFace) architecture [77] is based on SE-ResNet-50 and employees Additive Angular Margin Loss to obtain highly discriminative features. It explores the $BN - Dropout - FC - BN$ structure after the last convolutional layer to obtain the final $512 - D$ embedding feature. In this work, we used the ArcFace model pre-trained on the MS-Celeb-1M [78] dataset.

3.4.2.5 MobileFaceNet

The MobileFaceNet model [79] is a small convolutional neural network with less than 1 million parameters and a $4.0MB$ size. Its architecture is based on MobileNetV2 [80] and employs the ArcFace loss function. It is trained on the refined MS-Celeb-1M and achieves significantly improved accuracy along with a speedup of more than two times over MobileNetV2. The number of features it generates is the same as ArcFace.

3.5 Video descriptor computation

After preprocessing each frame in the video sequence and getting their aligned faces, as shown in 3.1, we used both hand-crafted and deep learning methods in our work.

In the hand-crafted approaches, to obtain the spatio-temporal feature vector which represents the whole video for the desired descriptor, we start by extracting the feature representation to obtain a $b = l(l+1)(2l+1)/6$ block over each face image. Then, we fed each region of the face to the desired hand-crafted descriptor (DESC) to obtain b feature vectors. Concatenating these features gives an intermediate feature vector which represents the current face (see Figure 3.18). Then, we merge these intermediate feature vectors in order to obtain the final spatio-temporal feature vector which represent the information of a whole video sequence.

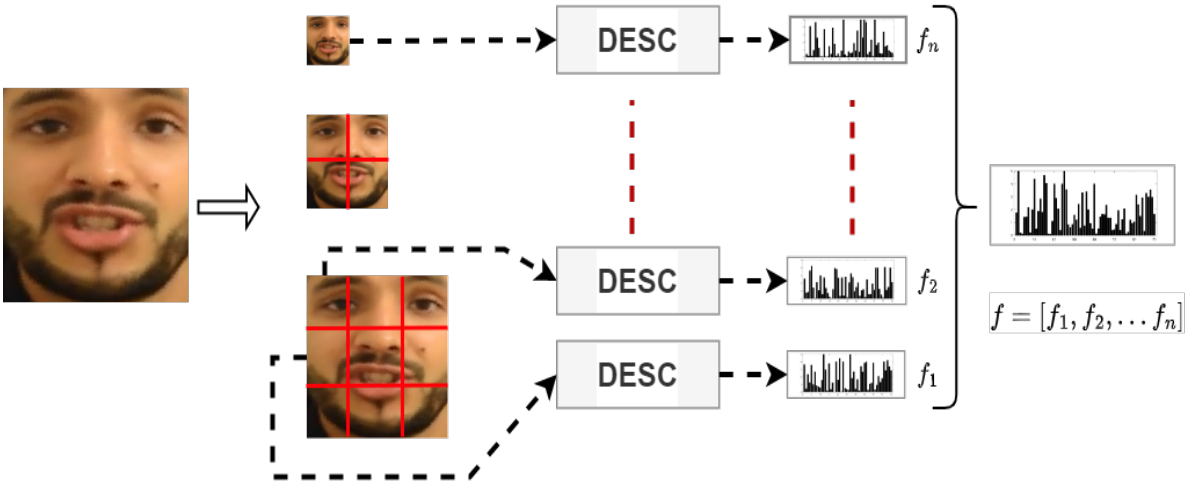


Figure 3.18: Feature extraction using the desired hand-crafted descriptor (DESC).

For deep learning-based approaches, intermediate spatial-temporal feature vectors are obtained by feeding aligned faces over a video sequence into the desired network and merged together to produce the final spatial-temporal feature vector.

As part of our work, we combine the video sequence information of hand-crafted and deep learning approaches using 12 statistical descriptors These statistical descriptors are: Mean, Variance, Skewness, Root Mean Square (RMS), Peak min, Peak max, Crest factor min, Crest factor max, Crest pulse min, Crest pulse max, Kurtosis, and the Shape factor. Utilizing these descriptors is designed to find the most efficient means of encoding spatial and temporal information about face features. A detailed comparison of these statistical descriptors is presented in the following chapter.

MEAN: The mean, or average, is likely the most frequently employed statistic for describing central tendency. To calculate the mean, simply add all the values and divide by the total number of values. This procedure is depicted by Equation 3.18.

$$Mean(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.18}$$

VARIANCE: The variance is the expected squared deviation of a random variable from the mean of its population. It quantifies the deviation of a set of numbers from their mean value and is given by Equation 3.19.

$$Var(x) = \frac{1}{N}\sum_{i=1}^{N} (x_i - Mean(x))^2 \tag{3.19}$$

SKEWNESS: The degree of the data's disproportion from the sample mean is expressed by the skewness. The data has a greater distribution to the left of the mean if the skewness is negative. Data with a positive skewness tends to lean to the right. If a distribution is normally distributed or entirely symmetric, then its skewness is zero. The skewness is defined by Equation 3.20.

$$Skewness(x) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - Mean(x))^3}{Std(x)^3} \tag{3.20}$$

where, the Std is the standard deviation and it is defined by Equation 3.21.

$$Std(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - Mean(x))^2} \tag{3.21}$$

ROOT MEAN SQUARE (RMS): The RMS of a set of values x_i is calculated by taking the square root of the mean square (see Equation 3.22).

$$RMS(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{3.22}$$

PEAK MIN & PEAK MAX: The peak value of the input data is defined by the Equation 3.23 for the peak min, and 3.24 for the peak max.

$$Peaks\_Min(x) = \bigvee_{i=1}^{N} min(peaks(x_i)) \tag{3.23}$$

$$Peaks\_Max(x) = \bigvee_{i=1}^{N} max(peaks(x_i)) \tag{3.24}$$

where the min, max functions return the minimum and the maximum value of the data respectively. The $peaks$ function returns a vector with the local maxima (peaks) of the input feature sample (see Figure 3.19).

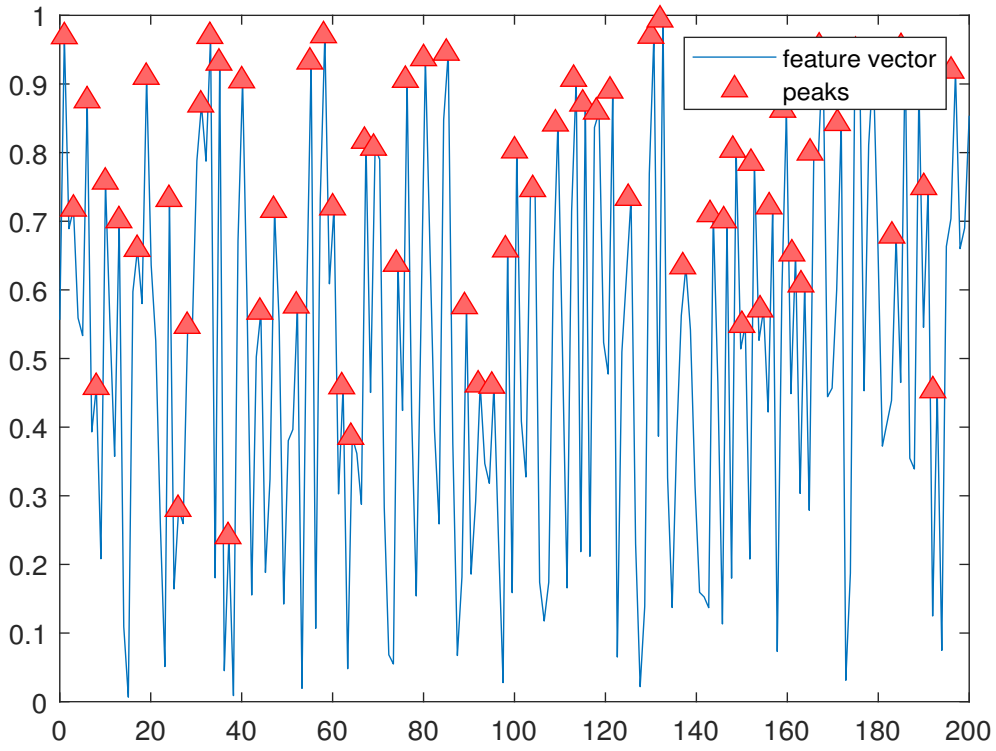


Figure 3.19: The local maxima (peaks) of the input feature vector.

CREST FACTOR MIN & CREST FACTOR MAX: Crest factor indicates how extreme the peaks are in an input data. It is the ratio of the peak value to its RMS. Equation 3.25 define the crest factor min, and Equation 3.26 defines the crest factor max.

$$Crest\_factor\_min(x) = \frac{Peaks\_Min(x)}{RMS(x)} \tag{3.25}$$

$$Crest\_factor\_max(x) = \frac{Peaks\_Max(x)}{RMS(x)} \tag{3.26}$$

CREST PULSE MIN & CREST PULSE MAX: The crest pulse also indicates the strength of the input data's peaks. It is the ratio of the maximum value to the mean. (see Equations 3.27, and 3.28)

$$Crest\_pulse\_max(x) = \frac{Peaks\_Max(x)}{Mean(x)} \tag{3.27}$$

$$Crest\_pulse\_max(x) = \frac{Peaks\_Max(x)}{Mean(x)} \tag{3.28}$$

KURTOSIS: Kurtosis measures the likelihood of an outlier occurring in a distribution. The kurtosis of the normal distribution is 3. Distributions that are more outlier-prone than the normal distribution have a kurtosis greater than 3; distributions that are less outlier-prone have a kurtosis less than 3. The kurtosis of a distribution is defined by Equation 3.29.

$$Kurtosis(x) = Kurtosis(x) == \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - Mean(x))^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - Mean(x))^2\right)^2} \tag{3.29}$$

SHAPE FACTOR: The shape factor is the proportion of the RMS value to its mean, and it is defined by the Equation 3.30.

$$Shape\_factor(x) = \frac{RMS(x)}{Mean(x)} \tag{3.30}$$

In this work, we apply $L2$ feature normalization over the spatio-temporal feature vector in order to normalize the distribution of data, and keep the overall error small. $L2$

is the square root of the sum of the squared vector values. It is also called the Euclidean norm, and it is given by Equation 3.31.

$$L^2 norm(x) = \sqrt{\sum_{i=1}^{N} x_i^2} \tag{3.31}$$

3.6 Feature selection

In computer vision, feature selection plays a crucial role. As a result of feature selection, redundant or irrelevant data are reduced in dimensionality without losing any of their information value. This makes the time it takes to train the model shorter while keeping or improving its ability to solve classification or regression problems.

We aimed to identify the best subset of features. Three future selection ranking methods are used to determine which one is best for our case. The ranked features were used in the selection of the most relevant features for each personality trait separately. Thus, for each personality trait, we select the best feature subset based on these ranked feature weights. The selected subset of features is determined by taking all features with high weights until reaching the mode of the histogram of the weights.

3.6.1 Relief Algorithm

The Relief algorithm [81] figures out the values of attributes based on how well they separate instances that are close to each other. It punishes predictors that assign different values to neighbors with similar response values and award predictors that assign different values to neighbors with distinct response values.

Relief-based feature selection methods have been improved so that they work better when there is noise, can be used for multi-class problems, can be used for regression problems [82], and can handle insufficient data better.

The Relief algorithm start by initializing all predictor weights W_j to zero. The algorithm then repeatedly picks an arbitrary observation x_r. Then, for each group of features, it identifies the k-nearest observations to x_r and updates all the weights for the predictors F_j for each nearest neighbor x_q by using Equation 3.32, when x_r and x_q have similar response values, or by using Equation 3.33 when they have distinct response values.

$$\mathrm{W}_j^i = \mathrm{W}_j^{i-1} - \frac{\triangle_j \left(x_r - x_q\right)}{m} d_{rq} \tag{3.32}$$

$$\mathrm{W}_j^i = \mathrm{W}_j^{i-1} + \frac{P_{y_q}}{1 - P_{y_r}} \frac{\triangle_j \left(x_r - x_q\right)}{m} d_{rq} \tag{3.33}$$

where:

- W_j^i is the weight of the predictor F_j at the i_{th} iteration step.
- P_{y_r} is the prior probability of the response value to which x_r belongs, and P_{y_q} is the prior probability of the response value to which x_q belongs.
- m is the number of iterations specified by the updates.
- $\triangle_j \left(x_r - x_q\right)$ is the difference in the value of the predictor F_j between observations x_r and x_q, it is defined by Equation 3.34 for discrete F_j, and by Equation 3.35 for continuous F_j, where x_{rj} is the value of the j_{th} predictor for observation x_r, and x_{qj} is the value of the j_{th} predictor for observation x_q.
- d_{rq} is a distance function given by Equation 3.36, d_{rq} is subjected to the scaling defined by Equation 3.37, where the $rank(r,q)$ is the position of the q_{th} observation among the nearest neighbors of the r_{th} observation, sorted by distance, k is the number of nearest neighbors, and sigma (σ) is the scaling factor.

$$\triangle_j \left(x_r - x_q\right) = \begin{cases} 0, & x_{rj} = x_{qj} \\ 1, & \text{otherwise} \end{cases} \tag{3.34}$$

$$\triangle_j \left(x_r - x_q\right) = \frac{\left|x_{rj} - x_{qj}\right|}{max(F_j) - min(F_j)} \tag{3.35}$$

$$d_{rq} = \frac{\tilde{d_{rq}}}{\sum_{l=1}^{k} \tilde{d_{rl}}} \tag{3.36}$$

$$\tilde{d_{rq}} = e^{-(rank(r,q)/\sigma)^2} \tag{3.37}$$

3.6.2 Minimum redundancy maximum relevance

The Minimum redundancy maximum relevance (MRMR) algorithm [83] identifies the best set of features that are mutually and maximally divergent to represent the response variable effectively. It uses the mutual information of variables to measure redundancy and relevance. This makes a feature set less redundant and more relevant to the response variable.

The objective of the MRMR algorithm is to discover an ideal subset S of features that maximizes V_S, the relevance of S with respect to a response variable y, and minimizes W_S, the redundancy of S, where V_S and W_S are defined using I as follows:

$$V_S = \frac{1}{|S|} \sum_{x \in S} I(x, y) \tag{3.38}$$

$$W_S = \frac{1}{|S|^2} \sum_{x,z \in S} I(x, z) \tag{3.39}$$

where $|S|$ represent the number of features contained in S. Finding an ideal subset S requires examining all $2^{|\Omega|}$ combinations, where Ω is the full set of features. Using the mutual information quotient (MIQ) value (see Equation 3.40), the MRMR algorithm ranks features through the forward addition scheme, which needs $O(|\Omega||S|)$ calculations.

$$MIQ_x = \frac{V_x}{W_x} \tag{3.40}$$

where V_x is the relevance of features and is given by Equation 3.41, while W_x is redundancy of a feature and given by Equation 3.42.

$$V_x = I(x, y) \tag{3.41}$$

$$W_x = \frac{1}{|S|} \sum_{z \in S} I(x, z) \tag{3.42}$$

The MRMR algorithm scores all Omega features in order of importance and returns an index of ranked features. Using an algorithm, it evaluates the importance of a feature and generates a score. A high score indicates that the corresponding prediction is important. A low score indicates a lack of confidence in the feature selection. MRMR algorithm ranks features as follows:

1. Select the feature with the largest relevance, $\underset{x\in\Omega}{max} V_x$ Add the selected feature to an empty set S.

2. Find the features with nonzero relevance and zero redundancy in the complement of S, S^c

   - If S^c does not include a feature with nonzero relevance and zero redundancy, go to step 4.
   - Otherwise, select the feature with the largest relevance, $\underset{x\in S^c, W_x=0}{max} V_x$ Add the selected feature to the set S.

3. Repeat Step 2 until the redundancy is not zero for all features in S^c.

4. Select the feature that has the largest MIQ value with nonzero relevance and nonzero redundancy in S^c, and add the selected feature to the set S as follow:

$$\underset{x\in S^c}{max} MIQ_x = max\ \frac{I(x,y)}{\frac{1}{|S|}\sum_{z\in S} I(x,z)}) \tag{3.43}$$

5. Repeat Step 4 until the relevance is zero for all features in S^c.

6. Add the features with zero relevance to S in random order.

7. Skip any step if the MRMR algorithm cannot find a feature that satisfies the conditions described in other steps.

3.6.3 Neighborhood component analysis

Neighborhood component analysis (NCA) [84] is a non-parametric learning method for estimating the feature weights. NCA sorts the features according to their relevance by performing feature ranking with regularization to learn feature weights, minimizing an objective function that measures the average leave-one-out classification or regression loss over the training data .

Considering a training set, given n observations as follow:

$$S = \{(x_i, y_i), i = 1, 2, \ldots, n\} \tag{3.44}$$

Where $x, y \in \mathbb{R}^p$ are the input and output variables. The aim is to learn an objective function $f : \mathbb{R}^p$ that predict the output y_i for the given the training set S. Considering a

randomised model $Ref(x)$ that randomly picks a feature vector x_j from S with its corresponding output value y_j. In NCA, the reference point is picked arbitrarily and any points in S has a chance of becoming the reference point. The probability $P\left(Ref(x) = x_j|S)\right)$ that point x_j is chosen from S as the reference point for x is greater if x_j is closer to x. This is determined by the distance function d_w which is represented by Equation 3.45.

$$d_w(x_i, x_j) = \sum_{r=1}^{p} \mathrm{w}_r^2 \left|x_{ir} - x_{jr}\right| \tag{3.45}$$

w_r are the feature weights. Assume that $P\left(Ref(x) = x_j|S)\right) \propto k\left(d_w\left(x, x_j\right)\right)$. k is some kernel or a similarity function that assumes large values when $d_w\left(x, x_j\right)$ is small. The sum of $P\left(Ref(x) = x_j|S)\right)$ for all j must be equal to 1. Therefore, $P\left(Ref(x) = x_j|S)\right)$ is given by Equation 3.46 [84].

$$P\left(Ref(x) = x_j|S)\right) = \frac{k\left(d_w\left(x, x_j\right)\right)}{\sum_{j=1}^{n} k\left(d_w\left(x, x_j\right)\right)} \tag{3.46}$$

In the same context, considering the leave-one-out application of this randomized regression model. Predicting the response for (x_i using the data in S^{-i}, the training set S excluding the point (x_i, x_j) The probability that point x_j is chosen as the reference point for x_i is given by Equation 3.47.

$$p_{ij} = P\left(Ref(x_i) = x_j|S^{-i})\right) = \frac{k\left(d_w\left(x_i, x_j\right)\right)}{\sum_{j=1, j\neq i}^{n} k\left(d_w\left(x_i, x_j\right)\right)} \tag{3.47}$$

Now, let $\widehat{y}_i$ be the response value the randomized regression model predicts and y_i be the actual response for (x_i. And let $l : \mathbb{R}^2 \to \mathbb{R}$ be a loss function that measures the disagreement between $(y_i, \widehat{y}_i)$. Then, the average value of $l(y_i, \widehat{y}_i)$ is given by Equation 3.48.

$$l_i = E\left(l(y_i, \widehat{y}_i)|S^{-i}\right) = \sum_{j=1, j\neq i}^{n} p_{ij} l(y_i, y_j) \tag{3.48}$$

NCA by default uses mean absolute deviation loss function ($l(y_i, \widehat{y}_i)$). It is defined by Equation 3.49, and to prevent the randomised regression model from being overfitted, a regularisation term λ is introduced to the final objective function. The objective function $f(w)$ for minimisation after adding a regularisation term λ is given by Equation 3.50.

$$l(y_i, \widehat{y}_i) = |y_i - \widehat{y}_i| \tag{3.49}$$

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} l_i + \lambda \sum_{r=1}^{p} \mathrm{w}_r^2 \tag{3.50}$$

3.7 Personality traits estimation

In order to estimate the scores of the Big-Five personality traits, we fed the five features subsets, which we got after feature selection to five Support Vector Regressors (SVRs) [22], one for each. These SVRs use hyper-parameter optimization to improve the final performance and standardize the features using their corresponding weighted means and weighted standard deviations.

Support Vector Regressor (SVR) [22] is a machine learning technique build based on support vector machine (SVM), unlike SVM which deal with classification problems, SVR is for regression based problems, SVR is considered a nonparametric technique because it relies on kernel functions. It embed data into a high dimensional feature space. Giving the training data $S = \{(x_i, y_i) \,|\, x_i \in \mathbb{R}^n, y_i \in \mathbb{R}; i = 1, ..., N\}$ where x_n is a multivariate set of N observations with observed response values y_n. The goal of SVR is to find a function $f(x)$ that deviates from y_n by a value no greater than ϵ for each training point x, and at the same time is as flat as possible, which means is to build a hyperplane, close to as many of the training points as possible by maximising the margin.

Suppose $f(x)$ takes the following form:

$$f(x) = w \cdot \phi(x) + b \tag{3.51}$$

Where $w \in \mathbb{R}^n$, and ϕ is the nonlinear function that transfer the training data x to a high dimensional feature space. We need to find w, and b that minimise the risk error and ensure that $f(x)$ it is flat as possible.

$$E(w, b) = C \sum_{i=1}^{N} L(y, f(x)) + \frac{1}{2} \|w\|^2 \tag{3.52}$$

Where $E(w, b)$ is a cost function, $L(y, f(x))$ is defined by Equation 3.53, the constant C is the box constraint, a positive numeric value that controls the penalty imposed on

observations that lie outside the ϵ margin and helps to prevent overfitting (regularization). This value determines the trade-off between the flatness of $f(x)$ and the amount up to which deviations larger than ϵ are tolerated.

$$L(y, f(x)) = \begin{cases} 0, & if\ |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon, & \text{otherwise} \end{cases} \tag{3.53}$$

The solution of the minimisation problem of Equation 3.52 is given by:

$$\begin{aligned} f(x) &= \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)(\phi(x_i) \cdot \phi(x)) + b \\ &= \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b \end{aligned} \tag{3.54}$$

$$\begin{aligned} &\sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0 \\ &\forall\, i : 0 \leq \alpha_i \leq C \\ &\forall\, i : 0 \leq \alpha_i^* \leq C \end{aligned} \tag{3.55}$$

where α_i and α_i^* are Lagrangian non-negative multipliers subject to the constraints shown in Equation 3.55, $K(x_i, x)$ is a kernel function. This kernel function could be a Linear function given by Equation 3.56, Gaussian function given by Equation 3.57 or Polynomial function given by Equation 3.58, where x_j and x_k are points in the high-dimensional feature space, $\bullet$ is dot product and q is the degree of the kernel in the set $\{2, 3, ...\}$.

$$k(x_j, x_k) = x_j \bullet x_k \tag{3.56}$$

$$k(x_j, x_k) = \exp(-\left\| x_j - x_k \right\|^2) \tag{3.57}$$

$$k(x_j, x_k) = \left(1 + x_j \bullet x_k\right)^q \tag{3.58}$$

The implementations of SVR varies [22]. Quadratic programming is one of the most frequent optimization techniques, although it can be computationally expensive to use, particularly, since the Gram matrix may be too large to hold in memory. Using a decomposition strategy helps speed up the computation and prevent memory exhaustion. Sequential minimal optimisation (SMO) [85], [86] is the most common method for solving SVR optimization issues.

3.8 Interview variable estimation

The estimated five scores are then considered as a new feature vector, which we fed to a Gaussian process regression (GPR) [26] scheme in order to estimate the interview score. This GPR also uses hyper-parameter optimization and standardize the features to improve the interview score. The reason behind choosing GPR instead of SVRs for the interview estimation is due to its high accuracy when it comes to very low dimensional data and this was found experimentally.

Gaussian process regression (GPR) [26] is a supervised machine learning algorithm that relies on few parameters to make predictions. GPR models are nonparametric, kernel-based probabilistic models. They work well on small datasets and having the ability to provide uncertainty measurements on the predictions. Giving the training data $S = \{(x_i, y_i) \,|\, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}; i = 1, ..., N\}$ where (x_i, y_i) drawn from an unknown distribution. The goal is to predict the value of a response variable y_{new} given the new input vector x_{new} and the training data S. A linear regression model is of the form:

$$f(x) = x^T \beta + \epsilon \tag{3.59}$$

Where $\epsilon \sim N(0, \sigma^2)$, the error variance σ^2 and the coefficients β are estimated from the data. A GPR model explains the response by introducing latent variables, $f(x_i), i = 1, 2, ..., n$ from a Gaussian process (GP), and explicit basis functions, h. The covariance function of the latent variables captures the smoothness of the response and basis functions project the inputs x into a p-dimensional feature space.

A GP is a set of arbitrarily variables, where any finite number of them have a joint Gaussian distribution. If $\{f(x), x \in R^d\}$ is a GP, then given n observations $x_1, x_2, ..., x_n$, the joint distribution of the arbitrarily variables $f(x_1), f(x_2), ..., f(x_n)$ is Gaussian. A GP is described by its mean function $m(x)$ and covariance function, $k(x, x')$. That is, if

$\{f(x), x \in R^d\}$ is a GP, then $E(f(x)) = m(x)$ and $Cov\left[f(x), f(x')\right] = E\left[\{f(x) - m(x)\}\{f(x') - m(x')\}\right] = k(x, x')$.

$$h(x)^T \beta + f(x) \tag{3.60}$$

Taking into account the model presented in Equation 3.60, where $f(x) \sim GP(0, k(x, x'))$, that is $f(x)$ are from a zero mean GP with covariance function, $k(x, x')$. $h(x)$ are a set of basis functions that transform the original feature vector x in R^d into a new feature vector $h(x)$ in R^p. β is a $p - by - 1$ vector of basis function coefficients. This model represents a GPR model. An instance of response y can be modeled as Equation 3.61.

$$P(y_i | f(x_i), x_i) \sim N\left(y_i | h(x_i)^T \beta + f(x_i), \sigma^2\right) \tag{3.61}$$

Hence, a GPR model is a probabilistic model. There is a latent variable $f(x_i)$ introduced for each observation x_i, which makes the GPR model nonparametric. In vector form, this model is equivalent to $P(y|f, X) \sim N\left(y|H\beta + f, \sigma^2 I\right)$, where X, y, H and f are defined by Equations 3.62,3.63,3.64,3.65 respectively.

$$X = \begin{pmatrix} \mathrm{x}_1^T \\ \mathrm{x}_2^T \\ \vdots \\ \mathrm{x}_n^T \end{pmatrix} \tag{3.62}$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{3.63}$$

$$H = \begin{pmatrix} h(\mathrm{x}_1^T) \\ h(\mathrm{x}_2^T) \\ \vdots \\ h(\mathrm{x}_n^T) \end{pmatrix} \tag{3.64}$$

$$f = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} \tag{3.65}$$

The joint distribution of latent variables $x_1), f(x_2), \cdots, f(x_n)$ in the GPR model is $P(f|X) \sim N(f|0, K(X,X))$ close to a linear regression model, where $K(X,X)$ is given by the format of the Equation 3.66. The covariance function $k(x,x')$ is usually parameterized by a set of kernel parameters or hyperparameters, θ. The optimization attempts to minimize the cross-validation loss (error) for GPR by varying this hyper-parameters. $k(x,x')$ is often written as $k(x,x'|\theta)$ to explicitly indicate the dependence on θ. The covariance function $k(x,x'|\theta)$ can be defined by various kernel functions, such as: Squared Exponential Kernel given by Equation 3.67, ARD Squared Exponential Kernel given by Equation 3.68, ARD Rational Quadratic Kernel given by Equation 3.69 or many others [26] [87].

$$K(X,X) = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_n) \\ \vdots & \vdots & \vdots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \cdots & k(x_n,x_n) \end{pmatrix} \tag{3.66}$$

$$K(x_i,x_j|\theta) = \sigma_f^2 \exp\left[-\frac{1}{2} \frac{(x_i - x_j)^T (x_i - x_j)}{\sigma_l^2} \right] \tag{3.67}$$

$$K(x_i,x_j|\theta) = \sigma_f^2 \exp\left[-\frac{1}{2} \sum_{m=1}^{d} \frac{(x_{im} - x_{jm})^2}{\sigma_m^2} \right] \tag{3.68}$$

$$K(x_i,x_j|\theta) = \sigma_f^2 \left(1 + \frac{1}{2\alpha} \sum_{m=1}^{d} \frac{(x_{im} - x_{jm})^2}{\sigma_m^2} \right)^{-\alpha} \tag{3.69}$$

Where, σ_l is the characteristic length scale, σ_f is the signal standard deviation, and α is a positive-valued scale-mixture parameter.

3.9 Conclusion

In this chapter, we described the overall architecture as well as the techniques used to estimate the five personality traits and the interview variable, which are either hand-crafted or based on deep learning. First, several feature representation methods are introduced, which serve as an input for the hand-crafted based techniques, which are: LBP, LPQ, BSIF, LDP, and COV. Then, in the approaches based on deep learning, we explained the used pretrained models, which are: VGG16, ResNet-50, SE-ResNet-50, ArcFace, and MobileFaceNet. Afterward, we give a brief description of the feature selection techniques used in order to identify the best subset of features. Finally, we detailed how the spatio-temporal features of a video sequence are computed over twelve statistical descriptors. In the next chapter, we will discuss the effects of these statistical descriptors, as well as the effects of PML level, and the feature selection techniques on the information learned by SVR and GPR, not to forget the good outcomes of using the hyper-parameters optimisation.

CHAPTER 4

EXPERIMENTS AND RESULTS

4.1 Introduction

In this study, we used the ChaLearn LAP 2016 APA dataset [9]. This dataset consists of 10,000 short clip video sequences with an average duration of 15 seconds each, the resolution of the videos varies between $682x406$ and $720x1280$, and the number of frames varies between [49 – 456] (see Table. 4.1). These video sequences were retrieved from YouTube and include more than 3,000 subjects. The subjects spoke English in front of a camera. The subjects depicted in the clips have different ages, genders, nationalities, and ethnicities.

The competition consisted of two phases, a validation phase and a testing phase. In the first phase, participants had access to 6,000 labeled video sequences, representing 60% of the dataset as a training set, and 2,000 unlabeled videos, representing 20% as a validation set. In the second phase, participants had access to the labeling of the previous validation set, and access to an additional 2,000 unlabeled videos as a test set.

This chapter describes the proposed framework's context and its construction techniques. We begin by investigating the influence of statistical descriptors on hand-crafted and deep learning methods. Then, we evaluate the impact of PML level on hand-crafted approaches. Next, we explore how feature selection affects the final results. Furthermore, we compared our framework's results to those of hand-crafted and deep learning approaches. Finally, we compared the performance of the PML-COV framework with that of the state-of-the-art automatic personality estimation approaches.

Number of frames	Count of videos
Equals 49 frames	1
Less than 100 frames	4
Less than 150 frames	18
Less than 200 frames	53
Less than 250 frames	333
Less than 300 frames	373
Less than 350 frames	430
Less than 400 frames	2714
Less than 456 frames	2826
Equals 456 frames	7174

Table 4.1: Database number of frames statistics

4.2 Experimental settings

In this study, we present a novel framework for evaluating the Big-Five personality traits and screening attributes of job candidates from facial videos. This framework is based on the Covariance descriptor (COV) for face image analysis, which can extract rich and distinct low-level face features.

In this section, we discuss the research methodology that led to the development of the PML-COV framework. We begin by identifying the most prominent statistical description, which leads us to identify the mean as the best choice. Then, we explore different feature selection techniques to emphasize the decision of using the NCA technique in our framework. Finally, we illustrate how the hyper-parameter optimisation could improve the final results.

4.2.1 Effect of statistical descriptors on hand-crafted and deep learning methods

We used twelve statistical descriptors to merge our features with the aim of avoiding the loss of the local and temporal information and to identify the optimal set of features which are suitable for our work. These descriptors are: Mean, Variance, Skewness, Root Mean Square (RMS), Peak min, Peak max, Crest factor min, Crest factor max, Crest pulse min, Crest pulse max, Kurtosis, and the Shape factor.

The results shown in Figure 4.2 for hand-crafted methods and Figure 4.1 for the deep learning ones demonstrate our preference of choosing the mean descriptor over other statistical descriptors. Note that the RMS descriptor is also as good as it gets to the mean, however in this study we only used the mean. We fixed the PML level for the hand-crafted methods to $l = 7$. This choice is discussed later in Section 4.2.2.

In the hand-crafted methods (see Figure 4.2), the COV descriptor combines the dimensional features by mapping $d = 19$ channels as follows:

$$
\begin{aligned}
F(x,y) = \big[& x, y, I, Ix, Iy, Ixx, Iyy, \\
& LDP(k=3), LDP(k=5), \\
& LDP(k=7), BSIF(f=9\cdot 9), \\
& BSIF(f=11\cdot 11), BSIF(f=13\cdot 13), \\
& LPQ(ws=7), LPQ(ws=9), \\
& LPQ(ws=11), LBP(r=1,n=8) \\
& LBP(r=2,n=8), LBP(r=2,n=16)\big]^T,
\end{aligned}
$$

where each descriptor in the dimensional feature image has been fed by a grayscale image which result to 19 2-D arrays as follows: x and y are the pixel location, I is the intensity, Ix, Iy, Ixx, Iyy are the first and second spatial intensity derivatives, $LDP(k)$ is LDP image obtained for a given $k = \{3,5,7\}$ most prominent directions, $BSIF(f)$ is BSIF image obtained for a given texture filter of size $f = \{9\cdot 9, 11\cdot 11, 13\cdot 13\}$ and $8bits$ length, $LPQ(ws)$ is LPQ image obtained for a given window size $ws = \{7,9,11\}$, and finally $LBP(r,n)$ is LBP image obtained for a given radius $r = \{1,2\}$ and number of neighboring points $n = \{8,16\}$, since the number of channels used is 19 then the COV descriptor for each region is described by $D = d\cdot(d+1)/2 = 190$ features. And the total number of features for the whole image is $B\cdot D = 140\cdot 190$, where B is the total number of blocks, D is the image descriptor size in each block.

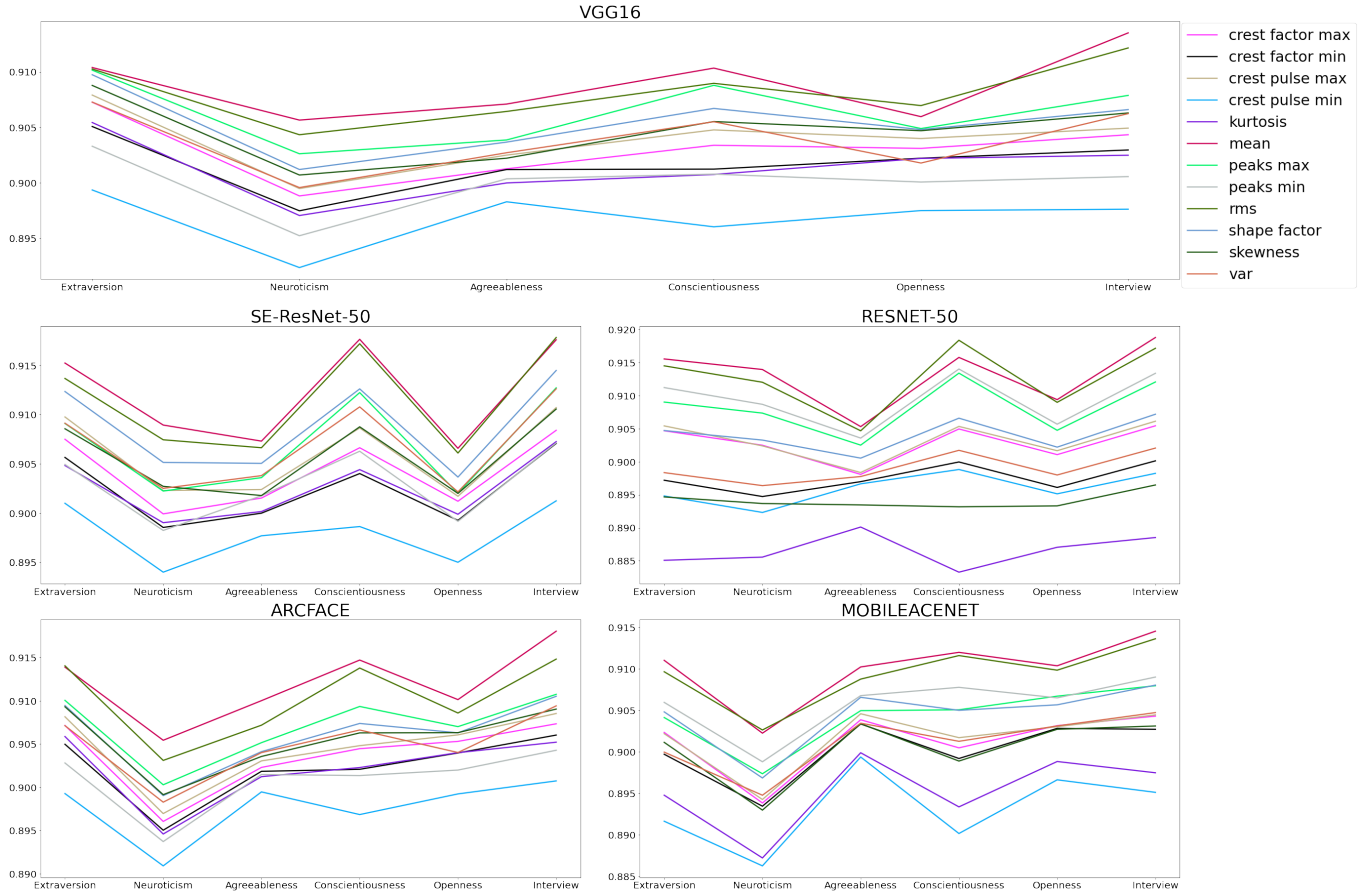


Figure 4.1: Effect of statistical descriptors on deep learning methods (PML $l = 7$)

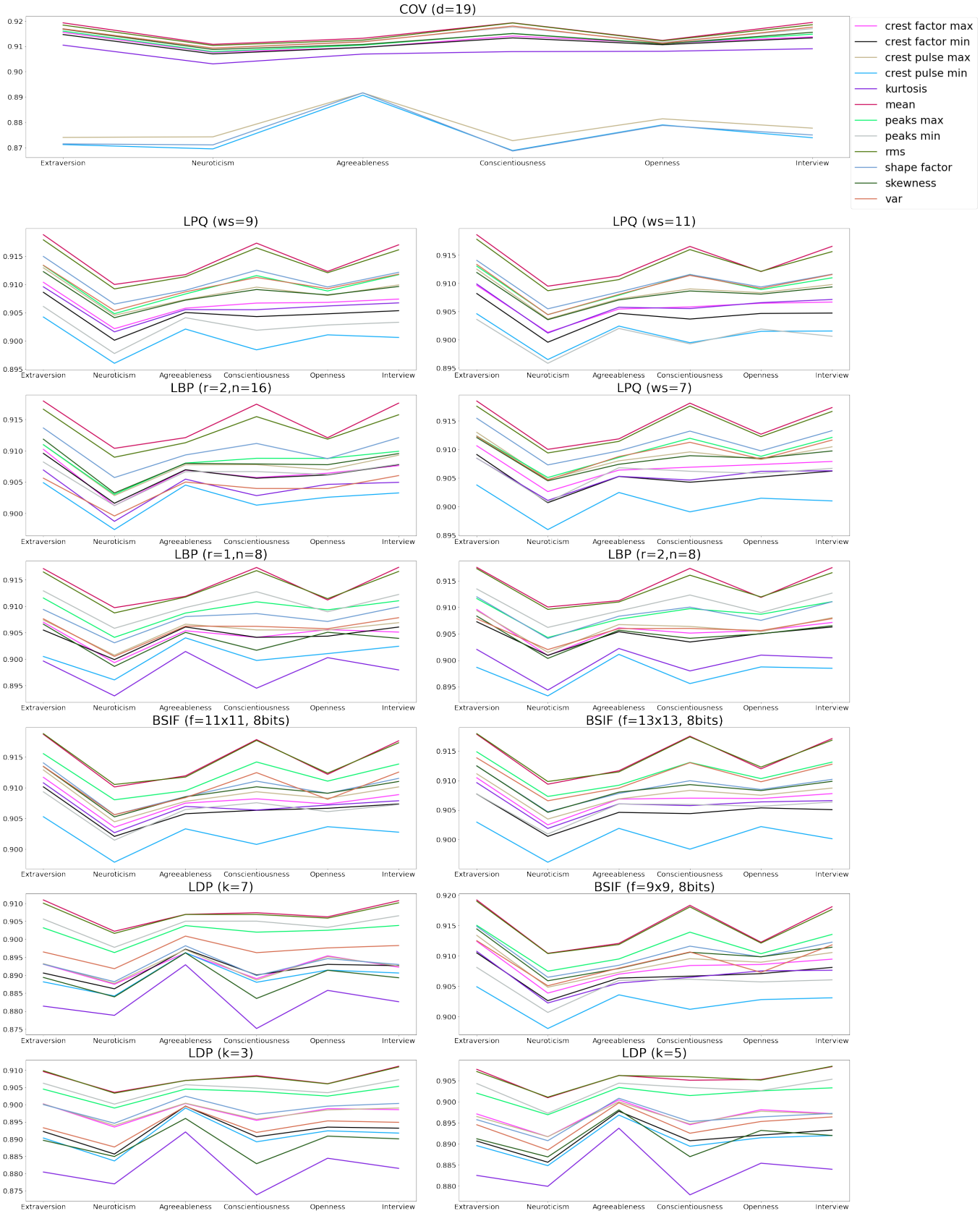


Figure 4.2: Effect of statistical descriptors on hand-crafted methods (PML $l = 7$)

4.2.2 Effect of PML level on hand-crafted methods

In this study, we used PML face representation due to its success in enriching the information of the face by generating multiscale multiblocks in which face part properties are efficiently encoded. This is shown in [4] [14] [88] . Figure 4.3 illustrates the effect of the PML level on the final results. As illustrated, PML with level $l = 7$ outperforms the others. Hence, we used PML with level 7 for the hand-crafted descriptors, including the PML-COV framework.

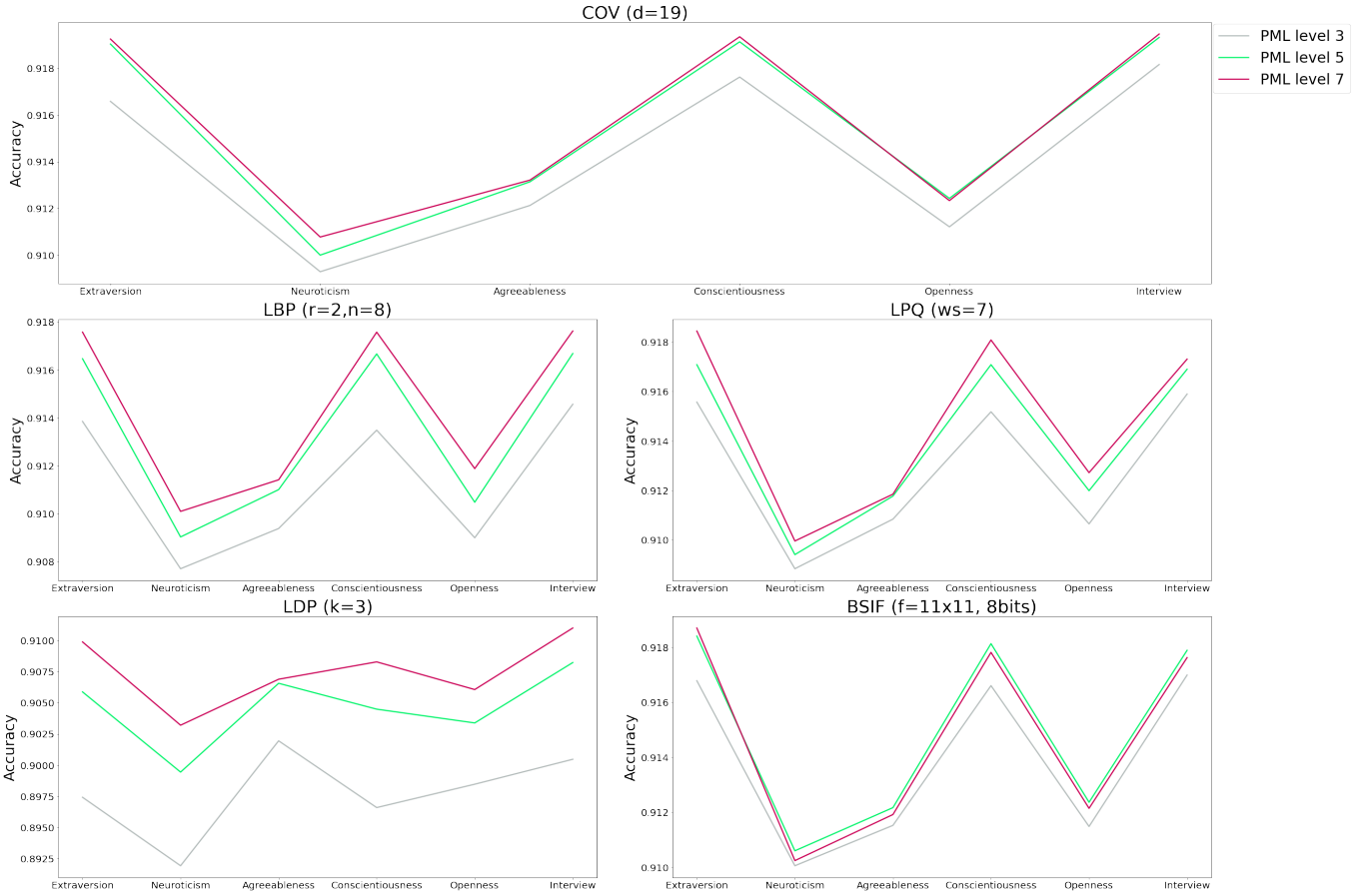


Figure 4.3: Effect of PML level on hand-crafted methods (PML $l = 7$)

4.2.3 Effect of feature selection

Feature selection plays an important role in choosing the important features and dropping the redundant ones. In this study, for the NCA algorithm, we used the mean absolute deviation loss function given by $l(y_i, \widehat{y}_i) = |y_i - \widehat{y}_i|$, with a regularisation parameter $\lambda = 5e - 4$.

NCA performed well in this study, as illustrated in Figure 4.5 for hand-crafted approaches and Figure 4.4 for deep learning-based approaches. It improved the results and outperformed the Relief and MRMR feature selection techniques. Therefore, we selected NCA as our main feature selection technique for this study.

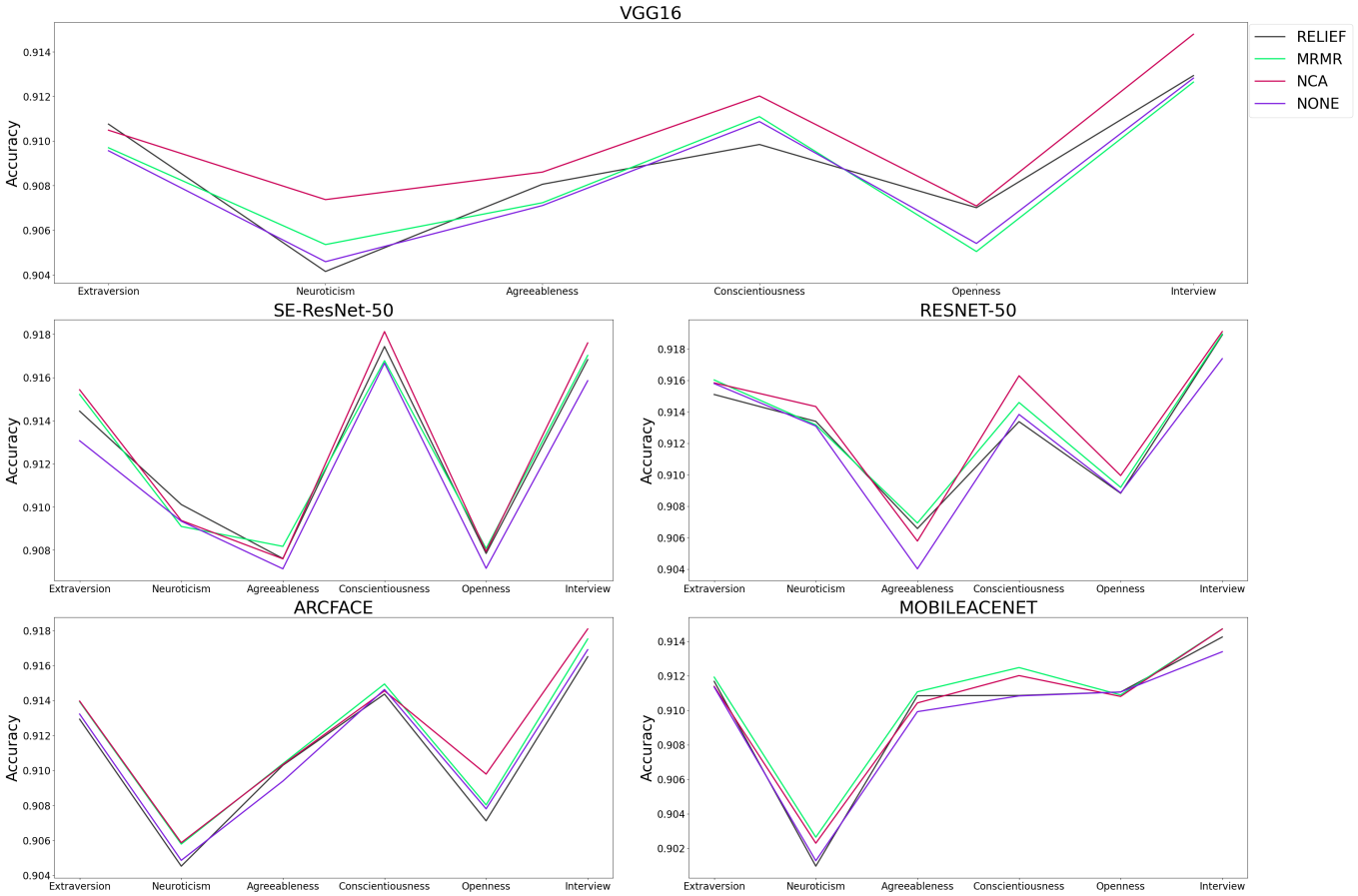


Figure 4.4: Effect of feature selection on deep learning methods

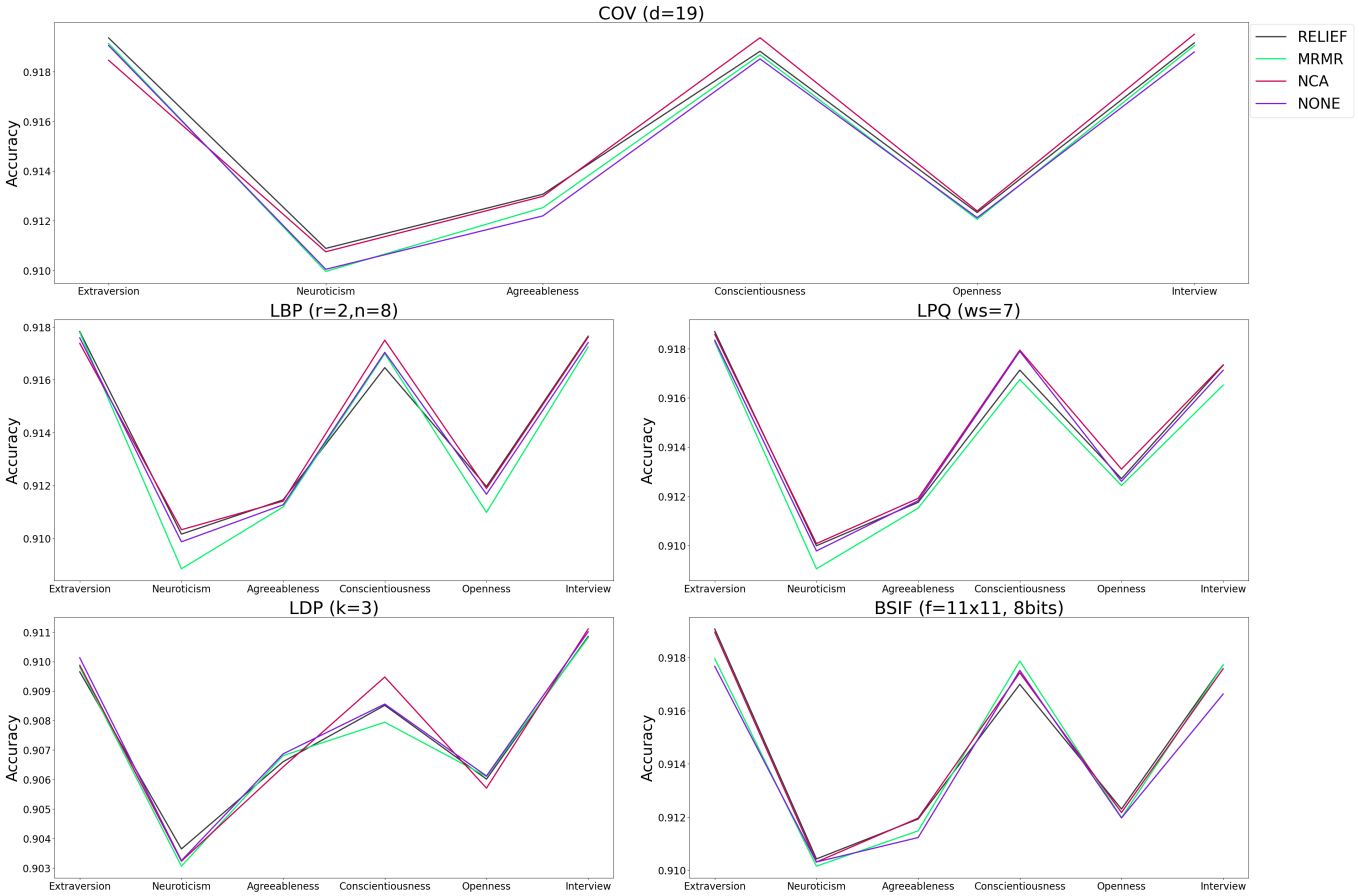


Figure 4.5: Effect of feature selection on hand-crafted methods

4.2.4 Effect of hyper-parameters optimisation

The regression problem in this study is addressed using five Support Vector Regressors (SVRs) for the Big-Five personality traits and one Gaussian Process Regressor (GPR) for the interview variable. Hyper-parameter optimization is applied to these SVRs and GPR to enhance the final results. The SVRs utilize the Gaussian kernel function and the Sequential Minimal Optimization (SMO) optimization solver algorithm. On the other hand, the GPR employs the Automatic Relevance Determination (ARD) Squared Exponential Kernel and optimizes only the standard deviation (σ) parameter. Figure 4.6 and Figure 4.7 demonstrate the significant difference between optimized and non-optimized SVRs and GPR.

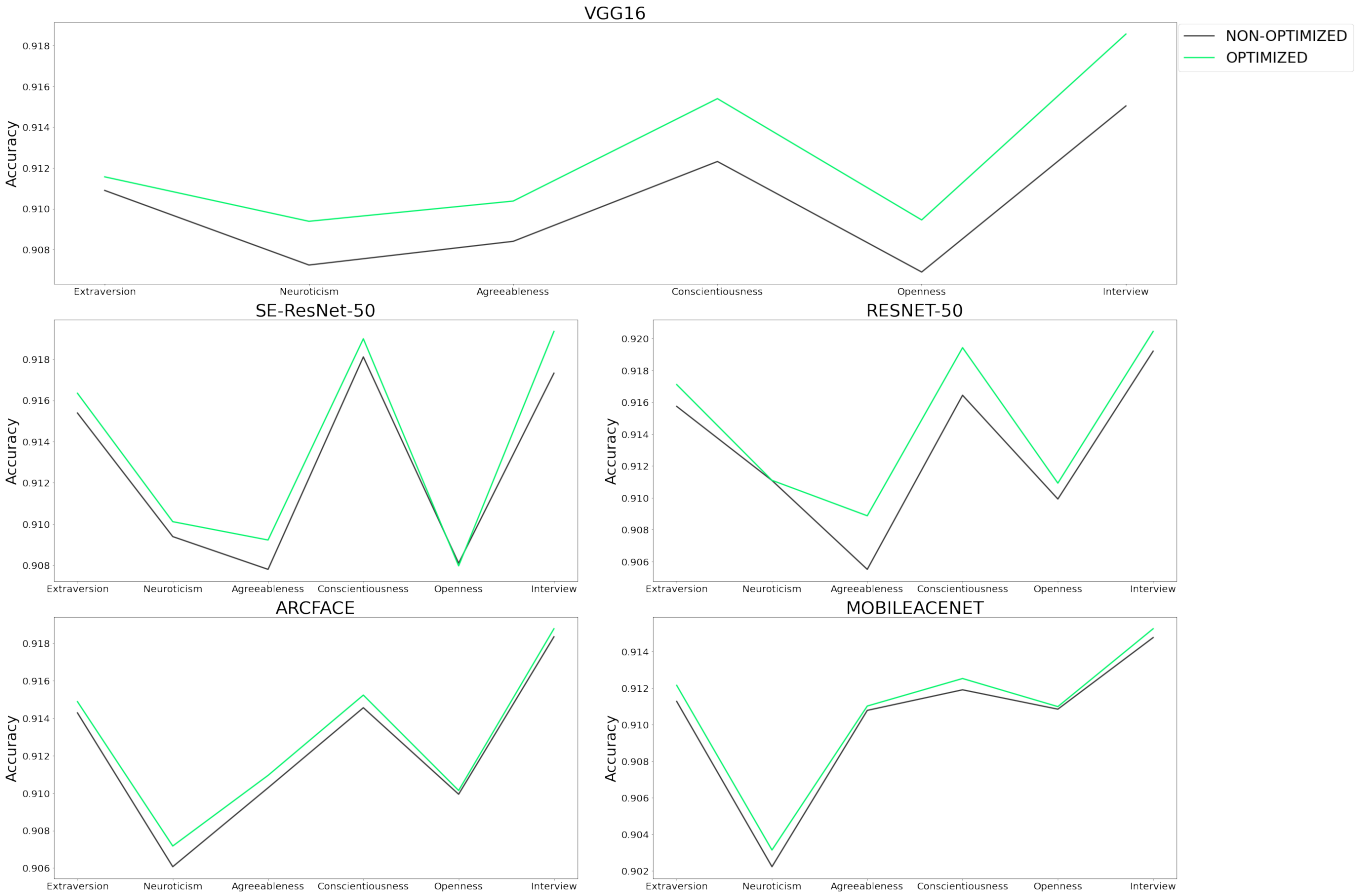


Figure 4.6: Effect of hyper-parameters optimisation on methods based on deep learning.

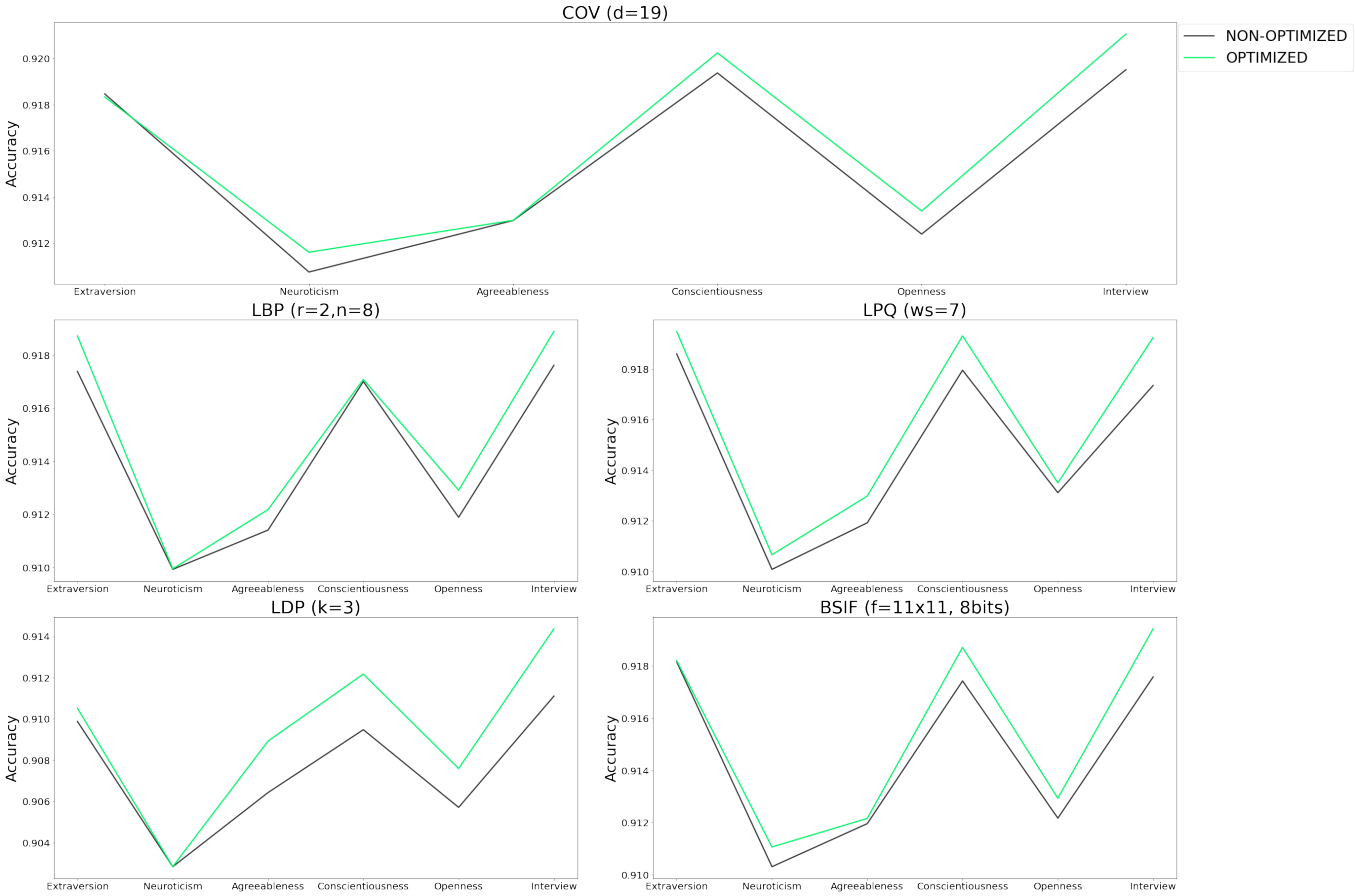


Figure 4.7: Effect of hyper-parameters optimisation on hand-crafted methods.

4.3 Results and discussion

In our experiments, we performed two types of video feature extraction. Initially, we employ five different hand-crafted descriptors with the same PML face image representation (l = 7). Then, we utilize five different CNN models. The purpose of this methodological variation is to decide which method is most applicable to our situation. Table 4.2 provides a summary of the performance on the test set. As demonstrated, the PML-COV descriptor beats the other approaches due to its ability to encode low-level facial features and its ability to combine multiple well-known image texture descriptors. PML-COV is extremely efficient at extracting the features of the video sequence; it requires less time and effort. The CNN-based method needs to be tuned to a specific or similar task as the target problem before extracting rich features from it. This training is time-consuming due to the extensive operations involved. The performance of the PML-COV descriptor on validation and test subsets is summarized in Table 4.3.

Table 4.2: Comparison of performances of the proposed PML-COV descriptor with other handcrafted and deep descriptors

	Method	AGRE	CONS	EXTR	NEUR	OPEN	MEAN	**INTER**
Deep	VGG-FACE	91.04	91.54	91.16	90.94	90.94	91.12	91.86
	ResNet-50	90.89	91.94	91.71	91.11	91.09	91.35	92.04
	SE-ResNet-50	90.92	91.90	91.63	91.01	90.80	91.25	91.93
	MobileFaceNet	91.10	91.25	91.21	90.31	91.10	90.99	91.52
	Arcface	91.09	91.52	91.49	90.72	91.01	91.17	91.88
Handcrafted	PML-LDP	90.89	91.22	91.05	90.28	90.76	90.84	91.44
	PML-LBP	91.22	91.71	91.87	90.99	91.29	91.42	91.89
	PML-LPQ	91.30	91.93	91.95	91.06	91.35	91.52	91.92
	PML-BSIF	91.22	91.87	91.82	91.11	91.29	91.46	91.94
	PML-COV	91.32	92.03	91.91	91.06	91.31	91.53	92.11

Table 4.3: PML-COV results for validation and test subsets.

Trait	Validation	Test
AGRE	91.67	91.32
CONS	91.93	92.03
EXTR	91.81	91.91
NEUR	91.34	91.06
OPEN	91.47	91.31
INTER	92.14	92.11

Figures 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14, 4.15, 4.16, 4.17 demonstrate the performance of the presented approach against other hand-crafted and deep learning methods. They show the correlation between the ground-truth and predicted scores (validation and test sets) for the interview variable. This correlation can be measured by a single measure, which is defined by the Pearson correlation coefficient (PC) [89]. PC is the most common method for calculating linear correlation. PC measures how strong the linear association is between two continuous variables (ground-truth scores and estimated scores). Given this pair of data as $S = \{(x_i, y_i) \,|\, x_i \in \mathbb{R}, y_i \in \mathbb{R}; i = 1, ..., n\}$, where x represents the ground-truth scores and y represents the estimated scores. PC is calculated as follows:

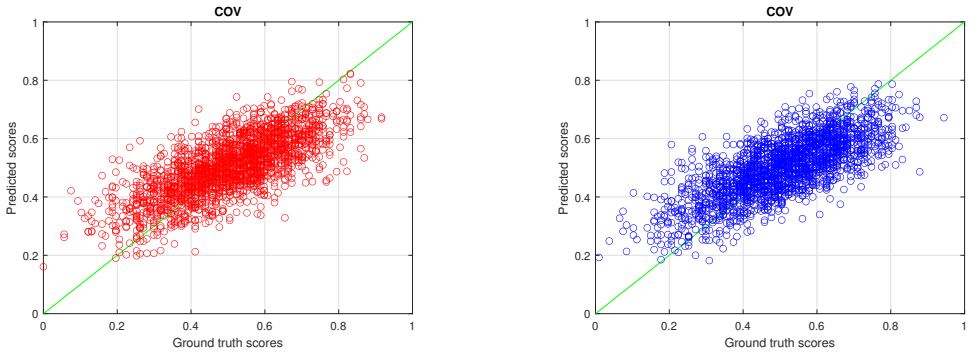
$$PC = \frac{\sum_{i=1}^{n}(x_i - \widehat{x})(y_i - \widehat{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \widehat{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \widehat{y})^2}} \tag{4.1}$$

where $\widehat{x}$ and $\widehat{y}$ are the mean values of x and y respectively.

Table 4.4: Interview (PC) for validation and test subsets.

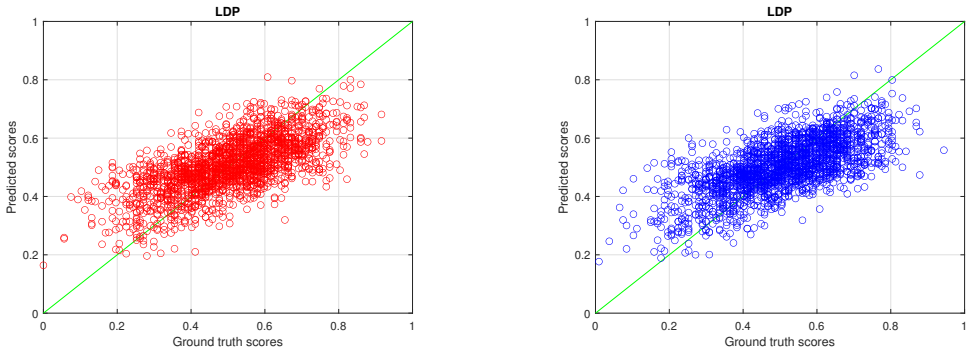
	Method	Validation	Test
Deep	VGG-FACE	0.7157	0.7143
	ResNet-50	0.7201	0.7239
	SE-ResNet-50	0.7233	0.7269
	MobileFaceNet	0.7194	0.7214
	Arcface	0.7165	0.7210
Handcrafted	PML-LDP	0.6665	0.6701
	PML-LBP	0.7205	0.7245
	PML-LPQ	0.7214	0.7231
	PML-BSIF	0.7226	0.7222
	PML-COV	0.7297	0.7335

PC ranges from -1 to 1, while a value of zero indicates that there is no correlation relationship between the two variables. A PC larger than zero shows a positive association between the two variables, in which an increase in the value of one variable results in an increase in the value of the other one. A PC less than zero shows a negative relationship between two variables, in which a rise in one variable decreases the other. Table 4.4 shows the PC of the hand-crafted methods and deep methods. As it is illustrated, the PML-COV have the best correlation and its PC is 0.7297 for the validation set, and 0.7335 for the test sets. This indicates a good linear correlation between prediction and ground truth, as a perfect prediction would have a PC equal to one.



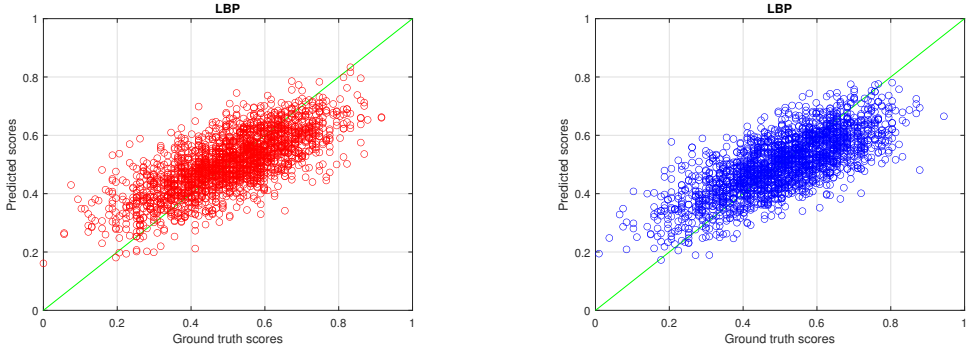
(a) Validation. **(b)** Test.

Figure 4.8: Correlations between true interview and estimated interview by the PML-COV descriptor.



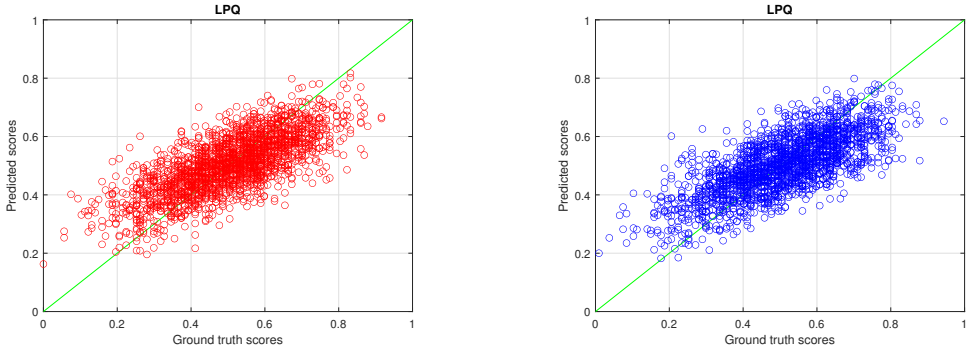
(a) Validation. **(b)** Test.

Figure 4.9: Correlations between true interview and estimated interview by the LDP descriptor.



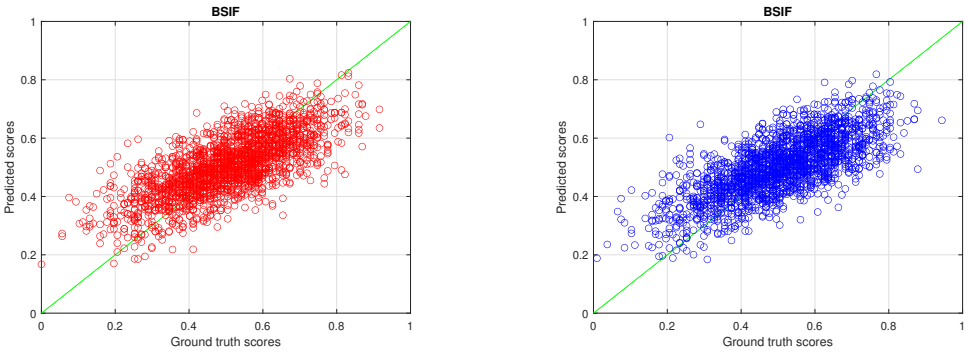
(a) Validation. **(b)** Test.

Figure 4.10: Correlations between true interview and estimated interview by the LBP descriptor.



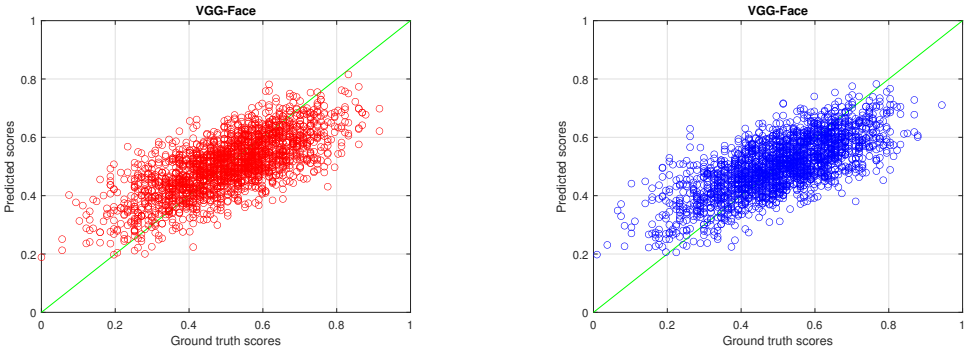
(a) Validation. **(b)** Test.

Figure 4.11: Correlations between true interview and estimated interview by the LPQ descriptor.



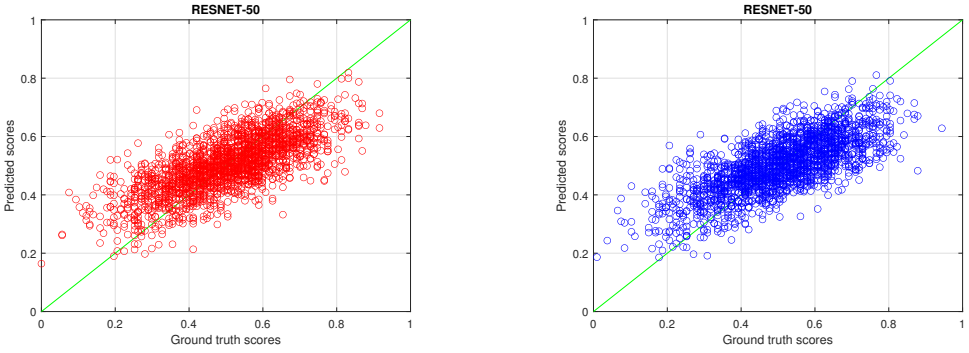
(a) Validation. **(b)** Test.

Figure 4.12: Correlations between true interview and estimated interview by the BSIF descriptor.



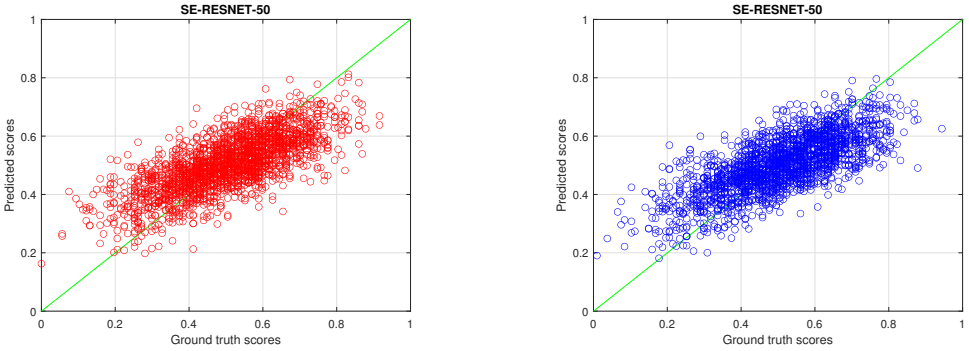
(a) Validation. **(b)** Test.

Figure 4.13: Correlations between true interview and estimated interview by the VGG16 model.



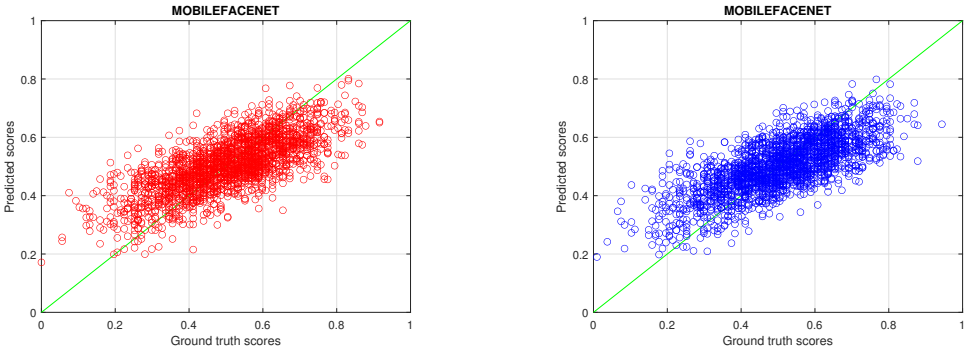
(a) Validation. **(b)** Test.

Figure 4.14: Correlations between true interview and estimated interview by the ResNet-50 model.



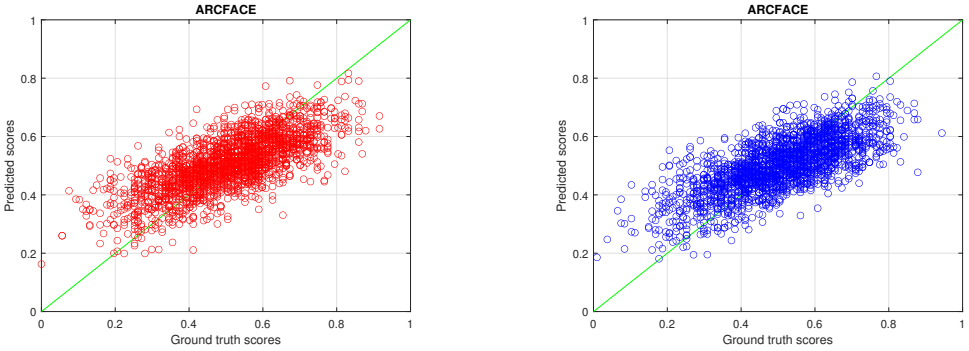
(a) Validation. **(b)** Test.

Figure 4.15: Correlations between true interview and estimated interview by the SE-ResNet-50 model.



(a) Validation. **(b)** Test.

Figure 4.16: Correlations between true interview and estimated interview by the MobileFaceNet model.



(a) Validation. **(b)** Test.

Figure 4.17: Correlations between true interview and estimated interview by the ArcFace model.

In Table 4.5, we summarize the results obtained by state-of-the-art methods and schemes using the ChaLearn LAP 2016 APA dataset. PML-COV outperformed the other methods, although it uses only visual information, unlike most of the other competing methods that use multimodal feature fusion. In addition, many of these approaches relied on deep learning, a computationally very expensive and time-consuming method. The CPU time associated with the PML-COV framework for each stage is given in Table 4.6. For the computation time, both the total test (2000 videos) and the average (1 video) are given. The test was done on a custom Windows 10 workstation with an Intel Xeon Processor $E5-2658v3$, $30M$ Cache, $2.20GHz$, and $64GB$ of RAM.

Table 4.5: A comparison of the proposed approach with other automatic personality estimation approaches.

Approach	Deep Learning	AGRE	CONS	EXTR	NEUR	OPEN	MEAN	**INTER**
DAN+ [90]	YES	91.20	91.40	91.50	90.70	91.00	91.16	-
DRN-Baseline [91]	YES	91.02	91.38	91.07	90.89	91.11	91.09	-
evolgen [6]	YES	91.19	91.19	91.50	90.99	91.17	91.21	-
NJU-LAMDA [5]	YES	91.26	91.66	91.33	91.00	91.23	91.30	-
FDMB [18]	NO	89.10	86.59	87.88	86.32	87.47	87.47	87.21
ROCHCI [18]	NO	90.32	89.49	90.26	90.11	90.47	90.13	90.18
PML [4]	NO	91.03	91.37	91.55	90.82	91.00	91.15	91.57
Baseline [7]	YES	91.12	91.52	91.12	91.03	91.11	91.18	91.62
BU-NKU [8]	YES	91.37	91.97	92.12	91.46	91.70	91.72	92.09
PML-COV (ours)	NO	91.32	92.03	91.91	91.06	91.31	91.53	92.11

Table 4.6: CPU time (seconds) of the different stages of our proposed framework.

Stage	Task	Testing time (2000 videos)	Average (1 video)
Prepossessing	Detection and landmarks	8022.0	4.0110
	Alignment and crop	3302.1	1.6511
Feature extraction	Video descriptor computation	57219.0	28.6095
Estimation	BIG-5	180.6556	0.0903
	Interview	4.8262	0.0024
	Total	68728.5818	34.3643

4.4 Conclusion

In this chapter, we introduced the PML-COV framework and compared it to four other hand-crafted approaches and five deep learning approaches. In the first section, we gave a brief overview of the ChaLearn LAP 2016 APA dataset and its statistics. In the second section, we initially analyzed the effect of the statistical descriptors that were used to combine video feature vectors and concluded that the mean descriptor is the most suited for our purpose. Then we investigated the influence of the PML level and discovered that PML level 7 is more accurate than other levels. Afterword, we compared three feature selection techniques. NCA was our best choice. Next, after examining the effect of hyper-parameters on SVRs and the GPR techniques used to solve our regression problem, we discovered that the optimization improved our final findings. In the third section, we compared the performance of the proposed PML-COV descriptor with other hand-crafted and deep learning descriptors. Then, we explore and measure the correlation between the ground truth and predicted scores of the interview variable using the Pearson correlation coefficient (PC). PML-COV performed very well. Next, we compared the proposed PML-COV framework with various state-of-the-art automatic personality estimation approaches. PML-COV outperformed the other techniques, including those based on deep learning, despite the fact that PML-COV uses only visual information. Finally, we report the CPU time associated with the PML-COV framework for each stage.

CONCLUSIONS

Conclusion

In recent years, government and private-sector organizations have grown, as has the demand for job applications. Sorting through these CVs is time-consuming and requires hard work. Nowadays, we need a human-like computer vision system to perform this task on our behalf. *ChaLearn Looking at People CVPR 2017* addressed this challenge and introduced a new database to assist both recruiters and job candidates with speed interviews by using automatic recommendations based on multi-media CVs. The goal is to find out if a candidate has enough potential to be invited to an interview through exploring his apparent personality traits.

Personality analysis from videos is a challenging problem in computer vision. In this thesis, we have developed a new framework for evaluating Big-Five personality traits and screening attributes of job candidates using facial videos, and we show that it is capable of solving regression problems in comparison with other state-of-the-art approaches. The proposed approach achieves high accuracy that outperforms the state-of-the-art results, including deep CNNs. In addition, we conducted an extensive experiment to compare hand-crafted features with deep features. Our goal was to strike the right balance between accuracy, complexity, and the time required for training and testing. Despite the fact that deep learning approaches are effective at solving complex problems, including those related to time series, they have drawbacks due to their complexity, cost of computation, hyper-parameter tuning, choosing the right architecture, and the requirement of massive amounts of training data. Hand-crafted methods usually find a good balance between the complexity, the lack of data and the time required to train and test.

Limitations and future works

In addition to the current focus on video modality for system evaluation, the inclusion of auditory and textual modalities has the potential to enhance the overall effectiveness of the framework. Future applications of this framework could include pain evaluation, disguised face identification, and driver drowsiness detection.

One promising avenue for expansion is pain evaluation. By incorporating additional modalities, such as vocal expressions or self-reported pain assessments, alongside facial video analysis, the framework could be adapted to assess pain levels in individuals. This has significant implications for medical and healthcare domains, where reliable pain

evaluation is crucial for effective treatment and patient care.

Another potential application lies in disguised face identification. The framework could be utilized to detect and identify individuals who attempt to conceal their identity through disguises. This can be particularly valuable in security and surveillance settings, aiding in the identification of potential threats or persons of interest.

Furthermore, the framework could be extended to address driver drowsiness detection. By analyzing facial video and other relevant data, such as changes in speech patterns or the detection of yawns, the system can help identify signs of driver fatigue or drowsiness. This has significant implications for road safety, as early detection of drowsiness can facilitate timely interventions and prevent accidents caused by driver inattention.

By exploring these potential avenues of research and development, the framework can be adapted and applied to various domains, opening up new possibilities for improving human-computer interaction and enhancing the overall effectiveness of multimodal analysis systems.

Bibliography

[1] A. Chergui, S. Ouchtati, H. Telli, F. Bougourzi, and S. E. Bekhouche, "Lpq and ldp descriptors with ml representation for kinship verification," in *The second edition of the International Workshop on Signal Processing Applied to Rotating Machinery Diagnostics (SIGPROMD'2018)*, pp. 1–10, 2018.

[2] A. Chergui, S. Ouchtati, J. Sequeira, S. E. Bekhouche, F. Bougourzi, and H. Telli, "Deep features for kinship verification from facial images," in *2019 International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, pp. 64–67, IEEE, 2019.

[3] A. Chergui, S. Ouchtati, J. Sequeira, S. Bekhouche, and H. Telli, "Robust kinship verification using local descriptors," in *Proc. Third Int. Conf. Adv. Technol. and Electr. Eng*, 2018.

[4] S. E. Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed, "Personality traits and job candidate screening via analyzing facial videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1660–1663, July 2017.

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

[6] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal in *European Conference on Computer Vision*, pp. 337–348, Springer, 2016.

[7] Y. Güçlütürk, U. Güçlü, X. Baro, H. J. Escalante, I. Guyon, S. Escalera, M. A. Van Gerven, and R. Van Lier, "Multimodal first impression analysis with deep residual networks," *IEEE Transactions on Affective Computing*, 2017.

[8] H. Kaya, F. Gürpinar, and A. A. Salah, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pp. 1651–1659, IEEE, 2017.

[9] 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 *European conference on computer vision*, pp. 400–418, Springer, 2016.

[10] J. Willis and A. Todorov, "First impressions: Making up your mind after a 100-ms exposure to a face," *Psychological Science*, vol. 17, no. 7, pp. 592–598, 2006.

[11] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, pp. 175–215, 07 1992.

[12] A. E. Abele and B. Wojciszke, "Agency and communion from the perspective of self versus others.," *Journal of personality and social psychology*, vol. 93, no. 5, p. 751, 2007.

[13] R. Qin, W. Gao, H. Xu, and Z. Hu, "Modern physiognomy: an investigation on predicting personality traits and intelligence from the human face," *arXiv preprint arXiv:1604.07499*, 2016.

[14] S. Bekhouche, A. Ouafi, F. Dornaika, A. Taleb-Ahmed, and A. Hadid, "Pyramid multi-level features for facial demographic estimation," *Expert Systems with Applications*, vol. 80, pp. 297—310, 2017.

[15] A. Moujahid and F. Dornaika, "A pyramid multi-level face descriptor: application to kinship verification," *Multimedia Tools and Applications*, pp. 1–20, 2018.

[16] A. Moujahid and F. Dornaika, "Multi-scale multi-block covariance descriptor with feature selection," *Neural Computing and Applications*, pp. 1–12, 2019.

[17] H. Telli, S. Sbaa, S. E. Bekhouche, F. Dornaika, A. Taleb-Ahmed, and M. B. López, "A novel multi-level pyramid co-variance operators for estimation of personality traits and job screening scores," 2021.

[18] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Gucluturk, U. Guclu, X. Baro, I. Guyon, J. J. Junior, M. Madadi, *et al.*, "Explaining first impressions: Model-

ing, recognizing, and explaining apparent personality from videos," *arXiv preprint arXiv:1802.00745*, 2018.

[19] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, pp. 273–291, July 2014.

[20] J. Junior, C. Jacques, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. van Gerven, *et al.*, "First impressions: A survey on computer vision-based apparent personality trait analysis," *arXiv preprint arXiv:1804.08046*, 2018.

[21] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *International conference on image and signal processing*, pp. 236–243, Springer, 2008.

[22] V. Vapnik, "The nature stat. learning theory," 1995.

[23] T. Ruf, A. Ernst, and C. Küblbeck, "Face detection with the sophisticated high-speed object recognition engine (shore)," in *Microelectronic Systems*, pp. 243–252, Springer, 2011.

[24] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[25] J. Kannala and E. Rahtu, "Bsif: Binarized statistical image features," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 1363–1366, IEEE, 2012.

[26] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Adaptive computation and machine learning, MIT Press, 2006.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

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

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)* (X. X. M. W. Jones. and G. K. L. Tam, eds.), pp. 41.1–41.12, BMVA Press, September 2015.

[30] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 356–361, 2013.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[32] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (New York, NY, USA), p. 1459–1462, Association for Computing Machinery, 2010.

[33] L. V. Phan and J. F. Rauthmann, "Personality computing: New frontiers in personality assessment," *Social and Personality Psychology Compass*, vol. 15, no. 7, p. e12624, 2021.

[34] M. Fajkowska and S. Kreitler, "Status of the trait concept in contemporary personality psychology: Are the old questions still the burning questions?," *Journal of Personality*, vol. 86, no. 1, pp. 5–11, 2018.

[35] J. Costa, Paul T. and R. R. McCrae, "Age Differences in Personality Structure: a Cluster Analytic Approach1," *Journal of Gerontology*, vol. 31, pp. 564–570, 09 1976.

[36] 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, no. 1, p. 81, 1987.

[37] 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.

[38] P. T. Costa Jr and R. R. McCrae, *Neo Personality Inventory.* Oxford University Press, 2000.

[39] R. A. Power and M. Pluess, "Heritability estimates of the big five personality traits based on common genetic variants," *Translational psychiatry*, vol. 5, no. 7, pp. e604–e604, 2015.

[40] T. A. Judge, C. A. Higgins, C. J. Thoresen, and M. R. Barrick, "The big five personality traits, general mental ability, and career success across the life span," *Personnel psychology*, vol. 52, no. 3, pp. 621–652, 1999.

[41] R. R. McCrae and A. R. Sutin, "A five–factor theory perspective on causal analysis," *European Journal of Personality*, vol. 32, no. 3, pp. 151–166, 2018.

[42] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1, pp. I–I, IEEE, 2001.

[43] S. E. Bekhouche, A. Ouafi, A. taleb ahmed, A. Hadid, and A. Benlamoudi, "Facial age estimation using bsif and lbp," 12 2014.

[44] A. Chergui, S. Ouchtati, S. Mavromatis, S. E. Bekhouche, J. Sequeira, and H. Zerrari, "Kinship verification using mixed descriptors and multi block face representation," in *2019 International Conference on Networking and Advanced Systems (ICNAS)*, pp. 1–6, IEEE, 2019.

[45] I. Adjabi, A. Ouahabi, A. Benzaoui, and S. Jacques, "Multi-block color-binarized statistical images for single-sample face recognition," *Sensors*, vol. 21, no. 3, p. 728, 2021.

[46] S. E. Bekhouche, A. Ouafi, A. Benlamoudi, A. Taleb-Ahmed, and A. Hadid, "Facial age estimation and gender classification using multi level local phase quantization," in *2015 3rd International Conference on Control, Engineering & Information Technology (CEIT)*, pp. 1–4, IEEE, 2015.

[47] M. Zighem, A. Ouafi, S. Beckhouche, A. Benlamoudi, and A. Taleb-Ahmed, "Age estimation based on color facial texture," 2017.

[48] F. Bougourzi, S. Bekhouche, M. Zighem, A. Benlamoudi, and A. Taleb-Ahmed, "A comparative study on textures descriptors in facial gender classification," in *10me Confrence sur le Gnie Electrique*, 2017.

[49] A. Benlamoudi, F. Bougourzi, M. Zighem, S. Bekhouche, A. Ouafi, and A. Taleb-Ahmed, "Face anti-spoofing combining mllbp and mlbsif," in *Proceedings of the 10ème Conférence sur le Génie Electrique, Alger, Algerie*, vol. 30, 2017.

[50] A. Chergui, S. Ouchtati, J. Sequeira, S. Bekhouche, and F. Bougourzi, "Discriminant analysis for facial verification using color images," in *Proc. First Int. Conf. Electr. Eng*, 2018.

[51] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6-7, pp. 379–390, 2014.

[52] A. Benlamoudi, "Multi-modal and anti-spoofing person identification," *University of Kasdi Merbah*, 2018.

[53] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[54] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, "Fast high dimensional vector multiplication face recognition," in *2013 IEEE International Conference on Computer Vision*, pp. 1960–1967, 2013.

[55] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[56] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li, "Face detection based on multi-block lbp representation," in *International conference on biometrics*, pp. 11–18, Springer, 2007.

[57] B. Patel, R. Maheshwari, and R. Balasubramanian, "Multi-quantized local binary patterns for facial gender classification," *Computers & Electrical Engineering*, vol. 54, pp. 271–284, 2016.

[58] A. Hafiane, G. Seetharaman, and B. Zavidovique, "Median binary pattern for textures classification," in *International Conference Image Analysis and Recognition*, pp. 387–398, Springer, 2007.

[59] H. Lu, M. Yang, X. Ben, and P. Zhang, "Divided local binary pattern (dlbp) features description method for facial expression recognition," *JOURNAL OF INFORMATION &COMPUTATIONAL SCIENCE*, vol. 11, no. 7, pp. 2425–2433, 2014.

[60] J. Lizé, V. Débordès, H. Lu, K. Kpalma, and J. Ronsin, "Local binary pattern and its variants: application to face analysis," in *International conference on smart Information & communication Technologies*, pp. 94–102, Springer, 2019.

[61] T. Jabid, M. H. Kabir, and O. Chae, "Local directional pattern (ldp)–a robust image descriptor for object recognition," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 482–487, IEEE, 2010.

[62] O. Tuzel, F. M. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *ECCV*, 2006.

[63] E. Vantaggiato, E. Paladini, F. Bougourzi, C. Distante, A. Hadid, and A. Taleb-Ahmed, "Covid-19 recognition using ensemble-cnns in two new chest x-ray databases," *Sensors*, vol. 21, no. 5, p. 1742, 2021.

[64] F. Bougourzi, F. Dornaika, N. Barrena, C. Distante, and A. Taleb-Ahmed, "Cnn based facial aesthetics analysis through dynamic robust losses and ensemble regression," *Applied Intelligence*, pp. 1–18, 2022.

[65] M. Korichi, D. Samai, A. Meraoumia, and A. Benlamoudi, "Towards effective 2d and 3d palmprint recognition using transfer learning deep features and reliff method," in *2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)*, pp. 1–6, 2021.

[66] A. Benlamoudi, S. E. Bekhouche, M. Korichi, K. Bensid, A. Ouahabi, A. Hadid, and A. Taleb-Ahmed, "Face presentation attack detection using deep background subtraction," *Sensors*, vol. 22, no. 10, p. 3760, 2022.

[67] F. Dornaika, S. E. Bekhouche, and I. Arganda-Carreras, "Robust regression with deep cnns for facial age estimation: An empirical study," *Expert Systems with Applications*, vol. 141, p. 112942, 2020.

[68] A. Bardes, J. Ponce, and Y. LeCun, "Vicregl: Self-supervised learning of local visual features," *arXiv preprint arXiv:2210.01571*, 2022.

[69] F. Bougourzi, C. Distante, A. Ouafi, F. Dornaika, A. Hadid, and A. Taleb-Ahmed, "Per-covid-19: A benchmark dataset for covid-19 percentage estimation from ct-scans," *Journal of Imaging*, vol. 7, no. 9, p. 189, 2021.

[70] F. Bougourzi, F. Dornaika, and A. Taleb-Ahmed, "Deep learning based face beauty prediction via dynamic robust losses and ensemble regression," *Knowledge-Based Systems*, vol. 242, p. 108246, 2022.

[71] A. Chergui, S. Ouchtati, S. Mavromatis, S. E. Bekhouche, M. Lashab, and J. Sequeira, "Kinship verification through facial images using cnn-based features.," *Traitement du Signal*, vol. 37, no. 1, 2020.

[72] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[73] 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 (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[74] 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*, pp. 770–778, 2016.

[75] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 67–74, 2018.

[76] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[77] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, June 2019.

[78] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*, pp. 87–102, Springer, 2016.

[79] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*, pp. 428–438, Springer, 2018.

[80] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[81] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with relieff," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.

[82] M. Robnik-Šikonja and I. Kononenko, "An adaptation of relief for attribute estimation in regression," in *Machine learning: Proceedings of the fourteenth international conference (ICML'97)*, vol. 5, pp. 296–304, 1997.

[83] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[84] W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," *JOURNAL OF COMPUTERS*, vol. 7, no. 1, p. 161, 2012.

[85] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.

[86] P.-H. Chen, R.-E. Fan, and C.-J. Lin, "A study on smo-type decomposition methods for support vector machines," *IEEE transactions on neural networks*, vol. 17, no. 4, pp. 893–908, 2006.

[87] MATLAB, *Documentation (R2022b)*. Natick, Massachusetts: The MathWorks, Inc., 2022.

[88] F. Bougourzi, F. Dornaika, K. Mokrani, A. Taleb-Ahmed, and Y. Ruichek, "Fusing transformed deep and shallow features (ftds) for image-based facial expression recognition," *Expert Systems with Applications*, vol. 156, p. 113459, 2020.

[89] I. Olkin, J. W. Pratt, *et al.*, "Unbiased estimation of certain correlation coefficients," *The Annals of Mathematical Statistics*, vol. 29, no. 1, pp. 201–211, 1958.

[90] C. Ventura, D. Masip, and A. Lapedriza, "Interpreting cnn models for apparent personality trait regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 55–63, 2017.

[91] 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 *European Conference on Computer Vision*, pp. 349–358, Springer, 2016.