PERSONALITY RECOGNITION USING COMPOSITE AUDIO-VIDEO FEATURES ON CUSTOM CNN ARCHITECTURE

BY

ENG ZI JYE

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONS)

Faculty of Information and Communication Technology

(Kampar Campus)

JANUARY 2020

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: PERSONALITY RECOGNITION USING COMPOSITE AUDIO-VIDEO FEATURES ON CUSTOM CNN ARCHITECURE

Academic Session: JANUARY 2020

I ENG ZI JYE

(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,

(Author's signature)

(Supervisor's signature)

Address:
81 JLN NUSA PERINTIS,
TMN NUSA PERINTIS 2 10/7
81550 GELANG PATAH JOHOR

Dr. Aun Yichiet

Date: 24/04/2020

Date: 24/04/2020

PERSONALITY RECOGNITION USING COMPOSITE AUDIO-VIDEO FEATURES ON CUSTOM CNN ARCHITECTURE

BY

ENG ZI JYE

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONS)

Faculty of Information and Communication Technology

(Kampar Campus)

JANUARY 2020

DECLARATION OF ORIGINALITY

I declare that this report entitled "**PERSONALITY RECOGNITION USING COMPOSITE AUDIO-VIDEO FEATURE ON CUSTOM CNN ARCHITECTURE**" is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature : ____________________________

Name : _Eng Zi Jye_________________

Date : _24/4/2020__________________

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisors, Dr. Aun Yichiet and Dr. Savio Tse who has given me this bright opportunity to engage in a Deep learning project. I would also like to thanks everyone who had given helps throughout this project.

In addition, I express my gratitude to my family for giving me the support throughout the course.

ABSTRACT

Automatic personality recognition is becoming more prominent in the domain of intelligent job matching. Traditionally, individual personality traits are measured through questionnaire carefully design based on personality models like the big-five or MBTI. Although the attributes in these models are proven effective; data collection through surveys can result in biased scoring due to illusory superiority. Machine-learning based personality models alleviate these constraints by modelling behavioural cues from videos annotated by personality experts; For example, the ECCV ChaLearn LAP 2016 challenge seek to recognise and quantise human personality traits. Using variants of CNN(s), existing methods attempt to improve model accuracy through adding custom layers and hyper-parameters tuning; trained on the full ChaLearn LAP 2016 datasets that are compute-intensive. This project proposes a rapid behavioural modelling technique for short videos to improve model accuracy and prevent overfitting while minimizing the amount of training data needed. The contribution of this work is two folds: (1) a selective sampling technique using the first seven-seconds of video for training and (2) Using limited amount of dataset to model a personality trait recognition model with optimum performance. By applying selective sampling technique and inclusion of multiple modalities, the model performance able to achieve 90.30 in testing result with almost 600% smaller training data.

TABLE OF CONTENT

APPENDIX A FYP WEEKLY REPORT

PLAGIARISM CHECK RESULT

CHECKLIST FOR FYP2 THESIS SUBMISSION

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial intelligence
ANN	Artificial neural network
API	Application Programming Interface
CNN	Convolutional neural network
DISC	Dominance, inducement, submission and compliance
ECCV	European Conference on Computer Vision
HRM	Human Resource Management
LAP	Looking at people
LSTM	Long short term memory
MBTI	Myers Briggs Type Indicator
et al.	And other
ResNet	Residual Network

Chapter 1 Introduction

1.1 Problem statement and motivation

First impression judgment of a people can be judged from various human characteristics, it can be varying from clothing to facial expression (Teijeiro-Mosquera, Biel, Alba-Castro, and Gatica-Perez, 2015) Study of personality computing aims at quantifying observable human difference based on stable, possibly measurable, individual characteristics. (Vinciarelli and Mohammadi, 2014). Significant advances in the field attract the attention of the researchers and practitioners to put in their effort in the personnel selection as personnel selection was a significant criterion that determines the overall performance of the organization. At the turn of the 20th Century, the research describes how personality and biographical characteristics can influence or even predict job search success. In detail, conscientiousness and extraversion attributes for the job applicant had been specific as the predictor for performance in job interviews by research. (Boudreau et al., 2001). However, although there is increasing empirical evidence stated that personality attributes are correlated with the performance of the job interview, a psychological mechanism to identify relationships is nearly unknown.

In the field of Computer Science, in specific, deep learning with the characteristics that provide a broad spectrum of statistical methods that could automatically recognize human personality. Many methods had been proposed under personality computing for recognizing the personality trait from visual to audio. Several characteristics of human varying from clothing to body gesture, all contribute to the individual apparent personality's judgment. Nevertheless, until today, there is no persistent data corpus or benchmark models that had been introduced, which is one of the primary motivators of this research. A model that can recognize the human personality trait can bring much potential applications to society.

As of today, there are several state-of-art models had been proposed in the ImageNet Large Scale Visual Recognition Competition(ILSVRC), that can outperform in several computer vision task, particularly in the image classification task. In 2015, a human labeller had reported having a 5.1% error rate while performing classification in the ImageNet while at the same time, the ResNet (152 layers) had achieved an error rate of 3.57%. Transfer learning had been proposed to put those state-of-art models into practical use.

Transfer learning can improve the model learning through transferring the knowledge from a related task that has already learned. By adopting the transfer learning technique, the model can achieve better performance even without enough dataset.

1.2 Project Scope

This research strives to model a profiling technique that can examine a human's personality by using a short video taken from the candidates. The Big 5 personality model will be applied as the guideline in which the candidate's personality traits will be divided into five dimensions include conscientiousness, extraversion, emotional stability, openness to experience, and agreeableness. The dataset used in this research obtains from the apparent personality analysis at the ECCV ChaLearn LAP 2015 competition. Riding on the big data analytic trends, these studies will apply transfer learning by referring to the state-of-art model – ResNet 50, in order to perform prediction based on the visual modality from the videos. On the other hand, audio modality will be handled by the Artificial Neural Network model. After that, the model will be an ensemble and perform prediction. By fine-tuning the ensemble model, the model is expected to have significant performance. In the meanwhile, this research also emphasis on design of a pre-processing technique and fine-tuning model so that only minimum dataset required to get optimum performance.

1.3 Project objective

- **To design a data pre-processing technique based on the First Impression Rules (7-11 rules) for weakly supervised learning**
- **To design custom personality recognition model using spatial and temporal features.**
    a. **To train a personality recognition model with CNN using human upper body features**

   b. **To train a personality recognition model using acoustic cues extracted with Librosa**

- **To systematically tune the proposed neural networks architecture for improved classification**
   a. **To add drop layers on the ResNet-LSTM model to prevent overfitting**
   b. **To add more pooling layer to extract richer feature**
   c. **To add weight decay to accelerate the training process.**
   d. **To average result from multiple neural network to get maximum exposure of information**

1.4 Impact, significance and contribution

This research adopts multiple human characteristics features to model a profiling technique that can quantize the employee's or interviewee's personality score. The method proposes to help the organization to profile the interviewee so they can view the weakness and the strength for the following interviewee. After profiling the interviewer, the organization can make fast, strategic, fact-based decisions when it comes to personnel selection in order to find the best person to fit into the desired position. This model will add in the human upper body feature, including the human's face, body, and also their voice as the data corpus. The face represents complex multidimensional meaningful visual stimuli that can ease the performance of personality computing. Besides that, the dataset implements the pre-processing technique based on the First Impression (7-11 rule). The First Impression 7-11 rule stated that the human would build the first impression on another human within 7 seconds and form 11 conclusion on a person include the education level, economic level, ethnic background, and et al. Besides, with the inclusion of 7-11 rule, it allowed the dataset required to train the model become smaller while achieving optimum performance. The relationship of the modalities and amount of information exposure will be evaluated in this research with the purpose of minimizing information required for the model by adding the modalities while preserving the performance .

1.5 Background Information

1.5.1 Personality

In the last decade, many of the researchers and psychologists suggest that the range of lexical personality trait for a human can be summarized into five dimensions. **(**Caldwell & Burger, 1998). While there are many of the models that can be used to describe personality, one of the most adopted and most modern theoretical framework for determining the personality trait for the human is the Big 5 model. Conscientiousness, extraversion, emotional stability, openness to experience, and agreeableness are 5 of the dimension have been mentioned in the Big 5 personality model to describe people. Although the Big 5 model is being used, it not mentioned that whether which personality trait might be different from varied job position and each of the personality traits have a different degree of influence on the job search and performance. Once the personality of the interviewer entitled to be examined, a model can be built based on the analysis, subsequently provided a more trustworthy guideline for the Human Resource Management(HRM) to maximize the usage of talent in their organization.

1.5.2 Personality Assessment

A personality test is a selection procedure to measure the personality characteristics of applicants that related to future job performance. This set of analyses on the personality of the interviewer had been widely used by the global organization and operated as a primary operation process when making an employment decision. The use of this test also exhibits how a global organization views personality as a predictor for their employee's performance and behavior. Personality assessment can be categorized into two categories which are the objective assessment and projective assessment. For the objective type of personality inspection, the target will be given a set of questions with a set of limited externally provided responses that can help to examine the person's personality. This kind of personality assessment does not require the test-maker to observe on the test-taker but only relies on the scoring of the test-taker on the test paper. The test taker needs first to interpret the question and evaluate himself before answer the question. This kind of test sometime will generate bias since the observer might give a response on the test that supports him to get a more satisfying score to get the job.

For the projective type of personality assessment, the test-taker will be given a set of activities or tasks that require him to perform himself with practically no constraint and a minimal amount of guidance given by the test-maker to impose that the response remained natural. This kind of assessment requires the tester to identify the pattern or characteristics of the test taker. Some studies have shown that the video interview had to change the modify the way of this employment process. (Silveira Jacques Junior et al., 2019). Projective type of assessment also can be conducted in another way which the observation process is no longer needed to be conducted by the human anymore. In this respect, any technology involving in understanding, predicting and synthesis of human behavior by means of computational and deep learning techniques from a different source is identified as personality computing.

1.5.3 Convolutional Neural Network(CNN)

When coming to the image-recognition task, Convolutional Neural Network(CNN) is a well-known deep learning algorithm that can deliver exceptional performance on the image classification task. It consists of several computing layers include several convolutional layers and the pooling layer stacked up together. It is one of the best models that has been well-known as a powerful class of models for understanding the image content. CNN is very similar to the ordinary neural network, both of the made-up from neurons that have learnable weights and biases. Each neuron will receive some input, perform a dot product, and optionally follow it with a non-linearity. However, CNN is more suitable in image recognition tasks since it constraint the architecture in a more sensible way. CNN made up of three main layers: convolutional layer, pooling layer, and fully connected layer. The convolutional layer will preserve some of the spatial information regarding the image, and adjusting the learnable weight through the learning process. The final output obtains by multiplying the input with a weight that had been obtained through the learning process. In the last few years, some of the research on personality recognition had shown some inferred with reasonable accuracy. Figure below shows the underlying architecture for the CNN.

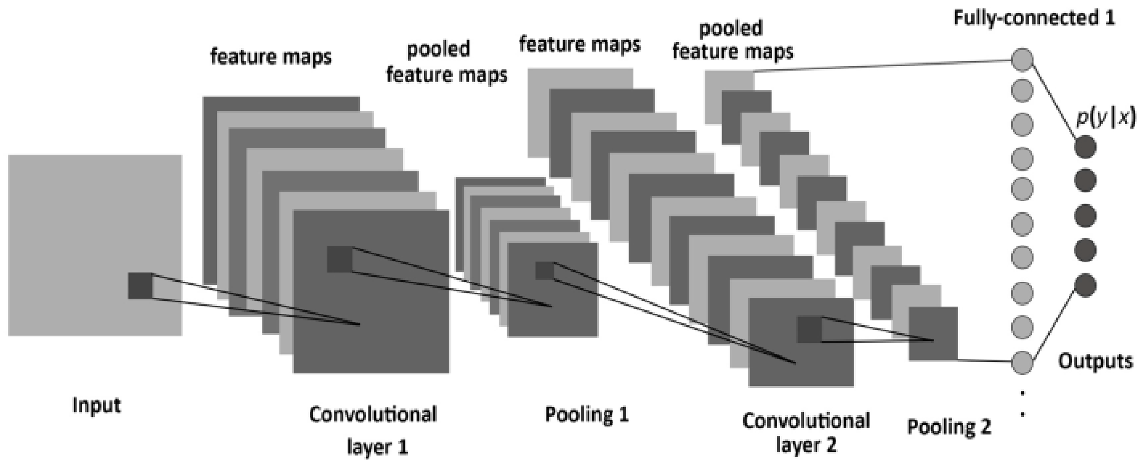


Figure 1-1 CNN's architecture

By utilizing these technologies, the organization can automate the personality assessment process by video processing, which will eventually enhance the personnel selection process effectiveness and efficiency. The company can convey the video interview with the interviewee, in the meantime process their personality. The global organization with tons of interviewees can use this technology to filter out some of the interviewees that are not suitable for the position.

Entire personnel selection's process can be prepared from pre-instructional information, observe patterns from the observer, finding the pattern exists within the process, scoring, transformation of scores concerning norms and other databases. (Jackson, 1985)

1.6 Report Organisation

The details of the research will be listed in the following chapter respectively. The literature review on personality model, personality computing, transfer learning and review on critical work on the European Conference on Computer Vision (ECCV) ChaLearn Looking at People (LAP) 2016 challenge are discussed in chapter 2. The research will also discuss on the methodology and describe the overview of the research's module in chapter 3. Chapter 4 will describe the implementation on the selective sampling technique on the video data based on the First Impression 7-11 rule and experiment on Audio-LSTM model. Chapter 5 reports the conclusion for this research.

Chapter 2 Literature Review

Personality had been conceptualized from a variety of theoretical perspectives and various levels of abstraction or breadth (Pervin & John,1999). It has been stated as the most significant aspect that concern by researchers and psychologists, as it always been used as the predictor for many aspects of human life such as the job's performance or academic success. (Azucar, Marengo and Settanni, 2018). The self-reported questionnaire has become one of the well-accepted ways of evaluating one's personality. However, with the advancement of technologies, machine type of personality computing was introduced for understanding the personality trait for an individual. Numerous personality computing's methods had been proposed in the past for studying the individual's personality trait from the nonverbal aspect of verbal communication, body movement (Pianesi et al., 2008), combining acoustic with visual cues or physiological with visual cues (Abadi et al., 2015) and face expression (Biel, Teijeiro-Mosquera and Gatica-Perez, 2012). The first research in personality computing had been conducted by Vinciarelli and Mohammadi, which centered on three main objectives, automatic personality recognition synthesis, and perception. After having the foundation for personality computing, the personality computing adopted the term of apparent personality analysis, personality impression, or only first impression to refer to personality perception. The first impression is interpreted as the rapid evaluation of personality traits and the social status of an individual. (Ambady, Bernieri and Richeson, 2000). The apparent personality computing is to measure how similar the outcome made by the machine to the rating given by the labeller to a specific individual. It is not expected to predict the actual/real personality but is the general idea of how people label the individual.

2.1 Personality Model

Human personality had been a significant aspect in the psychological world, many of the theories had been proposed throughout the decade to find a way to categorize it. Some researchers state that the trait was the most adequate way to classify and measure a different aspect of human personality. Many of the traits models had been built with human judgment about semantic similarity that people used to describe themselves. There are many models exist to categorized the human personality trait while using different terms and definition. The Big 5 model adopts the premises of trait theory, which indicate that each of the people can be categorized by individual differences that are stable over time, consistent across the situation, and involve patterns of thought, affect, and behavior (White, Hendrick & Hendrick, 2004). This model also is the primary reference throughout this research to measure and quantized the personality trait. The big 5 model organized from broad constructs which is the 5 dimension of the personal trait to specific constructs including the facets of each factor. Table 2.1 show the 5 dimension of the personal trait and the facet along with each of the personality type.

Big 5 traits	Personality facets
Openness to experience (openness)	Idea, Fantasy, Aesthetics, Action, Ideas, Values
Conscientiousness	Competence, Order, Dutifulness, Achievement striving, Self-discipline, Deliberation
Extraversion	Warmth, Gregariousness, Assertiveness, Activity, Excitement seeking, Positive emotions.
Agreeableness	Trust, Straight-forwardness, Altruism, Compliance, Modesty, Tender-Mindedness
Neuroticism	Anxiety, Angry hostility, Depression, Self-consciousness, Impulsiveness, Vulnerability.

Table 2-1 Big 5 framework for personality trait

Other than the Big 5 model, there also another model that can be used to measure personality traits but with different definitions and terms. For example, the Myers-Brigg Type Indicator is also one of the well-known personality assessment models which consist of 16 personality type and organized into four-letter abbreviations.

Despite the universality of the MBTI, some of the research indicates that MBTI cannot be supported. Many of the fundamental psychological standards had been neglected by the MBTI rule. The researchers claim that there is no visible evidence that there are 16 unique categories that can fit all the people. They suggest that the MBTI standard is no suitable for the career planning type of personality assessment. (Pittenger,1993) Dominance, Influence, Steadiness, Conscientiousness(DISC) is another type of personality trait model which includes the four-factor trait. It was developed by William Marston, who was a psychologist in the early 1900s. DISC is a useful personality trait model, but it has not been studied as often as another personality model such as Big 5 and MBTI, meaning that it has less number of research that can support it.

2.2 Personality computing

Personality computing can be conducted by using varied sources such as text, social media, and video. Many of the methods of deep learning had been used to identify the human's personality. Some of the research use user's digital footprint as a predictor (Azucar, Marengo & Settanni, 2018), some of the researchers perceive the text written by the user (Majumder et al., 2017), some of the research tries to capture face feature from a person to identify their personality.

The personality recognition visual-based model usually will emphasize on the human's face feature, generally combining features at a different level and their relationship. (Junior et al.,2018). For the image-related task, Convolutional Neural Network is a deep learning algorithm which can take an image as input. CNN requires lower pre-processing steps compared to the other classification algorithms, which made it extremely suitable for an image classification task.

Research infers the social impression using the human facial expression feature through the integration of low, intermediate, and high-level perception image feature came out with a hierarchical model. It identifies the facial landmarks first and extracts the low-level features from the face before applying it to recognize a set of mid-level attributes such as the nameable attributes (e.g., gray-haired), the part appearance types, and their relative geometries. The social dimensions are then determined by all of these attributes. (Joo, Steen & Zhu, 2015) Similar research also conducted to study the relationship between the facial features and the personality trait with the difference between the low-level feature are extracted from different face regions, each with different feature extraction methods. (Yan et al., 2016)

With the advance of technology, social media also become one of the data corpus for apparent personality recognition. The research examined the relationship between how the user manipulates their profile image by applying the filter on social media and their personality trait. Their research face difficulty since some of the user does not apply the filter on their profile picture. (Ferwerda, Schedl & Tkalcic, 2016) Liu et al also use more than 66000 users' profile images as the input to determine their personality based on their tweet. (Liu et al,. 2016). In 2014, the research examined the Big 5 personality trait of the Facebook user based on their Facebook profile image. (Celli, Bruni & Lepri, 2014) The profile image is labeled with the gold standard self-assessed personality and interaction style labels. After collecting the data, they applied a different kind of machine learning approaches such as Naives Baues, support vector, and logistic regression, to examine the effectiveness of the feature to measure the personality trait score.

2.3 Transfer learning

Deep learning had been applied to various computer vision applications, especially in image classification using Convolutional Neural Network(CNN). CNN can make use of the massive amount of labelled images for training a robust model. However, to process such an amount of dataset, a tremendous amount of computing power is required. Hence, different deep learning techniques had been proposed to apply the CNN approach to practical use. Transfer learning is one of the deep learning techniques that adapt a trained model that had been trained on another large but different dataset to a specific problem where the dataset had a lesser amount of dataset. ImageNet is a project that focuses on computing research has organized ImageNet Large Scale Visual Recognition Challenge(ILSVRC). ILSVRC had dominated the CNN and deep learning techniques since 2012(Rosebrock, 2020) and become the benchmark for the computer vision classification algorithms. Many of state-of-the-art pre-trained network had been proposed on ILSVRC, which indicate substantial capability on generalizing images on almost all-kind of a dataset through transfer learning by performing feature extraction and fine-tuning the network.

2.3.1 Comparison between different pre-trained model

In 2012, AlexNet had outperformed other models in ILSVRC, which also was the first deep CNN that outperform in the classical image/object detection task. AlexNet consists of a total of 8 learned layers, which include five sequentially connected convolutional layers, each with decreasing filter size and three layers of fully connected layers.

In 2014, Simonyan and Zisserman had introduced the VGG network, use only a small filter (3*3) that allowed a deeper network to be trained (Simonyan & Zisserman,2014). VGG can reach 16, and 19 layers deep, which were considered very deep network in 2014.VGG reduces spatial dimension by applying max pooling to reduce the spatial dimension by half while doubling the channel. A stack of convolution layer is followed by the fully connected layer. VGGNet had achieved an error rate of 6.8% and stood as runner up in ISLVRC in 2014. Unfortunately, VGG is much slower to train compared to other models in ILSVRC. Besides, the VGG's network architecture weights are considered heavy, while there are only 16 layers. GoogLeNet also was one of the proposed models in ILSVRC 2014, achieved a top-5 error rate of 6.67%, which outperform VGGNet in 2014 ILSVRC.

At the ILSVEC 2015, Residual Neural Network(ResNet) had been proposed by Kaiming He et al. (K.He etal., 2016) After the success of the AlexNet and VGG net, Residual Network had become one of the revolutionary work in the computer vision field. By applying the concept of 'skip connection', the ResNet able to train an enormous amount of layer compared to the previous proposed model, while still achieving excellent performances. Since the ILSVRC 2014, the proposed architecture going deeper (152 layers deepest) while VGG network with 19 layers while GoogLeNet having 22 layers respectively. The core idea of the ResNet is introducing a 'skip connection' that can skip one or more layers that allowed the ResNet to extend the neural network to 152 layers while having lower complexity compared to VGGNet. Figure 2-2 below showed the residual block.

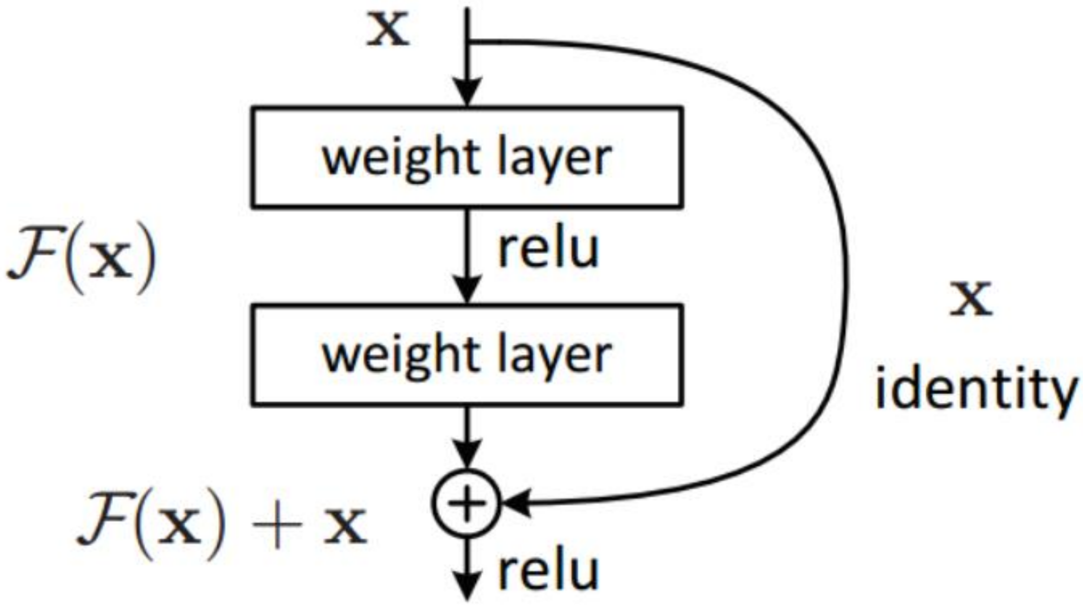


Figure 2-2 Residual block ResNet

Neural Networks are universal function approximates as the increasing number of layers indicate the increase in accuracy. But that not always the case as the three will be a limit of the number of layers added. However, vanishing gradients and curse of dimensionality problem exists as the neural network gets deeper. The deep neural network may not able to learn simple functions such as identity function. Hence, the residual block had been proposed to solve this problem. The residual block which also known as the identity shortcut connection, having the ability to overall trying to learn the true output. The residual network is built by adding those skip connections into the plain block (normal CNN block). Skip connection also allowed uninterrupted path for back-propagating gradient.

2.4 Review on related work

This research is dedicated to model a profiling technique that can quantify the apparent personality. The dataset is collected from the ECCV ChaLearn LAP 2016 challenge which consisted of a single track competition to quantitatively evaluate the recognition of the apparent Big Five personality traits. The data set has a total of 10000 clips which came from 3000+ different videos obtains from YouTube. The video collected from YouTube consists of 2 main characteristics that are facing on the camera and speaking English. The video picked from YouTube consists of people who are different in gender, age, nationality, and ethnicity to obtain the recognition of personality traits to be more interesting.

There is a total of 85 teams registered in the ECCV ChaLearn LAP 2016. Among all of the winning teams, the majority of them consider both audio and video modalities when constructing their model. Convolutional Neural Network (CNN) was the model used almost by all of the team in this event to exploit the visual modality.

Chen et al. team(NJU-Lambda) had achieved the best result for the competition with the regression accuracy of 0.913. In their research, they proposed a Deep Bimodal Regression (DBR) framework to model both visual and audio features. They modified the traditional CNN by exploiting critical visual cues. Most of the teams who participated in the competition had made separation within the background and face of the observant on the video data, but they allowed to get the best performance without considering the semantic assumption about the data. They extract a total of 100 frames from each of the training videos and model it using the Descriptor Aggregation Network(DAN). DAN is a modified version of CNN architecture by removing the fully connected layer and replace with the average and max-pooling layer. They applied early fusion for the epoch fusion to boost the DAN’s performance.

Arulkumar et al. (team evolgen) had proposed another different solution for the apparent personality recognition. In their research, they propose two bi-modal deep neural network architecture, one of the branches will be managing the visual feature while another managed for the audio feature. In their visual data pre-processing stage, they adopt the semantic assumption on data by taking the 3D aligned segmented face images as the input for the visual model. They indicate that by separating the face from the background, the model can be more accurate. While pre-processing the data for the visual model, they randomly select N input frame from the different partition of the raw video at the combination of input frame will be different in every epoch. After the input image had been selected for the following epoch, the extract the face feature of the observant and centring it in the middle of the final input image. The total frame extracted for the model was 200 frames.

Gucluturk et al. (team DCC) had proposed their audio-visual models for this dataset. Their model network consists of separate auditory and visual streams, each with a 17-layer of the deep residual network. The network is then followed by the audio-visual stream with one fully connected layer with hyperbolic tangent activation. The pre-processing method they're performed in the research by extracting a random frame (total 100 frames) from the video and fed to the model while training.

The majority of the approaches from the best performance model proposed in this challenge used both audio and video modalities. Convolutional Neural Network(CNN) was the dominant approach applied to video modality to learn the representation. Next, late fusion was applied to the multi-model methods before fed to different regression methods. Most of the team also made a semantic assumption about the data by separating it from the background but the winning team (NJU-LAMDA) did not apply those assumptions into it. Besides, for each of the top entry in the ChaLearn LAP 2016, the amount of frame extracted for each video are enormous. Table 2-2 show the summary of a method for the top teams that participate in the ECCV ChaLearn LAP 2016.

Team	**Pre-training**	**Pre-processing**	**Audio modality**	**Visual modality**	**Fusion**
NJU-Lambda	VGG-face	Extract 100 frame	Neural network	CNN	Late
Evolgen	-	Face-alignment, extract random 6 frame	RCNN	RCNN	Early
DCC	-	Extract random frame	ResNet	ResNet	Late
Ucas	VGG, AlexNet, ResNet	Face alignment	Spectral	Partial Least Square Regressor	Late
Pandora	-	Face alignment	LLD	CNN	early

Table 2-2 Review on participant's work on ECCV ChaLearn LAP 2016

Chapter 3 Methodology

3.1 Overview

This chapter will emphases on describing the general work procedure for this research. This research focuses on profiling the personality trait of an individual. The profiling of personality traits for the target achieved by recognizing the features available in a short video session. In this research, few models were proposed to model the audio and visual modality for predicting the apparent personality traits. Some of the models will only adopt certain modality. To determine the performance of each modality, each model is evaluated respectively. The personality score will be categorized using the Big 5 personality model which includes the conscientiousness, extraversion, emotional stability, openness to experience, and agreeableness. The input for the model will be a raw video with an only single person. The chapter include four topics which are Data-Collection, Data-Pre-process, Modelling, and Evaluation.

3.2 Data collection

The competition initially consists of 13,935 YouTube video and the organizer had manually filtered out some video that does not reach the requirement (speaking English and facing camera). After the filtering process, the remaining video had been used to generate a total of 32,139 clips with 15 seconds for each of the clips. They were also done another manual filtering which filters out the video with satisfying requirement (only a person, not much camera movement, no advertisement) and remains total 10,000 clips came from 3,060 unique videos. After collecting the video data, they also decide to use the human labelling of videos. To collect the label for the dataset, they used the Amazon Mechanical Turk (AMT). To reduce variance for the data, every video in the dataset had been giving multiple votes by the AWT worker. A custom interface with the videos and the question to define the personality score had been prepared for the AWT workers. Each of the AWT workers responsible for small batches of pairs of videos. The data can be obtained from the official website of the ChaLearn. There is a total of 6000 training videos, 2000 validation videos, and 2000 testing videos respectively, each video has a total length of approximately 15 seconds. Figure 3-1 shows the custom interface for data collection.

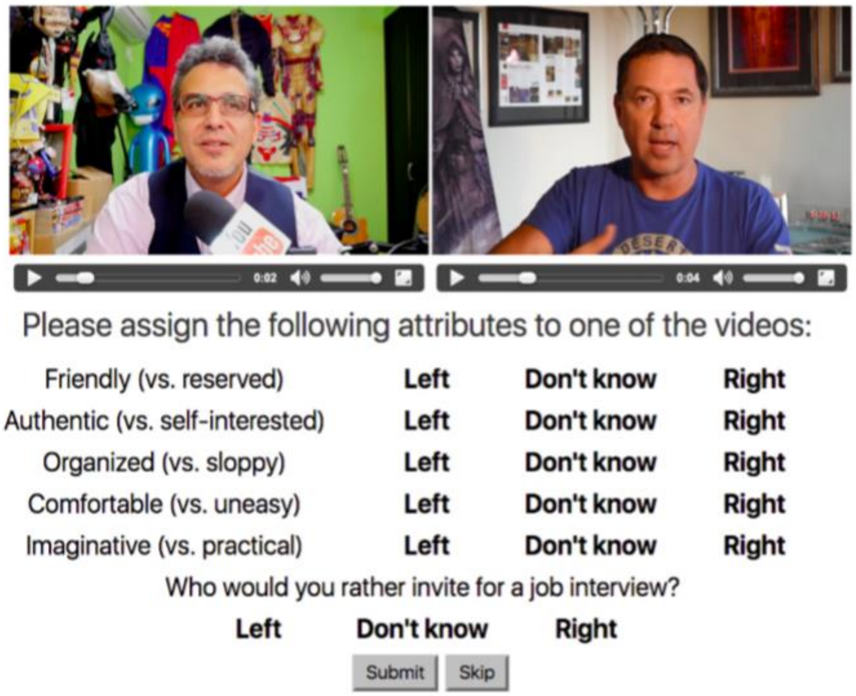
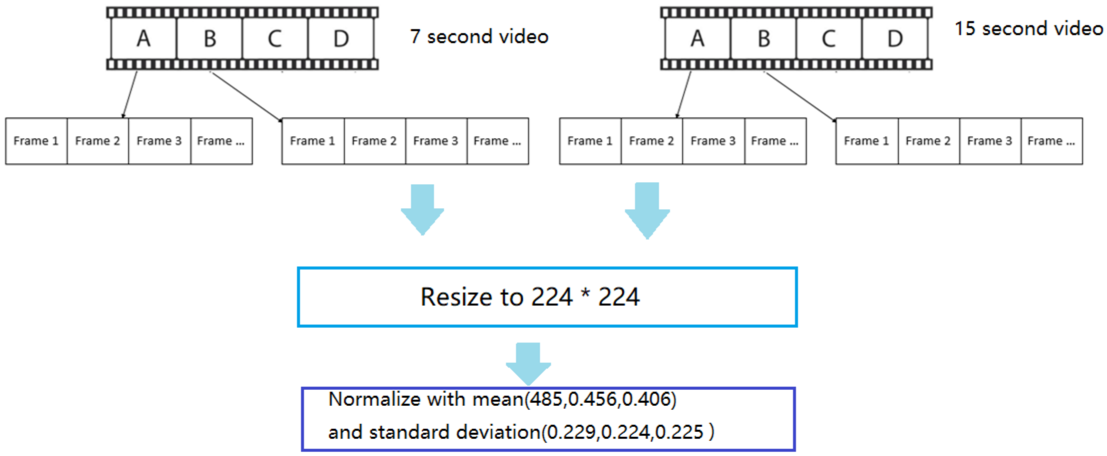


Figure 3-1 Custom interface for data collection

3.3 Data-pre-process

3.3.1 Visual feature pre-processing

The YouTube clips obtain from the ECCV 2016 need to be pre-processed to the desired format first before modelling it. The video data can be downloaded from the ChaLearn official website. The First Impression 7-11 rules are applied during this stage by cropping the video into 7 seconds. To verify the validity of the 7-11 eleven rule, another copy of the video data with 15 seconds will also be used in the later stage. N-number of images will be extracted from the videos (7 and 15 seconds). Each of the images extracted from the video will be resized with appropriate image resolution before passing to the profiling stage. All the images will be normalized by the mean (0.485,0.456,0.406) and variance (0.229,0.224,0.225) in each batch during the training to help to accelerate the training process by averaging the data within a range and reduce the skewness. Figure 3-2 shows the description of visual features processing in this research.



Image= image -mean/std

Figure 3-2 Data pre-processing on video data

3.3.2 Audio features pre-processing

The audio signal extracted from the videos are transformed into WAV-format in order to extract the audio feature from it. For each of the videos, the mean of certain audio properties of audio signal will be extracted. There are total 26 hand-crafted features extracted. Table below shows the features extracted from the audio signal.

Audio feature	Description
Chroma feature	A 12-element feature vector indicating how much energy of each pitch class
Root mean square energy	Root mean square energy for each frame, which represent the magnitude of the signal

Spectral(centroid, Bandwidth, roll off)	Spectral feature obtain by converting time based signal into the frequency domain using Fourier Transform
Zero Crossing Rate	Rate of sign-change along the signal which indicate the rate of signal change from positive to negative and back.
Mel-Frequency-Cepstral Coefficient(MFCC)	Short-period power spectrum of sound wave representation.

Table 3-1 Audio features extracted

3.4 Model Overview

The proposed (Audio-LSTM) model can be divided into three part. It comprises a Multilayer Perceptron(MLP) handling the audio features, a fine-tuned CNN (ResNet 50) modelling the visual features and a Long Short Term Memory(LSTM) model to handle the spatial-temporal information. Each video is divided into N-number of images and model by the visual model (ResNet 50 and LSTM model) while 26 extracted audio features will process by the MLP. Late fusion is applied to ensemble the output from audio and visual model. Figure below show the Audio-LSTM structure.

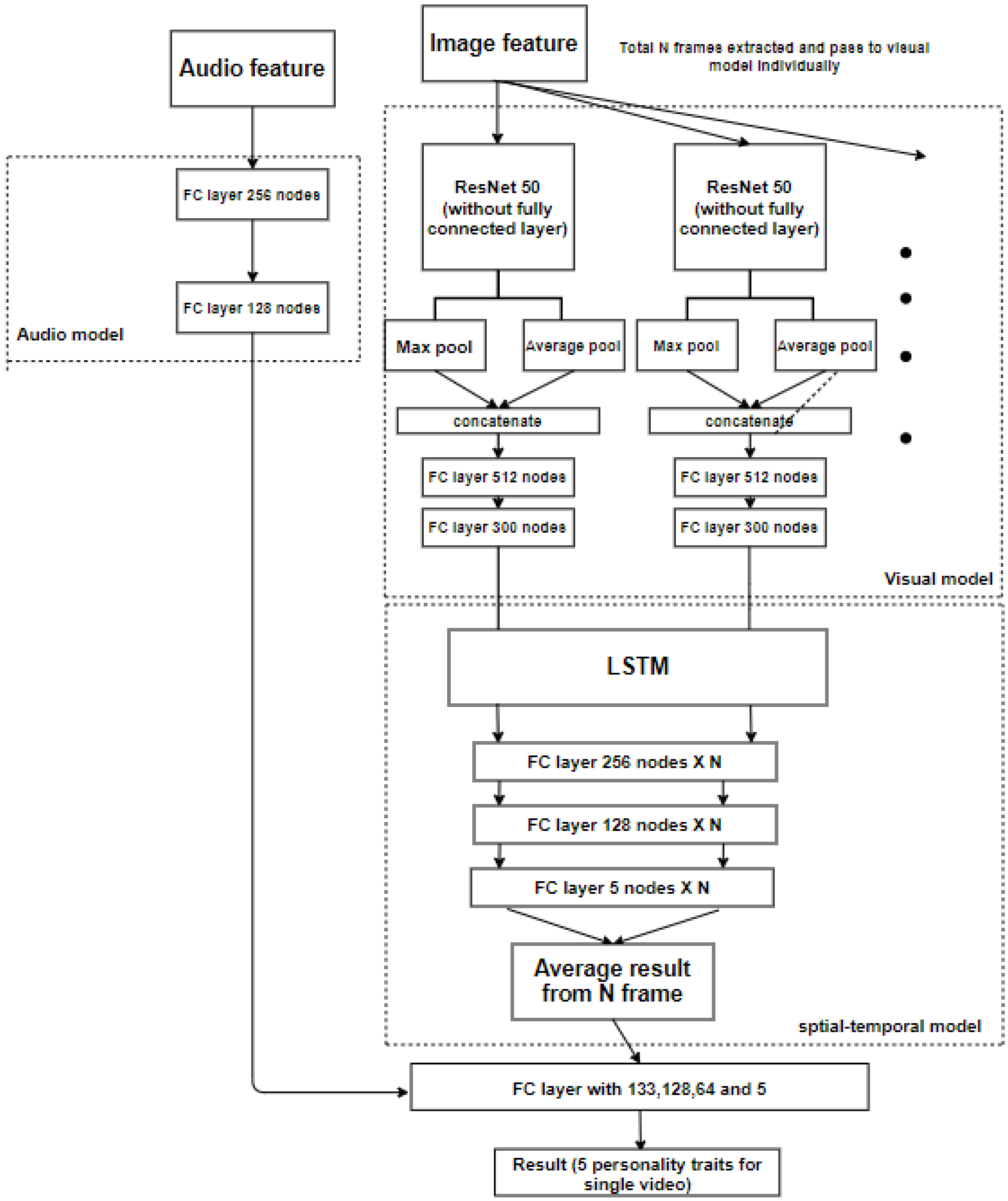


Figure 3-3 Audio-LSTM detail architecture

3.4.1 Fine-tuned ResNet50

ResNet 50 had been selected as the pre-trained CNN model for this research. The architecture of the ResNet 50 shown in Figure below.

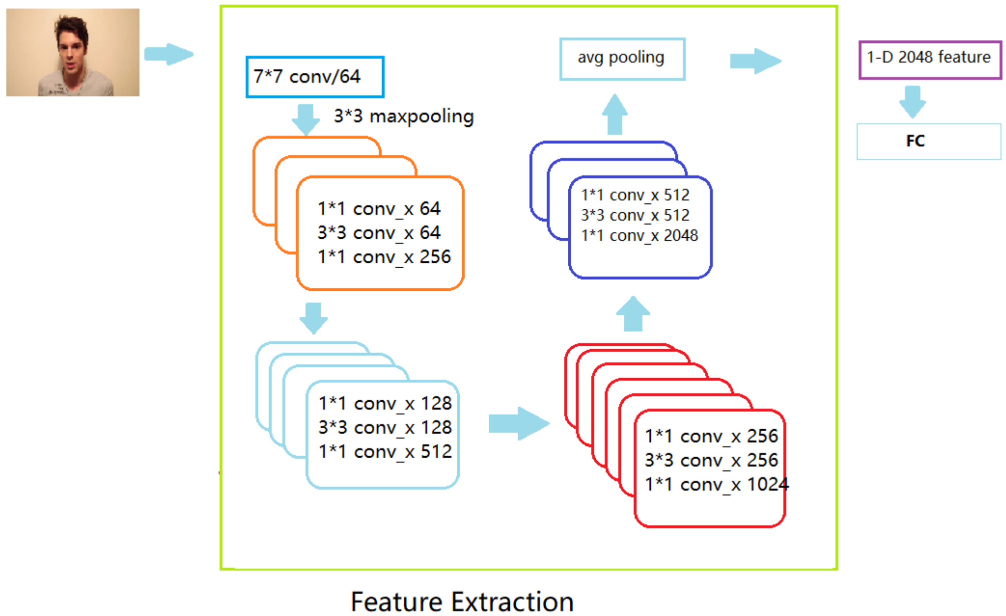


Figure 3-4 ResNet 50 architecture

To further enhance the model performance, the ResNet 50 model is further fine-tuned. The fully connected layers and averaging pooling layer of ResNet 50 are removed and replaced by both the average and max-pooling layer. After obtaining total of 2048 features from each of the pooling layers, the layers are then flattened and concatenated to form a linear layer. The layer is then connected to fully connected layers with 512-dim follow by 300-dim to weight each of the features. This allowed the modal to obtain richer features to be used in the later stages. Each fully connected layers are followed by the batch normalization (momentum=0.01) and drop-out (0.5). Figure 3-4 shows the fine-tuned ResNet architecture.

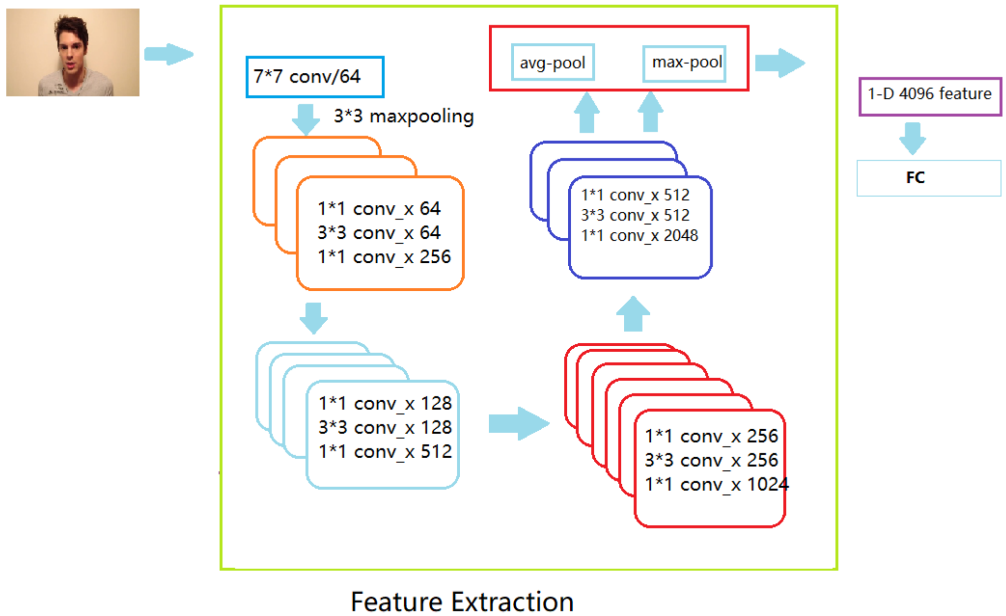


Figure 3-5 Fine-tuned ResNet 50 architecture

3.4.2 LSTM based model

LSTM based model is designed to learn the spatial-temporal information of the visual features in order to reduce overfitting problems. The architecture of the LSTM model is shown in Figure below.

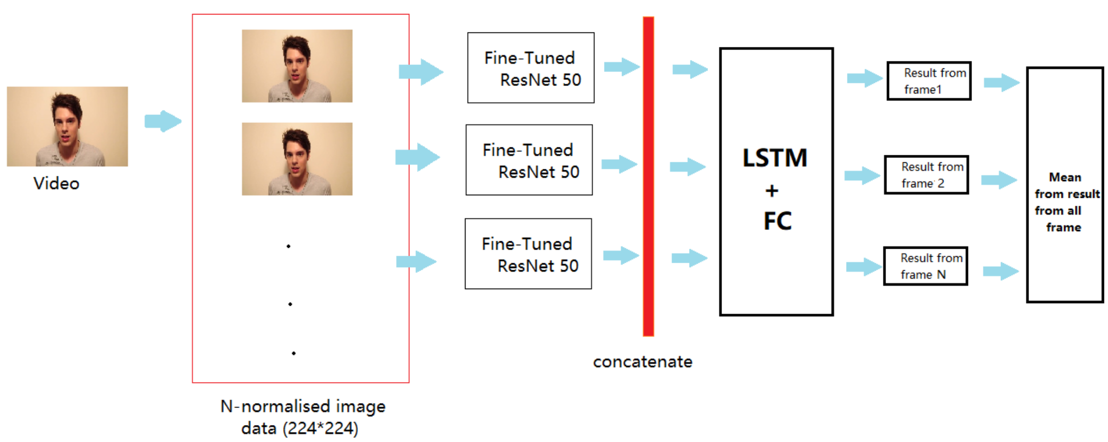


Figure 3-6 ResNet-LSTM model proposed

N-number of images from a single video will be model by the ResNet50 model. After all of the features are captured by the fine-tuned ResNet50, the weighted feature extracted from the ResNet's fully connected layers are passed to the LSTM model. The input size for the LSTM model will be 1*300 which followed by the last fully connected layer of the ResNet Model. The series of visual features from each image extracted will then pass to the LSTM model. The LSTM generates output for each of the frames extracted for a single video to maximize the information exposure. LSTM model is connected with fully connected layers with 256 nodes and followed by the ReLu and Drop out function. The output layers for the visual model will be a 5-dimensional fully connected layer followed by the sigmoid activation function. The LSTM model will process each of the images accordingly for a video and averaging N-frame from a single video to get the maximum amount of information exposure. To preserve the performance of the model, N is fixed as 15.

3.4.3 Audio Multilayer Perceptron(MLP)

There is a total of 26 features extracted on the data-pre-processing phase so the input for the Audio ANN will be 1*26. The MLP proposed will have a total of 2 layers of a fully connected layer (256,128) with an output layer of 128 and each layer is followed by the batch normalization (momentum=0.1) and ReLu activation function. Drop-out is applied to the fully connected layers as well to prevent overfitting. (drop out=0.5).

3.4.4 Audio-LSTM model

The Audio-LSTM model will then combined the fine-tuned ResNet 50 model, Audio MLP model and LSTM model. Output from the visual and audio models are ensemble and follow by fully connected layer with total 2 layers, each with 128,64 hidden nodes. The output layer for the final model will also be a 5 dimensional fully connected layer with sigmoid activation function. The overall structure is shown in Figure 3-6.

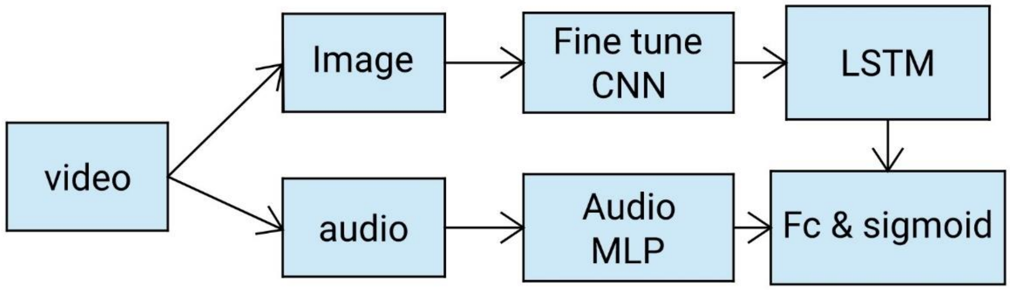


Figure 3-7 Overall structure of Audio-LSTM architecture

3.6 Research Tools and Technologies implementation

Librosa library will be used to extract audio feature mentioned in this research while OpenCV library will be used to pre-process the images. The other implementation detail is show at below:

Python Library	**Version**
Keras	2.1.3
Tensorflow-gpu	2.0
Image	1.5.27
PyTorch	1.1.0
Scikit-video	1.1.11
Scikit-image	0.15.0

Table 3-2 Library for personality model

All of the experiment conducted using the Google Colab. Table shows the settings for the Google Colab.

Specification	Description
Processor	2.3 GHz Intel Xeon Processor
GPU	NVIDIA Tesla K80(2496 cores)
GPU RAM	12 GB

Table 3-3 Google colab settings

The code and implementation guideline is available at https://github.com/ZackJye/Fyp-repository. The repository contains the source code of the model, dataset's link and other useful utilities function that helps to train and pre-process the model.

3.7 Evaluation metric

The performance of the model is evaluated by using 1- mean absolute error. The formula to measure the accuracy is state below.

$$Mean\ accuracy = \frac{1}{5N}\sum_{x=1}^{5}\sum_{i=1}^{n} 1 - |ground\ truth(i,j) - predicted\ value(i,j)|$$

The range of the result from the mean accuracy is from 0 to 1 which 1 indicate the best result and 0 indicate the worst result. N is the number of test video in the dataset and the $ground\ truth(i,j)$ represent the actual value for personality trait for the following image and the $predicted\ value(i,j)$ will be the predicted value that calculate from the model. Thus, this personality model will predict 5 output, consists of personality trait scores for 5 dimensions respectively. The mean accuracy's result will measure the correctness for five of the traits. This evaluation formula is used as it is the standard model that other participant who participated in ECCV 2016. The accuracy of each of the train will also be calculated using the same formula.

Chapter 4 Experimental Setup and Result

4.1 Exploring First Impression Rule

First of all, all of the datasets are downloaded from the official website of ChaLearn, which consists of a total of 75 zip files for training data and 36 zip files for validation data. The download process was completed by using the command line. After extracting all of the zip files, there was a total of 6000 video data for training data and another 2000 validation video data. Initially, all of the videos in the dataset has a total length of 15 seconds. Another copy of the training video data and validation video data of 7 seconds was created to practice the First Impression 7-11 rule. The training video data had been cut into frames. Each of the video data (both 7 second and 15-second video data) cut into 7 frames per video.

Two models had been created in this stage in order to verify the First Impression 7-11 rule. One modal was using the 7-second video data as input while another using 15-second video data to compute the model. The pre-train model ResNet50 had been used to train this model. Both of the 15-seconds and 7-second models apply the same setting by using 3000 training videos and 1000 validation video data to evaluate it. There was a total of 21000 images had been loaded into the modal to train and a total of 7000 images to evaluate the performance (7 frames per video). All of the image shapes had been resized into 224*224*3 to fit with the ResNet 50 model. The image data had been converted into a Tensor record format to make the retrieving and storing of these image data to become faster. The model applied Adam optimizer with a learning rate of 0.001 and 20 epochs for each of the models with 32 batch size. Same parameters and configuration were applied to both the 15-second model and 7-second model.

4.1.1 Evaluation

Both of the models (7 second and 15 second) had been evaluated by the validation data using the mean accuracy method. The result for the 7-second personality model on the validation data is around 87.38% while the result of the 15-second model is around 85.84 %. The accuracy of the model had been improved by 1.53% by using 7-second video data rather than 15-second video data. The 15-second and 7-second models loss is shown in Figure 4.2 and Figure 4.3

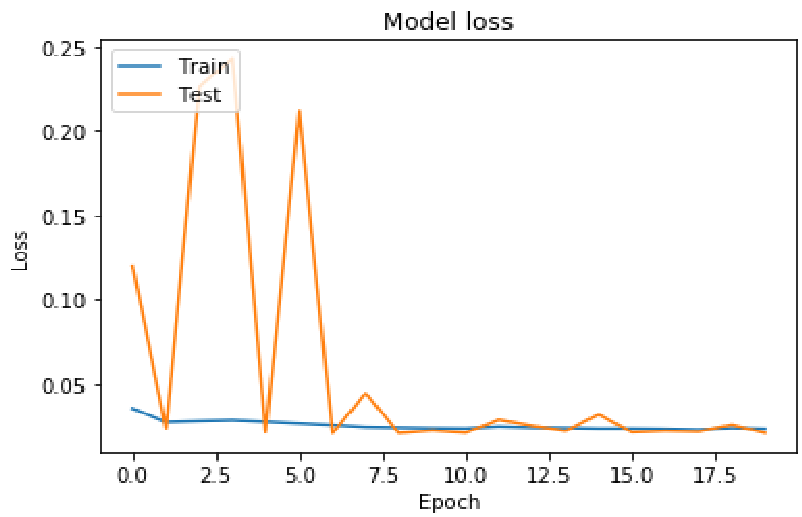


Figure 4-1 7-second model train and validation loss

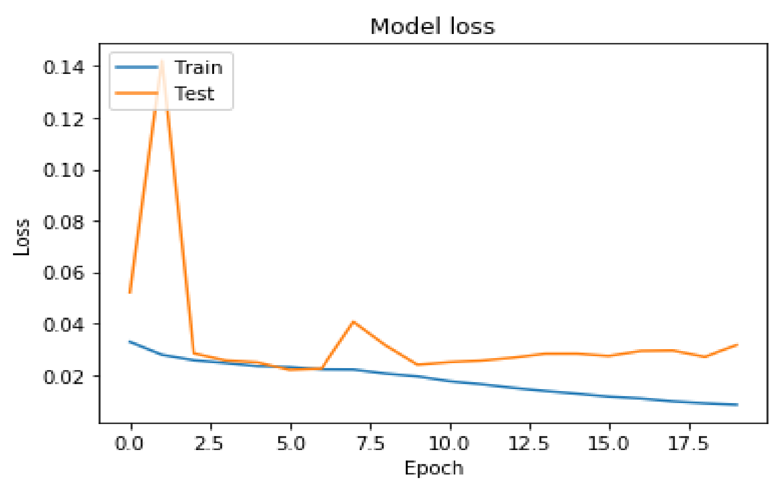


Figure 4-2 15-second model train and validation loss

From Figure 4-3, the modal's testing lost increased after the ten epoch while the training lost keep decreasing, which was a sign of overfitting. In comparison, the 7-second model does not have this issue, although the same parameters and configurations were applied in both of the models. Hence, this experiment concludes the model performance can be increased, and accuracy can be improved by applying the sampling technique through using the first seven-second of the video data on training.

The result obtains on the validation set also compared with the proposed model by another participant in ECCV ChaLearn LAP 2016. The result of the current personality model still no satisfying compared to the model proposed. Hence, the personality model proposed in this implementation will be improved by implementing the audio and spatial feature on the model.

4.2 Exploring personality recognition model architecture with different features

This sub-chapter lists the settings for the experiment conducted on the fine-tuned ResNet 50, ResNet50-LSTM, Audio-LSTM. Audio-LSTM model proposed in this research with the ability to model audio, visual as well as the spatial-temporal feature from video data. To verify the performance of added modality, a few experiments had been conducted.

In this phase, the images were extracted according to the frame per second from each of the videos. As a result of the 7-second model outperform 15-second model, N frames were equally extracted from the first 210 frames of the video (consider each of the videos has only 30 fps). All implemented model was evaluated using the validation set, containing a total of 2000 videos and total of 30000 images extracted from the videos. The table below shows the overall result for all the test performance for the model trained for all of the modality.

Model	Accuracy
Audio	86.10%
7-frame Fine-tuned ResNet	87.38%
15-frame Fine-tuned ResNet	88.72%
7-frame ResNet LSTM	89.95%
15-frame ResNet LSTM	90.13%
15-frame Audio-LSTM	90.30%

Table 4-1 Overall accuracy for all model

4.3 Exploring visual features using Fine-tuned ResNet50

As mention earlier, the fine-tuned ResNet 50 model used to model the visual features from the videos. 15 and 7 frames had been cropped from the raw video file and used in this phase. The frames cropped equally from the first 7 seconds only since First Impression Rule had been verified to improve the performances.

Total 90000 images used as the training image data for the fine-tuned ResNet 50 with a total of 15 epochs with 32 batch size, Adam optimizer with a learning rate of 0.001, weight decay of 0.00001. Figure 4-3 and 4-4 show the training accuracy and loss for fine-tuned ResNet 50 models.

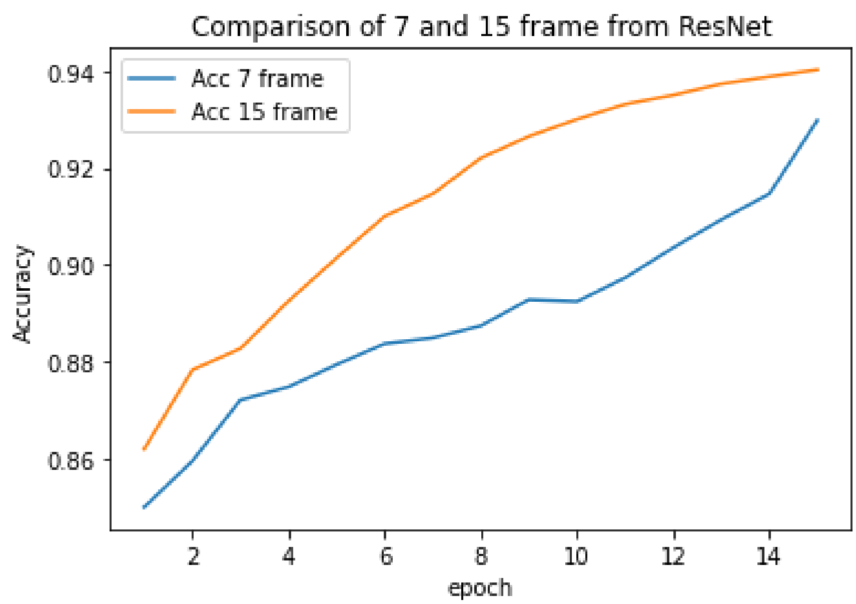


Figure 4-3 Training accuracy for 15 and 5 frame ResNet model

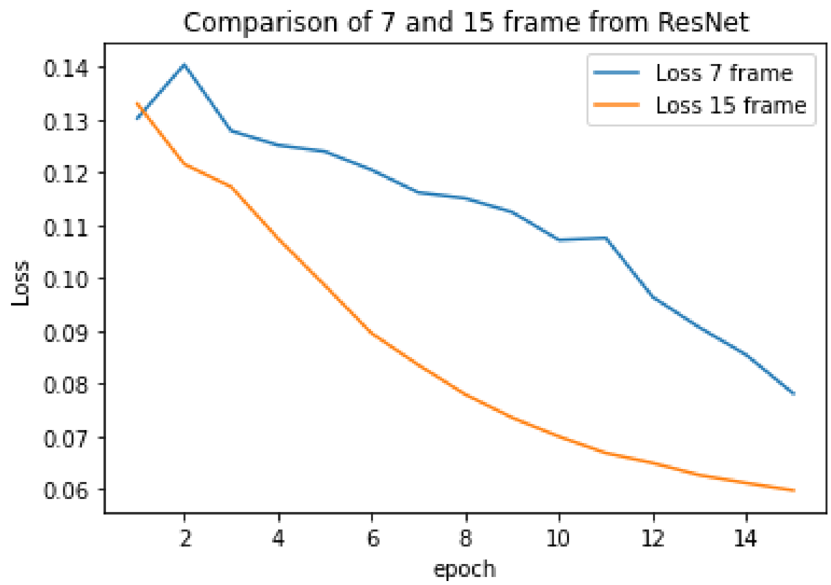


Figure 4-4 Training Loss for 15 and 7 frame ResNet model

4.3.1 Result and evaluation on visual feature

The evaluation result of the 7 frame model and 15 frame model was 87.38% and 88.72%. After comparing the training accuracy and loss of both to the model, the training can be more smooth when more frames per video were added into the training. Besides, significantly different between the training accuracy and validation accuracy can be observed from both of the models. This can be interpreting as an overfitting issue since the mean accuracy between the test performance and training performance was huge (around 4~5%). Other than that, by increasing the amount of information exposure, the performance allowed to increase significantly when only visual features were included.

Hence, to minimize the overfitting issue, spatial features was added-in to model the spatial-temporal pattern of the visual features.

4.4 Exploring visual-spatial-temporal features using ResNet

For the ResNet-LSTM model, the images for a video were encapsulated (15 images per video) and passed to the trained fine-tuned ResNet model. All the layers in the trained fine-tuned ResNet model were frozen during the training. This result only the LSTM model was trained to evaluate the effectiveness of spatial-temporal features in apparent personality recognition. Adam optimizer was used with a learning rate of 0.001, a weight decay of 0.00001. The training batch size was shrink to 12 due to the limit of computation power with the 10 epochs. In this section, the result was evaluated with varying the amount of data in visual-spatial-temporal modalities. The same settings were applied to another model with 7 frames extracted per video for evaluation.

Figure 4-5 and 4-6 show the training accuracy and loss while training the7 frames ResNet-LSTM model.

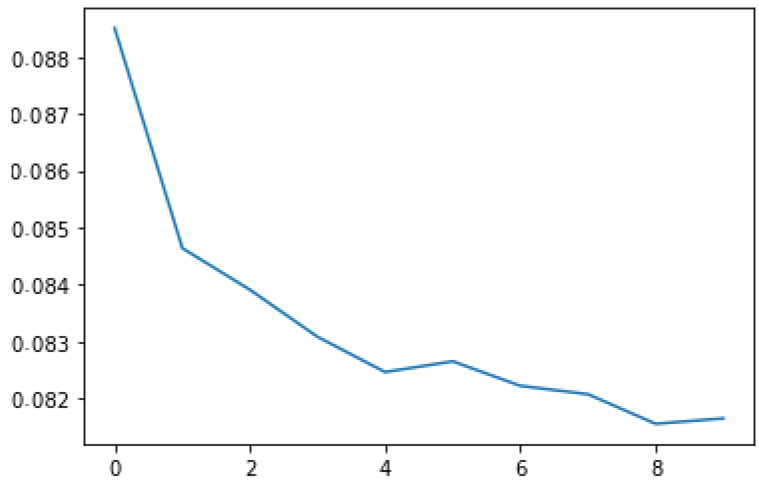


Figure 4-5 Training Loss for 7 frame ResNet-LSTM model

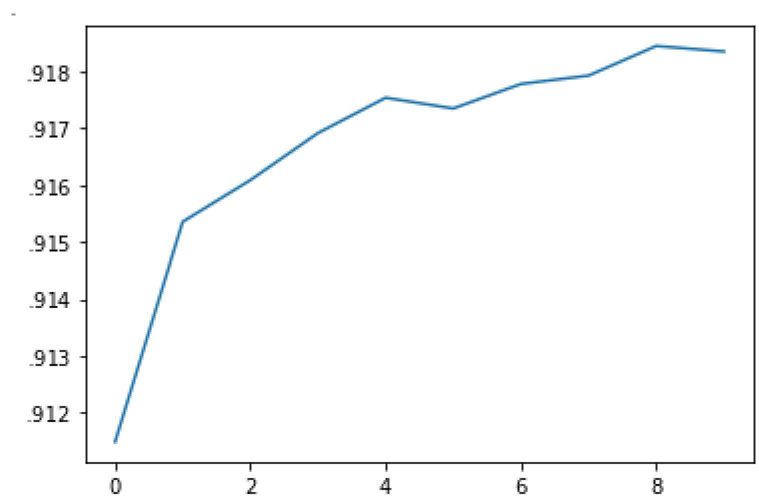


Figure 4-6 Training Accuracy for 7 frame ResNet-LSTM model

Figure 4.7 and 4-8 show the training accuracy and loss while training the 15 frames ResNet-LSTM model.

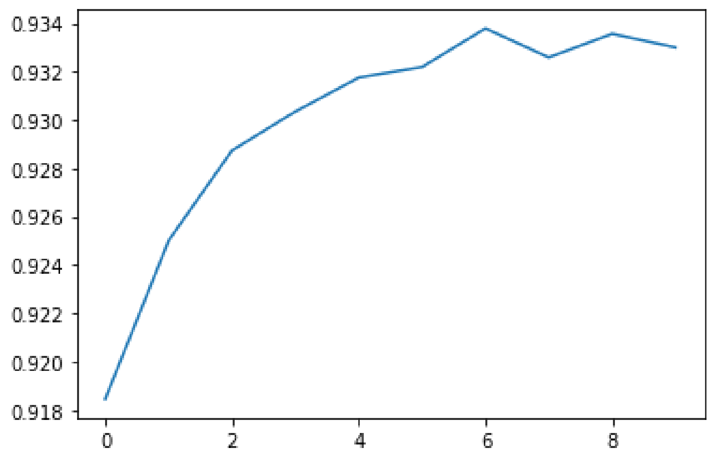


Figure 4-7 Training Accuracy for 15 frame ResNet-LSTM model

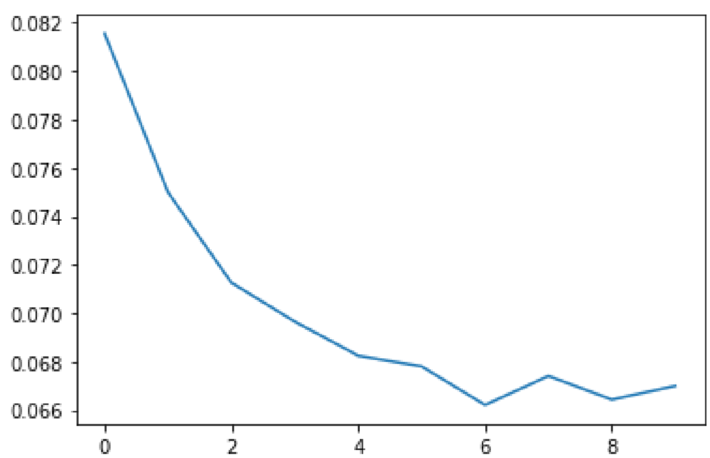


Figure 4-8 Training Loss for 15 frame ResNet-LSTM model

4.4.1 Result and Evaluation on visual-spatial-temporal features

The evaluation result for the 7 second and 15 second ResNet-LSTM model were 89.95% and 90.13%. This implied the inclusion of visual-spatial-temporal features can help to generalize the model well on smaller training set as well as larger dataset. With different amount of amount expose to the same model, the model able to generalize not while only with around 0.18% of difference in test performance. Besides, the different on training and validation accuracy also allowed to maintain on (2~3%) with both of the models.

4.5 Exploring visual-audio-spatial-temporal features with Audio-LSTM

In this experiment, the proposed Audio-LSTM model was built in this section. Audio-LSTM ensemble the trained LSTM-model and Audio MLP model. This experiment conducted with loading and freezing all the trained layers from the ResNet50-LSTM model with 15 frames extracted per video and trained only the Audio MLP layer and the ensemble module. Batch size (12), optimizer settings (learning rate=0.001, weight decay= 0.00001) and number of epochs applied.

4.6 Overall result and discussion

In this section, the result of apparent personality recognition model from different data with different modalities is shown. All the models were evaluated using the validation dataset. Results for fine-tuned ResNet, ResNet-LSTM and Audio-LSTM model show in Table 4-2.

Model	**Evaluation result(on validation set)**
Fine-tuned ResNet	87.38%
ResNet-LSTM	89.95%
Audio-LSTM	90.30%

Table 4-2 Overall validation result

The visual, visual-spatial-temporal, and audio-visual-spatial-temporal model achieved a result of 87.38%, 89.95%, and 90.21% respectively. There were few potential explanations for the change in model performance. First, only referring to information content from the visual modalities can achieve a significant amount of accuracy (87.38%). However, since the visual model only trained on a single image, this implies the information content that the visual model used to model the personality trait relies on the current exposure. However, the significant visual features might not exist at that moment lead to precluding accurate prediction. After adding the spatial-temporal feature(LSTM), the model can generalize better as well as the audio feature. The LSTM modalities allowed the model to obtain information from several frames which increase the exposure of information in a single video to the modal. By averaging the result from each of the frames, the model able to make more reliable predictions with a minimum amount of training set. From the experiment conducted, 15 frames per video extracted for the LSTM model able to achieve a significant test performance of 90.13% in the validation data. After that, the audio modalities also added to the LSTM model. Since the audio features were obtained by averaging the audio signals from the whole video, the model allowed to learned audio features and make better prediction. However, the audio modalities do not provide significant improvement to the performance might due to the video sample quality. Most of the video in the training sample contains unintelligible background sounds that becoming noise to the model but it does provide some reference for the model's overall.

4.7 Evaluation on Testing set

Other than validation data, another testing video data were used to evaluate the performance of the model. These testing data are independent from the validation and training data. Since the Audio-LSTM can achieve the best result out of three, the model was evaluated using the testing set. The proposed model achieved 0.9037 average mean in the test set. The Table 4-3 shows the comparison of result in each personality trait of Audio-LSTM model proposed in this research and the top ten entries in the challenge.

average	conscientiousness	extraversion	neuroticism	openness	agreeableness
0.9127	0.9166	0.9133	0.9100	0.9124	0.9126
0.9121	0.9119	0.9150	0.9100	0.9124	0.9119
0.9109	0.9138	0.9107	0.9089	0.9111	0.9102
0.9037	**0.9014**	**0.9102**	**0.9030**	**0.8997**	**0.9032**

Table 4-3 Comparison with top entries

The Audio-LSTM model achieved 90.3% while using lesser amount of data (~600 %) than the champion of the challenge while achieving only less than (~1%) of testing performance. By adding more modalities, the amount of information exposure of single video in personality trait recognition can be reduced while preserving the performance.

Figure below shows the amount of frame extracted during the training for the top entries.

Team	Average performance	Amount of frame extracted	Ratio of amount of frame extracted per video
NJU-LAMBDA	0.9126	100	100/15 =6.667
Evolgen	0.9119	200	200/15 = 13.33
Team DDC	0.9102	325	325/15 = 21.67
Audio-LSTM	**0.9037**	**15**	**1**

Table 4-4 Comparison with top entries in amount of frame

NJU-LAMBDA team as the champion in the ECCV ChaLearn LAP 2016 with the highest accuracy (0.9126%) but using a total of 100 frames extracted from each of the training videos as training data. On the other hand, the Audio-LSTM with only extracting 15 frames per video which is almost 600% smaller training set compared to NJU-LAMBDA proposed model, but only managed to maintain performance less than 1%. By extracting the salient information and ensemble of multiple modalities, the personality traits recognition model allowed to scale down the amount of information exposure to more than 700% while maintaining the performance with less than 1%.

Chapter 5 Conclusion

5.1 Overview

This research joins the visual, spatial-temporal, and audio modality to build a personality trait model that can quantify the observant personality traits based on the Big 5 personality model through video data. The First Impression 7-eleven rule is applied to improve the accuracy by sampling only the first 7 seven seconds of the video data for training. Other than implying First impression rule, the inclusion of various modalities allowed the proposed model to achieve significant performance while using a minimum amount of information. This research contribution comes from implementing a selective sampling technique that only includes a salient event from the video data while training that maintains the model's accuracy while only required a much smaller training set. This research also evaluates the relationship between the amount of information exposure for personality trait recognition and various modalities.

As a conclusion, by the inclusion of several vital features from the video with the application of selective sampling technique (First Impression rule), the model allowed to achieve optimum performance in quantifying human personality trait from the video data.

5.2 Future work

There are still many things could be done to improve the evaluation as well as the model performance. Due to the time limit, the setting of the model is limited. In the future, more images per video should be included for the training set that can boost performance further. Next, more audio features can be added to the Audio-LSTM model. Different parameter settings also can try out in the future by adjusting those parameters (learning rate) or increasing the depth for the neural network.

Bibliography

Abadi, M., Correa, J., Wache, J., Heng Yang, Patras, I. and Sebe, N. (2015). Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*.

Acuña, S., Gómez, M. & Juristo, N. (2009). How do personality, team processes and task characteristics relate to job satisfaction and software quality?. *Information and Software Technology*, 51(3), pp.627-639.

Ambady, N., Bernieri, F. and Richeson, J. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology Volume 32*, pp.201-271.

Azucar, D., Marengo, D. and Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, 124, pp.150-159.

Biel, J., Teijeiro-Mosquera, L. and Gatica-Perez, D. (2012). FaceTube. *Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12*.

Boudreau, J., Boswell, W., Judge, T. & Bretz, R. (2001) Personality and cognitive ability as predictors of Job Search Among Employed Mangers. *Personnel Psychology*, 54(1), p.25-50

Caldwell, D. & Burger, J. (1998). Personality Characteristics of Job Applicants abd success in screening interviews. *Personnel Psychology,* 51(1), pp. 119-136

Celli, F., Bruni, E. and Lepri, B. (2014). Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures. *Proceedings of the ACM International Conference on Multimedia - MM '14*.

Dorn, C., Skopik, F., Schall, D. & Dustdar, S. (2011). Interaction mining and skill-dependent recommendations for multi-objective team composition. *Data & Knowledge Engineering*, 70(10), pp.866-891.

Ferwerda, B., Schedl, M. and Tkalcic, M. (2016). Using Instagram Picture Features to Predict Users' Personality. *MultiMedia Modeling*, pp.850-861.

Jackson, D. (1985). Computer-based personality testing. *Computers in Human Behavior*, 1(3-4), pp.255-264.

Joo, J., Steen, F. and Zhu, S. (2015). Automated Facial Trait Judgment and Election Outcome Prediction: Social Dimensions of Face. *2015 IEEE International Conference on Computer Vision (ICCV)*.

Junior, Julio & Güçlütürk, Yağmur & Pérez, Marc & Güçlü, Umut & Andújar, Carlos & Baró, Xavier & Escalante, Hugo Jair & Guyon, Isabelle & Gerven, Marcel & Lier, Rob & Escalera, Sergio. (2018). First Impressions: A Survey on Vision-Based Apparent Personality Trait Analysis.

K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778.

Liu, L., Preotiuc-Pietro, D., Samani, Z.R., Moghaddam, M.E., & Ungar, L.H. (2016). Analyzing Personality through Social Media Profile Picture Choice. *ICWSM*.

Majumder, N., Poria, S., Gelbukh, A. and Cambria, E. (2017). Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE Intelligent Systems*, 32(2), pp.74-79. Minto, S. & Murphy, G. (2007). Recommending Emergent Teams. *Fourth International Workshop on Mining Software Repositories (MSR'07:ICSE Workshops 2007)*.

Pervin, L & John, O (1999). *Handbook of personality*. New York: Guilford Pass

Pianesi, F., Mana, N., Cappelletti, A., Lepri, B. and Zancanaro, M. (2008). Multimodal recognition of personality traits in social interactions. *Proceedings of the 10th international conference on Multimodal interfaces - IMCI '08*.

Pittenger, David. (1993). Measuring the MBTI … and coming up short. Journal of Career Planning and Employment. 54.

Silveira Jacques Junior, J., Gucluturk, Y., Perez, M., Guclu, U., Andujar, C., Baro, X., Escalante, H., Guyon, I., Van Gerven, M., Van Lier, R. and Escalera, S. (2019). First Impressions: A Survey on Vision-based Apparent Personality Trait Analysis. *IEEE Transactions on Affective Computing*, pp.1-1.

Wi, H., Oh, S., Mun, J. & Jung, M. (2009). A team formation model based on knowledge and collaboration. *Expert Systems with Applications*, 36(5), pp.9121-9134.

Yan, Y., Nie, J., Huang, L., Li, Y., Cao, Q., & Wei, Z. (2016). Exploring Relationship Between Face and Trustworthy Impression Using Mid-level Facial Features. *MMM*.

APPENDIX A FYP WEEKLY REPORT

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	**Study week no.: 1**
Student Name & ID: Eng Zi Jye 1605282	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE
2. WORK TO BE DONE **-** Research for dataloader
3. PROBLEMS ENCOUNTERED - Cannot fully load whole dataset
4. SELF EVALUATION OF THE PROGRESS **-** Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	**Study week no.: 2**
Student Name & ID: Eng Zi Jye 1605282	
Supervisor: Dr Aun Yichiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE
2. WORK TO BE DONE - Still research for dataloader - Need to change model from keras to pytorch
3. PROBLEMS ENCOUNTERED - Cannot fully load whole dataset - Need to learn pytorch structure
4. SELF EVALUATION OF THE PROGRESS -Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	**Study week no.: 3**
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE - Done developing dataloader
2. WORK TO BE DONE - Need to change model from keras to pytorch - Extract whole dataset
3. PROBLEMS ENCOUNTERED - Need to learn pytorch structure - Require computational power to extract image
4. SELF EVALUATION OF THE PROGRESS - Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	**Study week no.: 4**
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE - Done developing dataloader - Done extracting image
2. WORK TO BE DONE - Need to change model from keras to pytorch - Developing Visual model and fine tune
3. PROBLEMS ENCOUNTERED - Need to learn pytorch structure - Find skill to fine tune
4. SELF EVALUATION OF THE PROGRESS - Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 5
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE - Done Fine tuning model
2. WORK TO BE DONE - Evaluate model performance
3. PROBLEMS ENCOUNTERED - Need time to extract validation set
4. SELF EVALUATION OF THE PROGRESS **-** Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	**Study week no.: 6**
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE - Done evaluating the model
2. WORK TO BE DONE - Add spatial-temporal features - Create LSTM model
3. PROBLEMS ENCOUNTERED - No
4. SELF EVALUATION OF THE PROGRESS - Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 7
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE - Done creating LSTM model
2. WORK TO BE DONE - Evaluate LSTM model
3. PROBLEMS ENCOUNTERED - Accuracy not satisfies need fine tune
4. SELF EVALUATION OF THE PROGRESS - Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 8
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE - Fine tune new LSTM model with visual feature
2. WORK TO BE DONE - Evaluate LSTM model performance
3. PROBLEMS ENCOUNTERED - No
4. SELF EVALUATION OF THE PROGRESS - Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	**Study week no.: 9**
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE - Evaluate LSTM model
2. WORK TO BE DONE - Need fine tune more on LSTM
3. PROBLEMS ENCOUNTERED - Performance still not good enough
4. SELF EVALUATION OF THE PROGRESS - Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	**Study week no.: 10**
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE
- Fine tuned LSTM model
2. WORK TO BE DONE
- Evaluate performance on validation set
3. PROBLEMS ENCONTERED
- No
4. SELF EVALUATION OF THE PROGRESS
- Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	**Study week no.: 11**
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE
- Done evaluating all model
2. WORK TO BE DONE
- Model Audio LSTM model
3. PROBLEMS ENCOUNTERED
- No
4. SELF EVALUATION OF THE PROGRESS
- Still on track

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

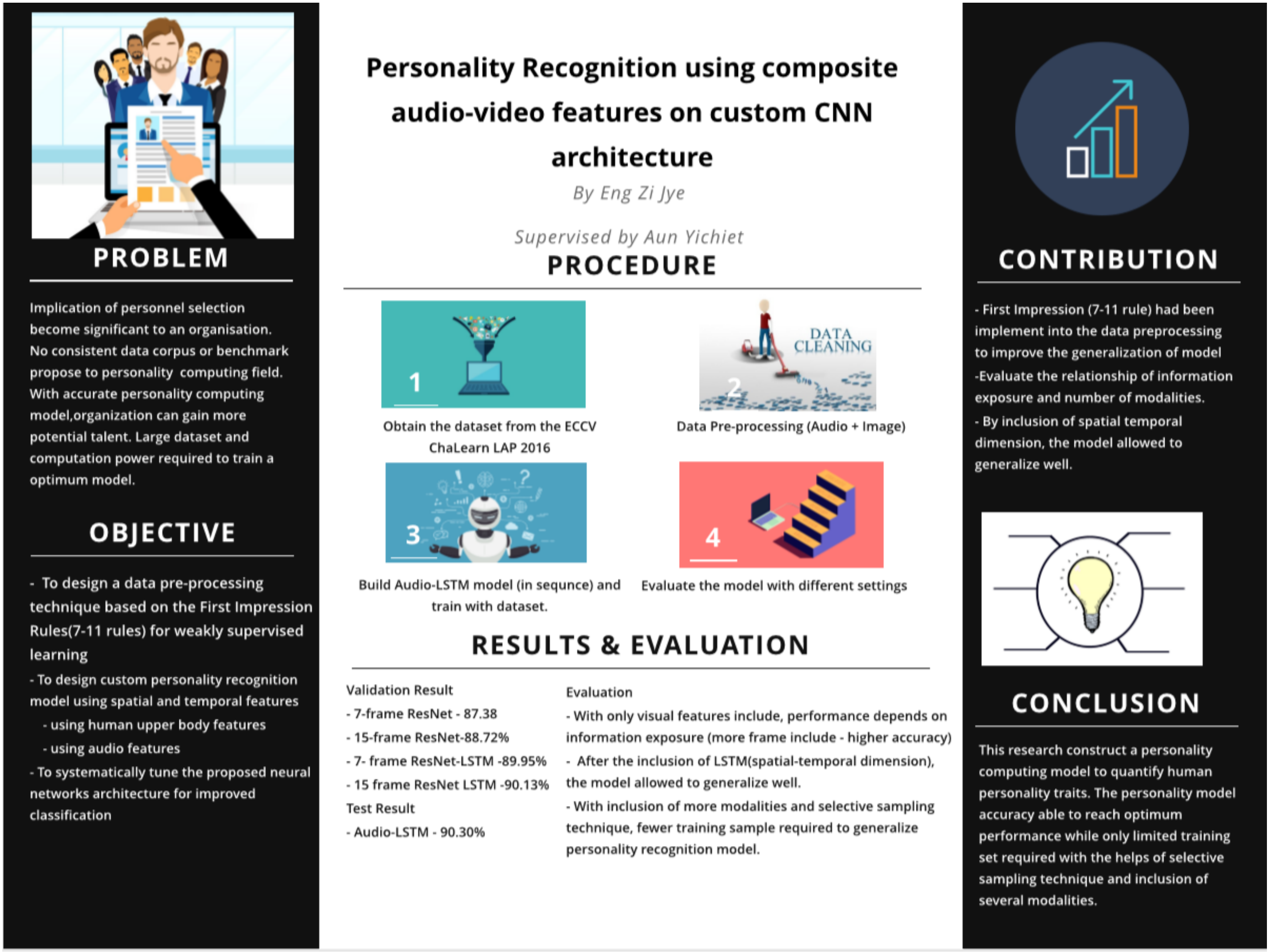
Trimester, Year: Y3S3	**Study week no.: 12**
Student Name & ID: Eng Zi Jye	
Supervisor: Dr Aun YiChiet	
Project Title: Personality Recognition using composite audio-video features on custom CNN architecture	

1. WORK DONE - Done building Audio Lstm model
2. WORK TO BE DONE - Evaluate performance
3. PROBLEMS ENCOUNTERED - No
4. SELF EVALUATION OF THE PROGRESS - Still on track

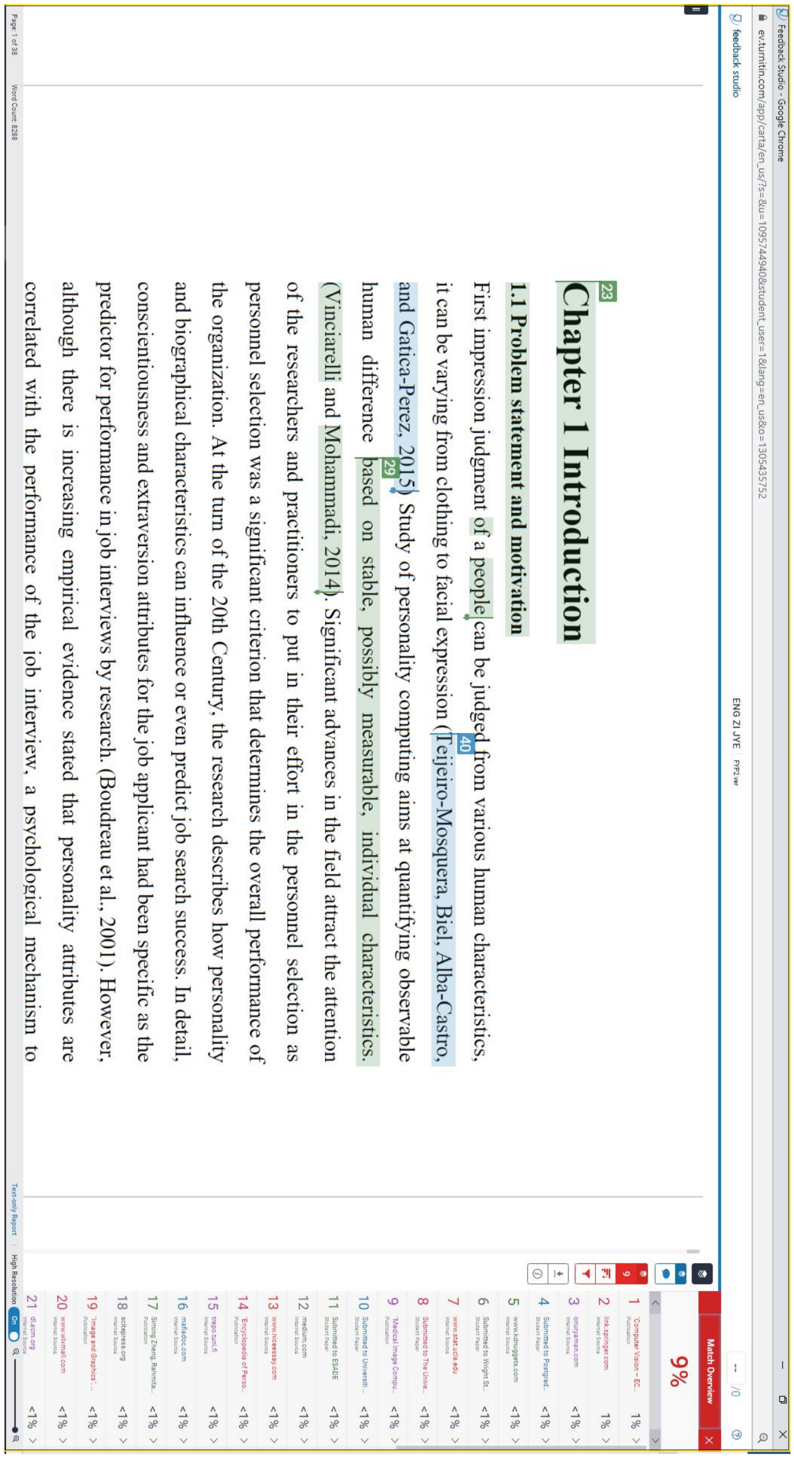
Supervisor's signature

Student's signature

POSTER


PROBLEM
Implication of personnel selection become significant to an organisation. No consistent data corpus or benchmark propose to personality computing field. With accurate personality computing model,organization can gain more potential talent. Large dataset and computation power required to train a optimum model.
OBJECTIVE
- To design a data pre-processing technique based on the First Impression Rules(7-11 rules) for weakly supervised learning
- To design custom personality recognition model using spatial and temporal features
- using human upper body features
- using audio features
- To systematically tune the proposed neural networks architecture for improved classification
Personality Recognition using composite audio-video features on custom CNN architecture
By Eng Zi Jye
Supervised by Aun Yichiet
PROCEDURE
1
Obtain the dataset from the ECCV ChaLearn LAP 2016
DATA CLEANING
2
Data Pre-processing (Audio + Image)
3
Build Audio-LSTM model (in sequnce) and train with dataset.
4
Evaluate the model with different settings
RESULTS & EVALUATION
Validation Result
- 7-frame ResNet - 87.38
- 15-frame ResNet-88.72%
- 7- frame ResNet-LSTM -89.95%
- 15 frame ResNet LSTM -90.13%
Test Result
- Audio-LSTM - 90.30%
Evaluation
- With only visual features include, performance depends on information exposure (more frame include - higher accuracy)
- After the inclusion of LSTM(spatial-temporal dimension), the model allowed to generalize well.
- With inclusion of more modalities and selective sampling technique, fewer training sample required to generalize personality recognition model.
CONTRIBUTION
- First Impression (7-11 rule) had been implement into the data preprocessing to improve the generalization of model
-Evaluate the relationship of information exposure and number of modalities.
- By inclusion of spatial temporal dimension, the model allowed to generalize well.
CONCLUSION
This research construct a personality computing model to quantify human personality traits. The personality model accuracy able to reach optimum performance while only limited training set required with the helps of selective sampling technique and inclusion of several modalities.

PLAGIARISM CHECK RESULT



Turnitin - Google Chrome

turnitin.com/newreport.asp?r=97.5649432784827&svr=58&lang=en_us&oid=1305435752&pbd=2&ft=1&ro=

preferences

turnitin Originality Report

Processed on: 23-Apr-2020 19:05 +08
ID: 1305435752
Word Count: 8288
Submitted: 1

FYP2 ver
By ENG ZI JYE

Similarity Index	Similarity by Source	
9%	Internet Sources:	4%
	Publications:	5%
	Student Papers:	5%

Document Viewer

include quoted include bibliography excluding matches < 8 words

mode: show highest matches together Change mode

Chapter 1 Introduction 1.1 Problem statement and motivation First impression judgment **of a people** 22

can be judged from various human characteristics, it can be varying from clothing to facial expression

(Teijeiro-Mosquera, Biel, Alba-Castro, and Gatica-Perez, 2015) 37

Study of personality computing aims at quantifying observable human difference

based on stable, possibly measurable, individual characteristics. (Vinciarelli and **Mohammadi, 2014).** 29

Significant advances in the field attract the attention of the researchers and practitioners to put in their effort in the personnel selection as personnel selection was a significant criterion that determines the overall performance of the organization. At the turn of the 20th Century, the research describes how personality and biographical characteristics can influence or even predict job search success. In detail, conscientiousness and extraversion attributes for the job applicant had been specific as the predictor for performance in job interviews by research. (Boudreau et al., 2001). However, although there is increasing empirical evidence stated that personality attributes are correlated with the performance of the job interview, a psychological mechanism to identify relationships is nearly unknown. In the field of Computer Science, in specific, deep learning with the characteristics that provide a broad spectrum of statistical methods that could automatically recognize human personality. Many methods had been proposed under personality computing for recognizing the personality trait from visual to audio.

Several characteristics of human **varying from clothing to** body gesture, all **contribute to the** 1

individual apparent personality's judgment. Nevertheless, until today, there is no persistent data corpus or benchmark models that had been introduced, which is one of the primary motivators of this research. A model that can recognize the human personality trait can bring much potential applications to society. As of today, there are several state-of-art models had been proposed in the

ImageNet Large Scale Visual Recognition Competition(ILSVRC), that can outperform **in** 48

several computer vision task, particularly in the image classification task. In 2015, a human labeller had reported having a 5.1% error rate while performing classification in the ImageNet while at the same time, the ResNet (152 layers) had achieved an error rate of 3.57%. Transfer learning had been proposed to put those state-of-art models into practical use. Transfer learning can improve the model learning through transferring the

knowledge from a related task that has already learned. By adopting the **transfer learning** 23

technique, the model can achieve better performance even without enough dataset. 1.2 Project Scope This research strives to model a profiling technique that can examine a human's personality by using a short video taken from the candidates. The Big 5 personality model will be applied as the guideline in which the candidate's

personality traits will **be** divided **into five dimensions** 14

include conscientiousness, extraversion,

1 1% match (publications)
"Computer Vision – ECCV 2016 Workshops", Springer Science and Business Media LLC, 2016

2 1% match (Internet from 02-Jul-2019)
https://link.springer.com/chapter/10.1007%2 3-319-49409-8_32

3 < 1% match (Internet from 24-Apr-2014)
http://onuryaman.com

4 < 1% match (student papers from 17-Jun-2019)
Submitted to Postgraduate Institute of Medicine

5 < 1% match (Internet from 31-Mar-2020)
https://www.kdnuggets.com/2020/02/audio-data-analysis-deep-learning-python-part-1.html

6 < 1% match (student papers from 30-Sep-2017)
Submitted to Wright State University

7 < 1% match (Internet from 08-Sep-2017)
http://www.stat.ucla.edu

8 < 1% match (student papers from 27-Apr-2018)
Submitted to The University of the South Pacific

9 < 1% match (publications)
"Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017", Springer Nature, 2017

10 < 1% match (student papers from 21-Apr-2020)
Submitted to ESADE

11 < 1% match (student papers from 07-Jan-2020)
Submitted to Universiti Teknologi MARA

FYP2 ver

ORIGINALITY REPORT

9%	4%	5%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Computer Vision – ECCV 2016 Workshops", Springer Science and Business Media LLC, 2016 Publication	1%
2	link.springer.com Internet Source	1%
3	onuryaman.com Internet Source	<1%
4	Submitted to Postgraduate Institute of Medicine Student Paper	<1%
5	www.kdnuggets.com Internet Source	<1%
6	Submitted to Wright State University Student Paper	<1%
7	www.stat.ucla.edu Internet Source	<1%
8	Submitted to The University of the South Pacific Student Paper	<1%

9	"Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017", Springer Nature, 2017 Publication	<1%
10	Submitted to Universiti Teknologi MARA Student Paper	<1%
11	Submitted to ESADE Student Paper	<1%
12	medium.com Internet Source	<1%
13	www.niceessay.com Internet Source	<1%
14	"Encyclopedia of Personality and Individual Differences", Springer Science and Business Media LLC, 2020 Publication	<1%
15	trepo.tuni.fi Internet Source	<1%
16	mafiadoc.com Internet Source	<1%
17	Siming Zheng, Rahmita Wirza OK Rahmat, Fatimah Khalid, Nurul Amelina Nasharuddin. "3D texture-based face recognition system using fine-tuned deep residual networks", PeerJ Computer Science, 2019	<1%

Publication

18	scitepress.org Internet Source	<1 %
19	"Image and Graphics", Springer Science and Business Media LLC, 2019 Publication	<1 %
20	www.wlxmall.com Internet Source	<1 %
21	dl.acm.org Internet Source	<1 %
22	Submitted to International University of Sarajevo Student Paper	<1 %
23	d-nb.info Internet Source	<1 %
24	Submitted to Cardiff University Student Paper	<1 %
25	Submitted to Harrisburg University of Science and Technology Student Paper	<1 %
26	nrl.northumbria.ac.uk Internet Source	<1 %
27	Submitted to Shri Guru Gobind Singhji Institute of Engineering and Technology Student Paper	<1 %

28	Xiu-Shen Wei, Chen-Lin Zhang, Hao Zhang, Jianxin Wu. "Deep Bimodal Regression of Apparent Personality Traits from Short Video Sequences", IEEE Transactions on Affective Computing, 2018 Publication	<1%
29	Submitted to National Defense University Student Paper	<1%
30	Submitted to South Bank University Student Paper	<1%
31	Submitted to International Hellenic University Student Paper	<1%
32	"Analysis of Images, Social Networks and Texts", Springer Science and Business Media LLC, 2018 Publication	<1%
33	Submitted to CSU, San Jose State University Student Paper	<1%
34	Submitted to American University in Cairo Student Paper	<1%
35	Submitted to Liverpool John Moores University Student Paper	<1%
36	Submitted to Universiti Tenaga Nasional Student Paper	<1%

37	Submitted to Kyungpook National University Student Paper	<1%
38	Hugo Jair Escalante, Meysam Madadi, Stephane Ayache, Evelyne Viegas et al. "Modeling, Recognizing, and Explaining Apparent Personality from Videos", IEEE Transactions on Affective Computing, 2020 Publication	<1%
39	Xiu-Shen Wei, Chen-Lin Zhang, Hao Zhang, Jianxin Wu. "Deep Bimodal Regression of Apparent Personality Traits from Short Video Sequences", IEEE Transactions on Affective Computing, 2017 Publication	<1%
40	Submitted to University of Kent at Canterbury Student Paper	<1%
41	"Advances in Multimedia Information Processing – PCM 2018", Springer Science and Business Media LLC, 2018 Publication	<1%
42	core.ac.uk Internet Source	<1%
43	Submitted to University of Malaya Student Paper	<1%
44	powerlab.fsb.hr Internet Source	<1%

45	Submitted to University of Hertfordshire Student Paper	<1%
46	www.motionmetrics.com Internet Source	<1%
47	Fabio Celli, Elia Bruni, Bruno Lepri. "Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures", Proceedings of the ACM International Conference on Multimedia - MM '14, 2014 Publication	<1%
48	Submitted to Universiti Teknologi Malaysia Student Paper	<1%
49	www.scribd.com Internet Source	<1%
50	Submitted to University of Southampton Student Paper	<1%

Exclude quotes	On	Exclude matches	< 8 words
Exclude bibliography	On		

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1of 1

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	Eng Zi Jye
ID Number(s)	16ACB05282
Programme / Course	Bachelor of Computer Science(BSC)
Title of Final Year Project	**PERSONALITY RECOGNITION USING COMPOSITE AUDIO-VIDEO FEATURES ON CUSTOM CNN ARCHITECURE**

Similarity	**Supervisor's Comments(Compulsory if parameters of originality exceeds the limits approved by**
Overall similarity index:___9_____ % **Similarity by source** Internet Sources: ___4____________% Publications: ____5________ % Student Papers: ___5__________ %	
Number of individual sources listed of more than 3% similarity: 0_________	

Parameters of originality required and limits approved by UTAR are as Follows:

(i) Overall similarity index is 20% and below, and

(ii) Matching of individual sources listed must be less than 3% each, and

(iii) Matching texts in continuous block must not exceed 8 words

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor	Signature of Co-Supervisor
Name: Aun Yichiet	Name:
Date: 24/4/2020	Date: 24/4/2020

UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	16ACB05282
Student Name	Eng Zi Jye
Supervisor Name	Dr Aun Yichiet

TICK (√)	**DOCUMENT ITEMS** Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
✓	Front Cover
✓	Signed Report Status Declaration Form
✓	Title Page
✓	Signed form of the Declaration of Originality
✓	Acknowledgement
✓	Abstract
✓	Table of Contents
✓	List of Figures (if applicable)
✓	List of Tables (if applicable)
	List of Symbols (if applicable)
✓	List of Abbreviations (if applicable)
✓	Chapters / Content
✓	Bibliography (or References)
✓	All references in bibliography are cited in the thesis, especially in the chapter of literature review
✓	Appendices (if applicable)
✓	Poster
✓	Signed Turnitin Report (Plagiarism Check Result – Form Number: FM-IAD-005)

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report. ________________________ (Signature of Student) Date: 24/4/2020	Supervisor verification. Report with incorrect format can get 5 mark (1 grade) reduction. ________________________ (Signature of Supervisor) Date:24/4/2020