Bachelor's Thesis
Academic year 2022

Textual summarization system of human personalities from video

Supervisor: Prof. Osamu Nakamura
co-advisors: Prof. Shigeya Suzuki, Prof. Achmad Husni Thamrin

Faculty of Environment and Information Studies
Keio University

Yubin Kwon

Abstract of Bachelor's Thesis

Academic year 2022

Textual summarization system of human personalities from video

Abstract

This project proposes a system that identifies and analyzes the human personalities from the long video with 2 persons having a conversation. This will be accompanied by a labeling system that correctly tracks and recognizes each person in the video, a recognition system that distinguishes human actions & facial expressions. The result will provide explanations of the reasons in a natural language for each person. In this field, there is an existing attempt to analyze personality from the interview video using speaker's intonation. However, there is a limitation that intonation would not be an enough indicator for personality. The main idea of this project is that it will be possible to derive a person's personality more accurately from the video that two people are having a conversation than the video that one person is speaking. 2 kinds of methods are selected to evaluate this project. From the first evaluation, it is assumed that the idea is reliable since more diverse labels have been detected from the videos of two people having a conversation, and the frequency and accuracy of detected labels are higher. Second evaluation was human evaluation to check the accuracy of inferred personality. However, due to the low number of test targets and mis-selection of the sample videos, the results are vague. There is still room to improve with this project. However, it has been successfully proved that the video of two people speaking is more suitable for inferring a person's characteristics than a single speech video throughout the project.

Keywords:

Video Summarization, Artificial Intelligence, Deep Learning, Human Activity Recognition, Facial Expression Recognition, Personality Test

Faculty of Environment and Information Studies, Keio University

Yubin Kwon

Contents

1 Introduction **6**
1.1 Background 6
1.2 Goal 7
1.3 Outline 8

2 Past Researches **9**
2.1 Video Summarization 9
2.2 Human Recognition 11
2.3 Personality Test 14

3 Problem **17**
3.1 Issues 17
3.2 Approach 18

4 Design **19**
4.1 Overview of the Structure 19
4.2 Video Splition 20
4.3 Recognition Model 20
4.4 Matching Personality Questionnaire 21
4.5 Integration 24

5 Implementation **25**
5.1 Environment Settings 25
5.2 Datasets 26
5.3 Implementation 27
5.3.1 Step 1 Data preprocessing 27
5.3.2 Step 2 Detection Model 27
5.3.3 Step 3 Human Action Recognition + Human Facial Expression Recognition Model 28
5.3.4 Step 4 Make a system to check a personality test questionnaire 31
5.3.5 Step 5 Integrate the each functions into one system 33

6 Evaluation **34**
6.1 Evaluation 1 34
6.1.1 Evaluation Design 34
6.1.2 Results 35
6.2 Evaluation 2 36
6.2.1 Evaluation Design 36
6.2.2 Results 37

7 Conclusion **39**
7.1 Limitations 39

7.2 Conclusion 39
7.3 Future Work 40

Acknowledgements **42**

References **43**

List of Figures

2.1.1 Video Summarizer Overview
2.1.2 Classification of Feature-based Approaches
2.2.2 Network structure of SlowFast
2.3.1 Questionnaire for the Big Five personality test

4.1.1 System Architecture
4.3.1 Flow of Recognition Model
4.4.1 Selected Questions from the Big Five test Questionnaire
4.4.2 Checklist

5.1.1 Tables for environment settings
5.2.1 Sample images for each labels in Kinetics Dataset
5.3.1 Spatio temporal action detection network
5.3.2 Sample result of spatio-temporal recognition
5.3.3 Sample result of spatio-temporal recognition with the system
5.3.4 Sample result of facial expressions recognition with the system
5.3.5 System flow example description
5.3.6 Score sample for agreeableness
5.3.7 Sample result

6.1.1 Comparison of the sample videos
6.1.2 Comparison for label frequency of the videos
6.2.2 Result of human evaluation

Chapter 1

1 Introduction

This chapter introduces the background and goal of this thesis and provides the outline of the following sections.

1.1 Background

Current society is an era of rich information. You can easily get the information you want through the Internet. Since the development of science and technology, especially in computers and IoT, massive amounts of data have accumulated regardless of the value of the data, and the existence of such a rich dataset has contributed greatly to the development of AI. However, there will also be a lot of unwanted information among them. In this information era, it is vital to think about how to easily obtain the desired information from the data and use it as valuable.

The format of data varies and one of them is a video. For example, YouTube is also serving as a search engine, in addition to its function as a social networking site. [1] Video is an electronic media of information combined with visual and audio information. However, it consumes time to obtain the desired information when watching videos. How can people effectively get the contents they need from the video? One representative method is video summarization.

There are several places where this automatic video summary technology is used. Among them, there is a tendency to analyze interviewees by combining AI technology during web interviews. This method has also been reported by several companies such as ‘Myinterview’ and ‘Curious Thing.’ The main theme is to analyze the expression and tone of the person in the single-person video. However, there are some limitations to this system. Will it be accurate to analyze a person's personality in a tone?

This paper proposes a way to evaluate a person's personality in a video with two or three people, based on the idea that a person's personality will be more visible when the one is interacting with others.

1.2 Goal

This project intends to develop a system that organizes the main contents of the video. The main content is inferred human personality of the people in the video. Furthermore, the provision of scenes on which the system determined the personality.

The goal of this research is to increase the time efficiency of information acquisition from the video, through automated video summarization, human action recognition, facial expression recognition, and voice analysis technologies. Also, another goal is to find a way to use the human-recorded video data meaningfully.

1.3 Outline

Overall outline of this thesis

This thesis is with 7 chapters.
Chapter 2 - Past research in this field will be introduced.
Chapter 3 - The problem will be defined.
Chapter 4 - The design of the project will be introduced.
Chapter 5 - The method of implementation will be introduced.
Chapter 6 - Evaluations of the system.
Chapter 7 - Conclusion with limitation and future work of the project.

Chapter 2

2 Past Researches

2.1 Video Summarization

This chapter will introduce Past Researches on 1. Video summarization, 2. Human activity & facial expression recognition, and 3. Criteria of personality test.

Summarizing video intents to pick out the most important scenes from the video. The main process is to create a set of video frames or fragments (video key-frames or key-fragments). Then, they are reconnected in sequential order, completing the summary as a short version of the video. There is also a method of summarizing videos using scenes from important parts in the form of storyboards, which is video skim. [2] The Figure 2.1.1 illustrates the overview of this procedure.

Getting the necessary information from a video takes a considerable amount of time. This is because users who want to get information need to watch the video until the essential part appears (also they will have to watch it until the end to make sure whether the target information appears again or not). This type of data is difficult to analyze and manage due to its temporal characteristics. Therefore, recognizing a specific feature from video data is a complicated technique. [3]

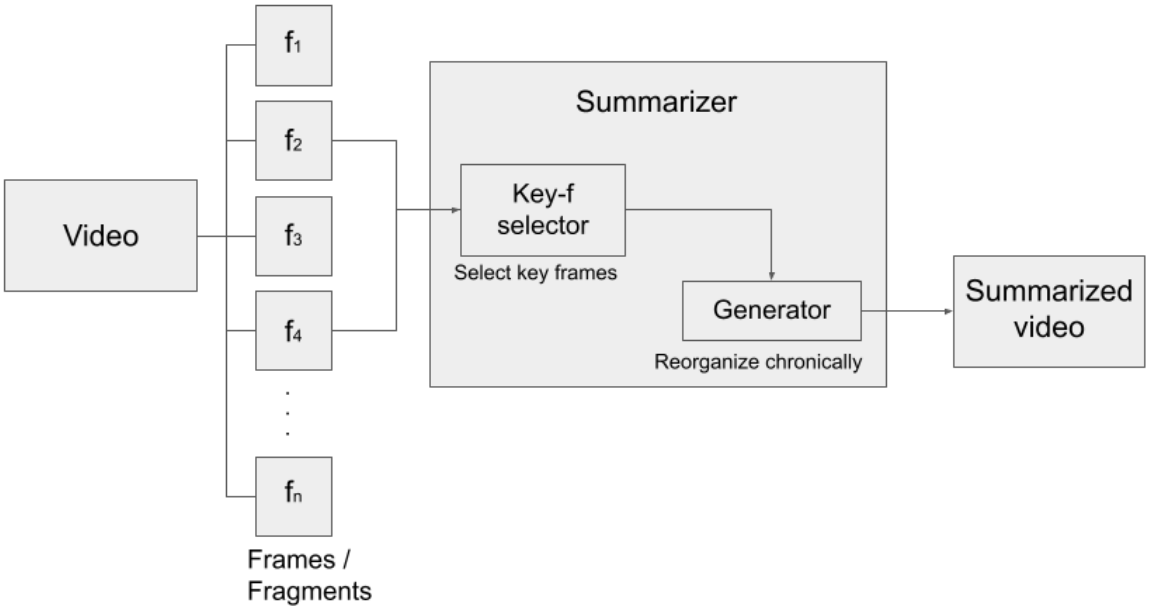

Figure 2.1.1. Video Summarizer Overview

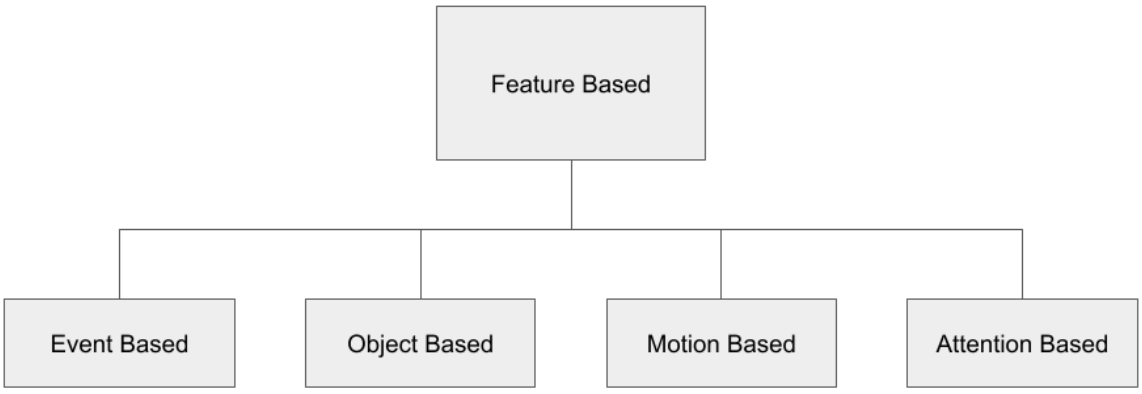

Figure 2.1.2 Classification of Feature-based Approaches

This project will focus on **feature-based** summarization, especially with the motion or event.

The video summarization technique is based on various elements. If the summarization procedure is done by focusing on the feature of the video, the features are as follows: *an object, attention, **event**, **motion**, etc* as Figure 2.1.2.

Event-based:

This approach is useful when the user wants to determine normal or abnormal events that exist in video content. Examples of the event-based approach are the detection of robbery, terrorism, etc. on CCTV. What these examples have in common is that the target scenes are sudden abnormal changes in the entire video. [3]

Motion-based:

Video summarization based on motion detection is a very important and representative task in many computer vision applications analysis. Its purpose is to extract an object moving at time t from a video sequence. The selection of motion detection as the main feature is also useful in summarizing sports videos, as in the paper in [4].

2.2 Human Recognition

Past researches on human activity recognition

- Human Action Recognition
- SlowFast Networks for video recognition
- Facial Expression Recognition

Human action recognition has been extensively developed in recent years. The use of convolutional networks (CNNs) was the cause of the great advancements. [5] CNNs are first introduced to this video classification field in [6]. Later, two-stream architectures [7] and 3D-CNN [8] are proposed, which could incorporate both appearance and motion features. Efforts have also been made to explore structures for a long-time video by the technologies such as temporal pooling or RNNs [9, 10, 11].

One of the most representative tools for recognizing human actions from the video is MMAction2. This is a part of the OpenMM Lab project, which is an open-source toolbox for the users to analyze video. The purpose of this project is to detect human actions in long, untrimmed videos. The system is based on the structured segment network (SSN). In [12], this project proposes a novel framework with a structured segment network (SSN), which models the temporal structure of each action instance through a structured temporal pyramid.

The MMAction2 tool supports the connection of various datasets and models. The contents are as [13]. Among them, PySlowFast [14] is one of the representative codebases. PySlowFast is an open-source video understanding codebase from FAIR that provides state-of-the-art video classification models with efficient training. The network structure of this codebase is as Figure 2.2.2.

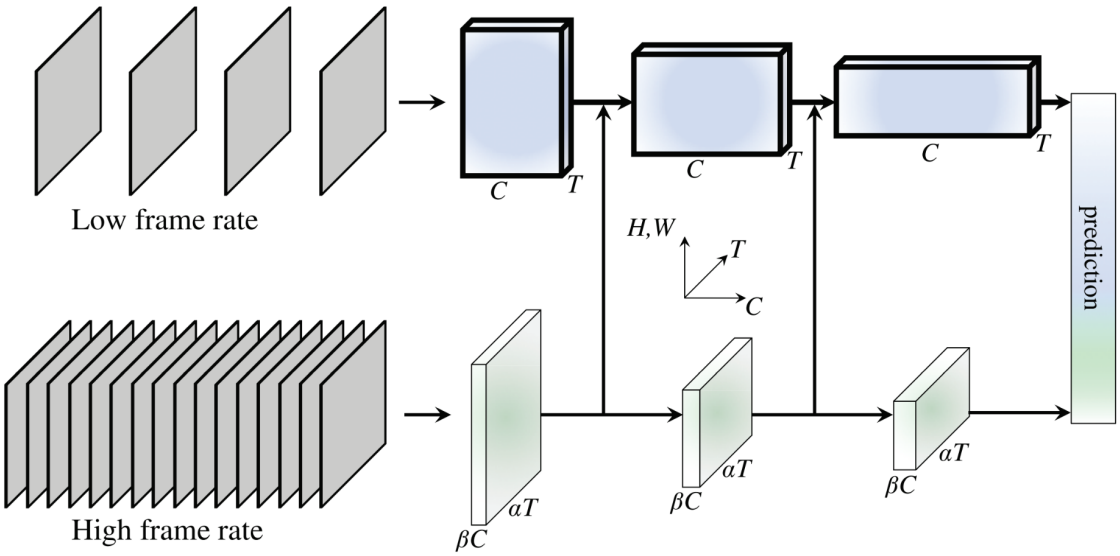

Figure 2.2.2 Network structure of SlowFast [14]

Facial expressions convey various meaningful messages in interaction with others. Therefore, several attempts to analyze people's facial expressions have been made for centuries. A human observer-based method that can be used to classify facial expressions was proposed at the early attempts. The Face Behavior Coding System (FACS) by Paul Eckman and Wallace V. Friesen is the most representative method for measuring facial expressions. However, the manual practice of this method is a labor-intensive task. The Automatic Facial Expression Recognition System (AFERS) was developed to automate the processes of FACS. This portable near-real-time system detects seven common emotional representations (Figure 2.2.3) and provides investigators with indicators during the interview process. [15]

2.3 Personality Test

In order to analyze a person's personality, the criteria for personality tests as an evaluation index are needed. For example, if there is a question ‘Are you good at making “small talk”?’, this is to figure out whether the person is sociable, and this question can be answered with the visual data. Let there is a video with 2 people, Alice and Bob, talking. If Alice started talking to Bob first more often in the video, we can assume that Alice is sociable. Like this, with personality indicators that can be evaluated from the data available from the video, one’s personality is able to be inferred without having to check it oneself.

The personality test to be introduced in this project is the Big Five personality test. The traits that can be assessed in this test are the best accepted and most commonly used model of personality in academic psychology. [16]

Those 5 personality traits are as follows; No.1 Agreeableness: The extent to which an individual values social harmony and getting along with others. No.2 Conscientiousness: The extent to which an individual is responsible, organized, dependable, and reliable. No.3 Extroversions: The extent to which an individual is gregarious, assertive, and comfortable around others. No.4 Openness: The extent to which an individual is imaginative and creative, as opposed to conventional. No.5 Stress Tolerance: The extent to which an individual will remain even-tempered and calm, as opposed to reacting emotionally to negative events.[16] The test consists of fifty items that the test subjects must rate on how true the questions are about them on a five-point scale where 1=Disagree, 3=Neutral and 5=Agree.

This test uses the Big-Five Factor Markers from the International Personality Item Pool, developed by Goldberg(1992)[17]. The Questionnaire list is shown in Figure 2.3.1.

	Disagree		Neutral		Agree
I am the life of the party.	○	○	○	○	○
I feel little concern for others.	○	○	○	○	○
I am always prepared.	○	○	○	○	○
I get stressed out easily.	○	○	○	○	○
I have a rich vocabulary.	○	○	○	○	○
I don't talk a lot.	○	○	○	○	○
I am interested in people.	○	○	○	○	○
I leave my belongings around.	○	○	○	○	○
I am relaxed most of the time.	○	○	○	○	○
I have difficulty understanding abstract ideas.	○	○	○	○	○
I feel comfortable around people.	○	○	○	○	○
I insult people.	○	○	○	○	○
I pay attention to details.	○	○	○	○	○
I worry about things.	○	○	○	○	○
I have a vivid imagination.	○	○	○	○	○
I keep in the background.	○	○	○	○	○
I sympathize with others' feelings.	○	○	○	○	○
I make a mess of things.	○	○	○	○	○
I seldom feel blue.	○	○	○	○	○
I am not interested in abstract ideas.	○	○	○	○	○
I start conversations.	○	○	○	○	○
I am not interested in other people's problems.	○	○	○	○	○
I get chores done right away.	○	○	○	○	○
I am easily disturbed.	○	○	○	○	○
I have excellent ideas.	○	○	○	○	○
I have little to say.	○	○	○	○	○
I have a soft heart.	○	○	○	○	○
I often forget to put things back in their proper place.	○	○	○	○	○
I get upset easily.	○	○	○	○	○
I do not have a good imagination.	○	○	○	○	○
I talk to a lot of different people at parties.	○	○	○	○	○
I am not really interested in others.	○	○	○	○	○
I like order.	○	○	○	○	○
I change my mood a lot.	○	○	○	○	○
I am quick to understand things.	○	○	○	○	○
I don't like to draw attention to myself.	○	○	○	○	○
I take time out for others.	○	○	○	○	○
I shirk my duties.	○	○	○	○	○
I have frequent mood swings.	○	○	○	○	○
I use difficult words.	○	○	○	○	○
I don't mind being the center of attention.	○	○	○	○	○
I feel others' emotions.	○	○	○	○	○
I follow a schedule.	○	○	○	○	○
I get irritated easily.	○	○	○	○	○
I spend time reflecting on things.	○	○	○	○	○
I am quiet around strangers.	○	○	○	○	○
I make people feel at ease.	○	○	○	○	○
I am exacting in my work.	○	○	○	○	○
I often feel blue.	○	○	○	○	○
I am full of ideas.	○	○	○	○	○

Figure 2.3.1 Questionnaire for the Big Five personality test

Chapter 3

3 Problem

This chapter introduces the issues of the problem and the approach to the solution.

3.1 Issues

Viewing (or checking) untrimmed long video data to get necessary information is time-consuming work. Accordingly, many video summarization technologies have been developed in recent years. The result of most video summarization technology is an objective list of events in the video. Looking at the example video summarizing AI introduced in the video [18], it accurately conveys how each figure-object is moving. For videos of non-human objects, this summary technique will be sufficient.

However, in the human-focused video, there is content that may be considered in more depth. With the summarized description of a person's facial expression or action behavior shown in the video, we can infer that person, their personality. What if AI gives the results of not only the objective facts that have happened in the video but also the inferred information?

This approach to the human-focused video analysis is also attracting attention in the AI-interview field. The company ‘Myinterview’ said they have the skill to judge the interviewee's personality through their intonation. They said the system is based on the Big Five Personality Test. However, there is a doubt in the system whether that intonation would be

enough of an indicator of human characteristics. [19] Therefore, there is room for improvement in this problem. What method can be used to increase objectivity when evaluating a person's personality? What data is required for an AI system to analyze a person's personality, and how should the data be pre-processed? And how should we use that data? This project would like to answer the above questions.

3.2 Approach

The Approach that I took to solve the suggested problem

For a specific and acceptable analysis of personality, this paper proposes that personality should be defined when people interact with others, so this paper suggests creating a system to analyze human personality from the video with two or more people appearing. The system proposed in this project will analyze the actions and facial expressions of one person when interacting with others. In addition, among the questions of the existing personality test, the list of items that can be answered with information that can be confirmed in the video - such as actions and facial expressions (visual data) - will be organized. The characteristics of people in the video are assessed with those data and questionnaires.

Chapter 4

4 Design

This Chapter will show the overview of the system structure.

4.1 Overview of the Structure

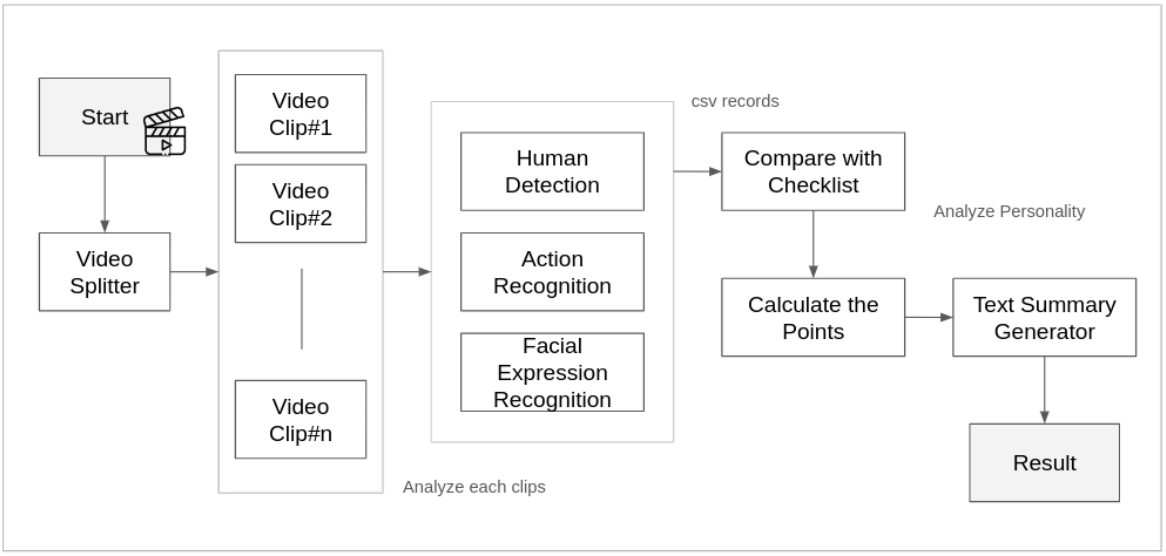

Figure 4.1.1 System Architecture

This system is divided into 4 main steps. First, it will split the video based on the target event or motion feature. The module for human distinction, action recognition, and facial expression recognition will run with the data. Then, the noted features will be compared to the personality questionnaire, eventually giving the inferred personality and explanation of the person.

4.2 Video Splition

When the user inputs the video, the system automatically separates the video and audio file by extracting the mp3 file from the original mp4 file. Then, it will divide the video into several clips, based on the frames. Later, it will merge the clips into one video with the same continuous event (ex. Person A speaking, Person B laughing). Therefore, there will be several short video clips from one video. Each short video contains event-feature information that can be summarized in one sentence.

4.3 Recognition Model

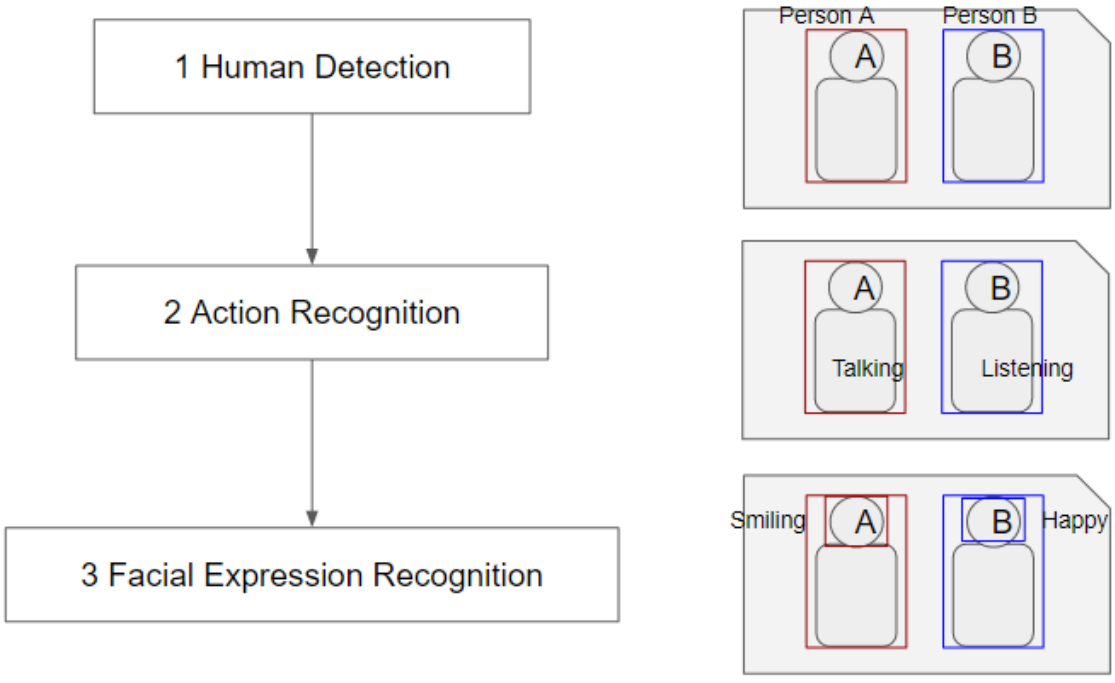

Figure 4.3.1 Flow of Recognition Model

The recognition model goes through in this flow. First detect humans so that it would not mistakenly give results that other objects are doing something. It will also prevent recognizing human-object interacting events. Then, the model will recognize human actions. For example, from clip1 to clip3, person A was talking and person B was listening. Lastly, the model will recognize the facial expressions per clip.

4.4 Matching Personality Questionnaire

4.4.1 List of the Questionnaire (from the Big Five Test)

https://ipip.ori.org/new_ipip-50-item-scale.htm

Among the total question contents, objective elements that can be automatically found in the video data were designated. The parts of the questions that require test subjects to make judgments based on their personality were excluded since they could not be identified through the video data.

Questionnaire: From 1~50 Questions, Select the questions which can be answered within visual/audio contents from video
(Light colored questions are presumed to be poorly analyzed, so they will be deleted later if shown unnecessary during the experiment.)

(1-)
6. Don’t talk a lot - action ‘talk to’
26. Have little to say - similar to no.6
46. Am quiet around strangers
(36. Don’t like to draw attention to myself)

(1+)
11. Feel comfortable around people - facial expressions
21. Start conversations - ‘talk to’, ‘asking questions’
31. Talk to a lot of different people at parties - ‘talk to’, ‘asking questions’
(41. Don’t mind being the center of attention)

(2-)
22. Am not interested in other peoples’ problems(2-) - eye contact ‘watch’
32. Am not really interested in others(2-) - ‘watch’, ‘listen to’

(2+)
7. Am interested in people - ‘asking questions’, ‘watch’, ‘listen to’
(17. Sympathize with others’ feelings - facial expressions)
(42. Feel others’ emotions - facial expressions)
47. Make people feel at ease

(3+)
3. Am always prepared - facial expressions
13. Pay attention to details - facial expressions

(4-)
4. Get stressed out easily - facial expressions
(14. Worry about things - facial : - possible error with 17.

(4+)
9. Am relaxed most of the time - facial expressions

Figure 4.4.1 Selected Questions from the Big Five test Questionnaire

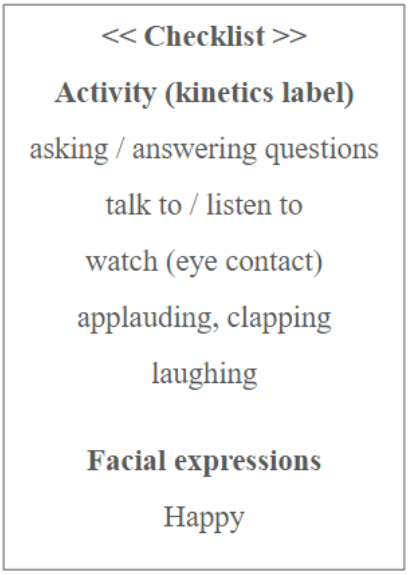

Figure 4.4.2 Checklist

The above list in Figure 4.4.2 was made by selecting questions as Figure 4.4.1 that could be judged by the data obtained from the video. The selection criteria are whether the question can be checked with yes (1) / no (0) as binary classification with visual/auditory evidence, among 50 questions. While preparing the list, No.5 Stress Tolerance was excluded because it was mainly about questions that had to be checked by itself (the evaluation target person), so it was judged that it was difficult to be visually evaluated. In addition, in the case of No. 3 Extroversion, it should evaluate the stress state shown in the facial expression and the tension felt in the tone, but there is only one corresponding item, so it is difficult to judge objectively. It will be recorded as an optional element.

Therefore, this project aims to analyze recognizable features such as human behaviors, facial expressions, and voices, to assess 1 Agreeableness, 2 Conscientiousness, and 3 Openness.

4.5 Integration

The last step of this system is to integrate the result in the previous procedure. There will be detected actions and facial expressions for each Person A and Person B. Also, those detected features will be matched with personality questionnaires. Those information and inferred personality based on the information will be given as a textual summary as a final result.

Chapter 5

5 Implementation

This chapter introduces the implementation procedure of the project.

5.1 Environment Settings

Environment Settings - Personal Computer	
GPU	GeForce GTX 1660 Ti
CUDA ver	10.1
Python	3.7
pytorch	1.9.0
torchvision	0.13.0
opencv	4.6.0

Environment Settings - Remote Server	
GPU	RTX A6000
CUDA ver	11.7
Python	3.7

Environment Settings - Google Colab	
GPU	NVIDIA Tesla K80
CUDA ver	11.1
Python	3.7

Figure 5.1.1 Tables for environment settings

5.2 Datasets

5.2.1 For human action recognition

Kinetics

This dataset is a large collection of high-quality URL datasets that can recognize 400/600/700 human actions depending on the version. The videos include human-object interactions such as shaking hands and hugging, as well as human-object interactions such as playing musical instruments and playing with sports equipment. It consists of up to 650,000 video clips, each of which lasts for about 10 seconds as a human annotated with a single action class. [20]

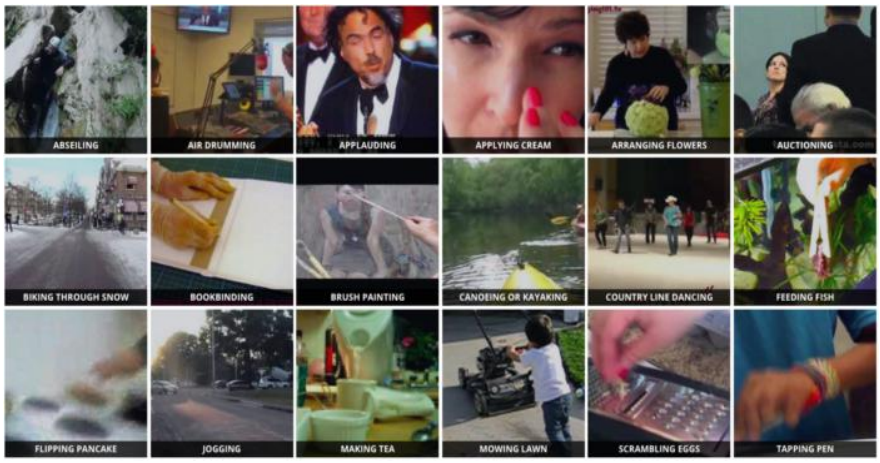

Figure 5.2.1 Sample images for each labels in Kinetics Dataset [20]

5.2.2 For facial expressions

Kaggle

This dataset is a processed version of the YouTube Faces dataset, which is basically with short videos of celebrities who are publicly available downloaded from YouTube. There are multiple videos (up to six) per each celebrity. Only the face perimeter from the original video is trimmed

and kept only up to 240 consecutive frames. This task was performed for disk space reasons and to make the dataset easier to use. [21]

5.3 Implementation

Steps following as:

5.3.1 Step 1 Data preprocessing

VIDEO PROCESSING: The video processing component of this application is responsible for sequencing the inputted video into 10 seconds long. The reason why it was divided into 10 seconds is that the basic type of kinetics data is clips of about 10 seconds length. Therefore, it was judged that the video of 10 seconds would be the most appropriate. In addition, too high resolution of the video results in too long running time. Therefore, it was adjusted to 320 pixels of low-quality videos, and width and height were also rearranged into 340 and 256 each.

5.3.2 Step 2 Detection Model

First of all, a person recognition model will go on as a base necessary procedure before the entire system is operational. Through this process, in videos featuring more than one person, it is clear who the model recognizes and records the data results. The person appearing in the video is labeled separately. [22]

5.3.3 Step 3 Human Action Recognition + Human Facial Expression Recognition Model

For Human Action Recognition, the model is referred from the spatio temporal action recognition model of mmaction2. In [23], they propose a simple but effective approach for spatio-temporal feature learning using deep 3D convolutional networks (3D ConvNets) trained on large supervised video datasets. The results of the study are as follows: 3 notes. 1) 3D ConvNets are better suited for spatio-temporal function learning than 2D ConvNets. 2) Homogeneous architectures with small 3x3x3 convolutional kernels in all layers are one of the best performance architectures for 3D ConvNets. 3) The features we have learned, namely Convolution 3D (C3D) cypher, with simple linear classes, outperforms state-of-the-art methods on four different benchmarks and is comparable to current best methods on two different benchmarks. Also, the function is compact. The paper suggests the achievement of 52.8% accuracy on UCF101 datasets in 10 dimensions alone and are also very efficient in computation due to the fast inference of ConvNets. Finally, they are conceptually very simple and easy to train and use. [23]

Using this network, the action of the person in the video is detected. According to the checklist, the contents of the action to be detected are as follows. The labels of actions as 'talk to', 'asking questions', 'listen to', 'answering questions', 'watch', and reactions as 'clapping'. Labels detected from each frame might differ. Therefore, in this process, all labels detected for more than 3 seconds in a clip with a length of 10 seconds are checked once per clip.

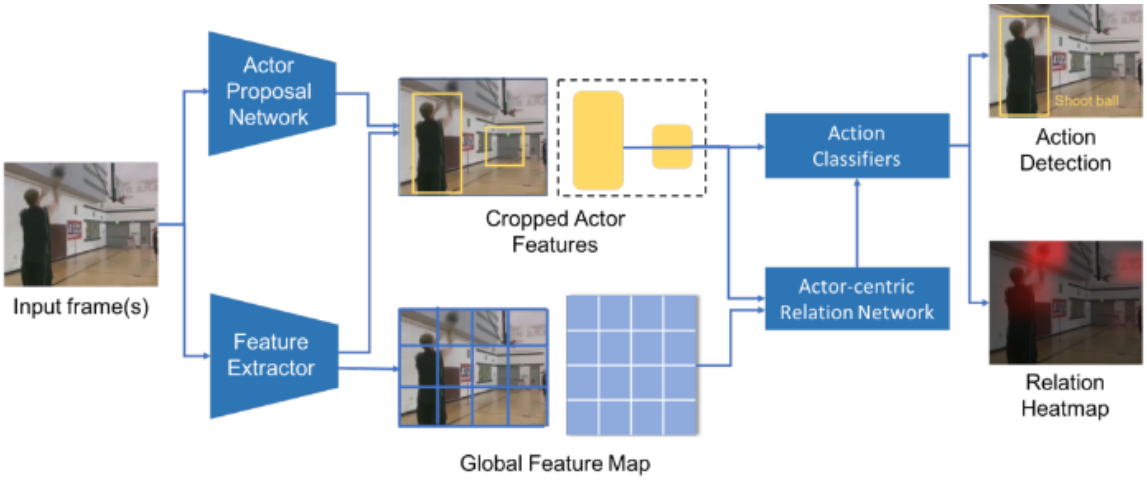

Figure 5.3.1 Spatio temporal action detection network [24]

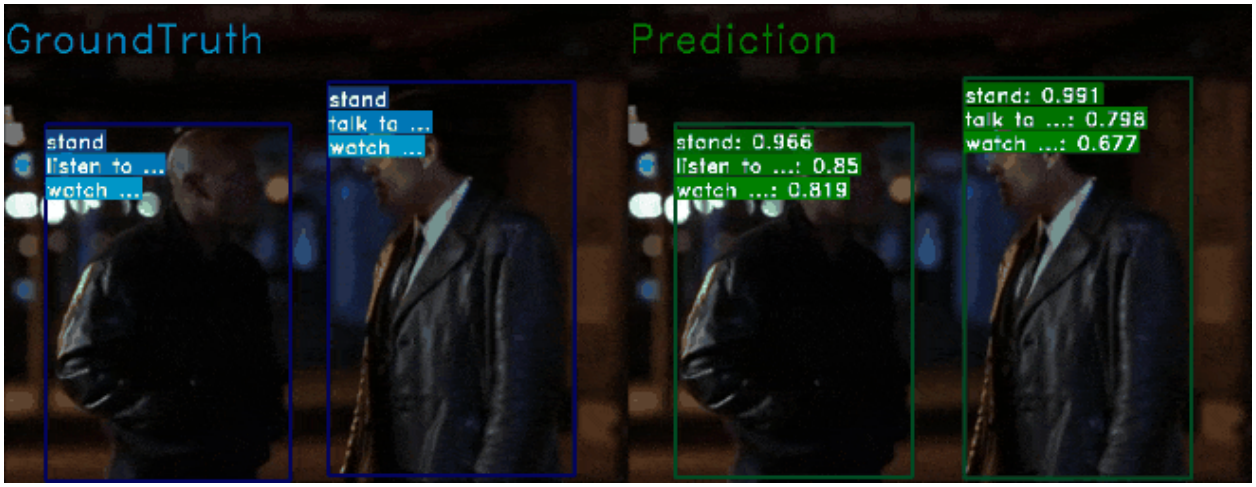

Figure 5.3.2 Sample result of spatio-temporal recognition [24]

Figure 5.3.3 Sample result of spatio-temporal recognition with the system

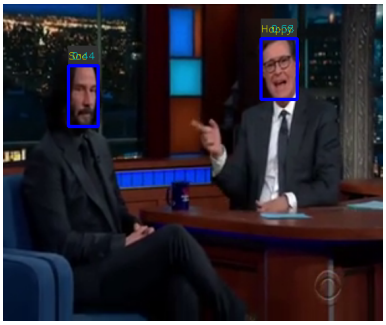

Figure 5.3.4 Sample result of facial expressions recognition with the system

In addition, after the procedure with the human recognition model ends, the facial expressions recognition model will go on. The model is referred from [15] [25]. Various facial expressions can be detected with the model, but only labels in the 'happy' state are used in this system. The purpose of using the 'happy' state is to provide an indicator for 'relaxed' in a questionnaire. As shown in Figure 5.3.4, facial expressions are recorded on a person's face. If a ‘happy’ state is recorded for more than 5 seconds in a clip video with a length of 10 seconds, it will be counted as showing a happy state.

5.3.4 Step 4 Make a system to check a personality test questionnaire

The video is truncated in 10 seconds. For example, if there is a video of 10 minutes (600 seconds), it will be divided into 60 clips. Among the contents, this model will check which of the characters A or B was more applicable to each of the questionnaire. All results are extracted as CSV files. Then, check who showed more indicators on the interaction questionnaire.

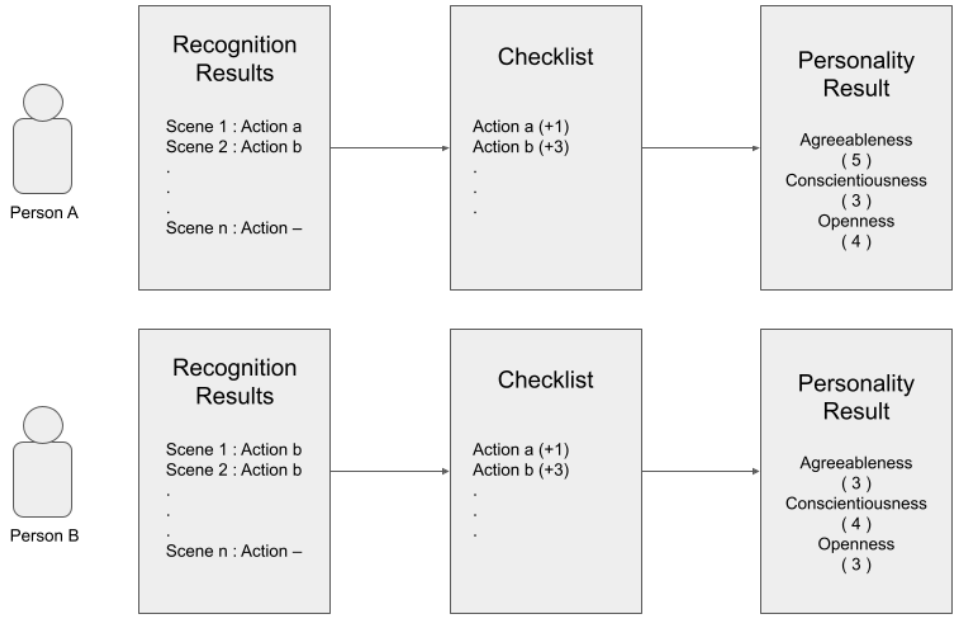

Figure 5.3.5 System flow example description

Calculation Method

For each agreeableness, constantness, and openness, the score calculation method for producing results is as follows. First of all, Agreeableness was measured through the number of times 'asking questions' and 'talk to' labels were recognized. And because of the nature of the

conversation, for example, in about 10 minutes of speaking, it would be common for each person to speak for 5 minutes. Therefore, the maximum number of times is assumed to be half of the total number of clips. As a result, if there are about 60 clips, 30 times is assumed to be the maximum number of times. And, the degree of agreeableness is divided into five stages. Divide each section by a maximum of 5 so that it is equally dispersed. The score classification of agreeableness when there are a total of 60 clips is shown in Figure 5.3.6.

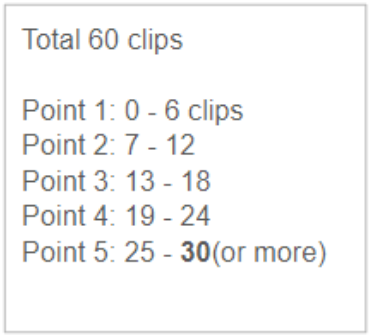

Figure 5.3.6 Score for agreeableness

Second, in the case of assessing Conscientiousness, the labels ‘listening to’, ‘answering to’ and ‘watch’ are used. As well as ‘asking questions’ or ‘talk to’, these indicators are normally recognized in half each of the whole video. Therefore, the above labels also make the maximum counts as a half of the entire clip numbers. In addition, there are labels that measure the reaction such as clapping or applauding. However, unexpectedly, these labels on the checklist did not appear frequently in the entire video. Therefore, if the behavior is checked several times during the entire video clip, the method of giving additional points was used.

Lastly, the measurement of Openness consists of how often facial expressions of the happy state are detected. Unlike speaking or listening actions, facial expressions can be found at any time. Therefore, the maximum is set equal to the total number of clips. It is to infer the degree of Openness by measuring how happy and relaxed the person was during the conversation.

5.3.5 Step 5 Integrate the each functions into one system

Integrate the functions of each model. The actions recorded in each scene are presented as results in sentences. It serves as a convincing basis for revealing who was more active in interaction in the video. The results will be presented as shown in Figure 5.3.7. The system will show the user a summary of what ratio of indicators (talking, reacting, and facial expressions) each person performed in the video and how many points were recorded for each personality section;
1) agreeableness, 2) conscientiousness, and 3) openness.

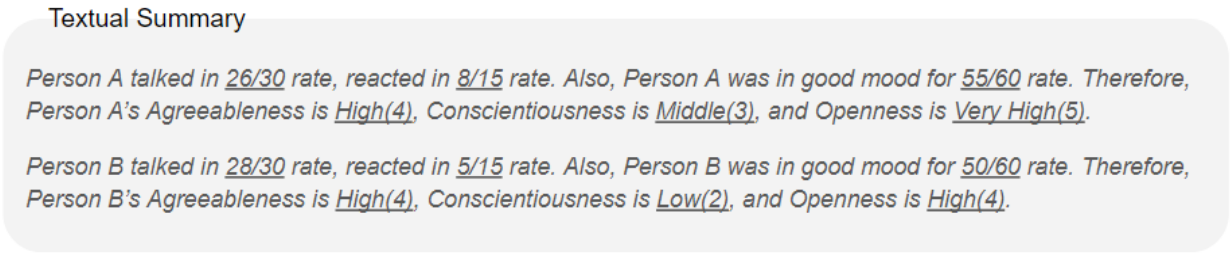

Figure 5.3.7 Sample result

Chapter 6

6 Evaluation

This Chapter will show how I evaluated the project. I evaluated this project in two methods.

6.1 Evaluation 1

6.1.1 Evaluation Design

The main content of this research project is that compared to the video in which one person appears, the video in which two people are talking will allow them to infer a person's personality more accurately. To prove the content, the following evaluation method was implemented. First, about the same celebrity, videos of speaking alone and talking with interviewers were collected. And through the action recognition model & facial expressions recognition model of the system, we compared which actions are measured.

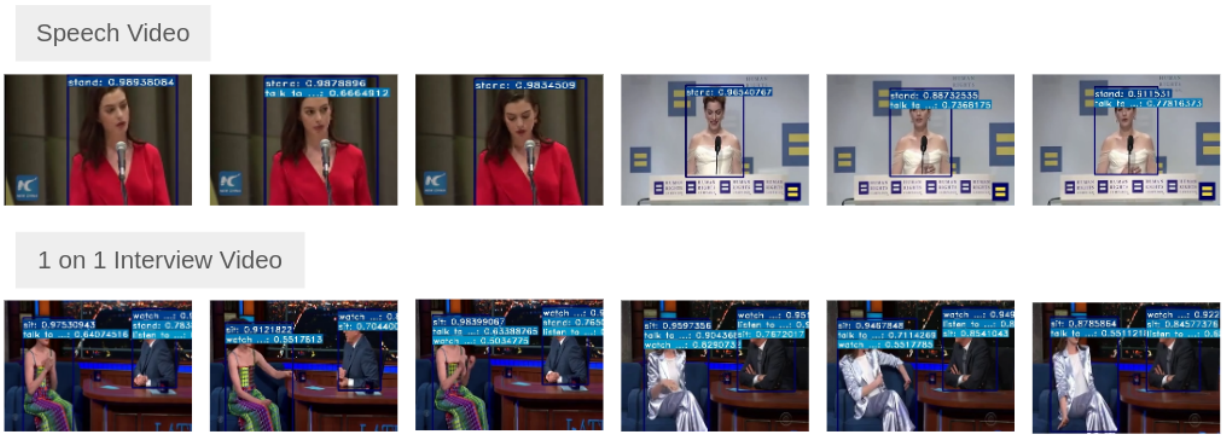

Figure 6.1.1 Comparison of the sample videos

6.1.2 Results

Through the comparison of the videos in Figure 6.1.1, it was possible to prove that this hypothesis was reliable. The comparison of labels is shown in Figure 6.1.2. In the case of the label 'talk to', it was detected in both videos. However, due to the nature of the one-person speech video, even 'talk to' label was not properly detected, regarding that the person must have kept talking. In the degree of half of the whole video, the person's action was not properly recognized and missed.

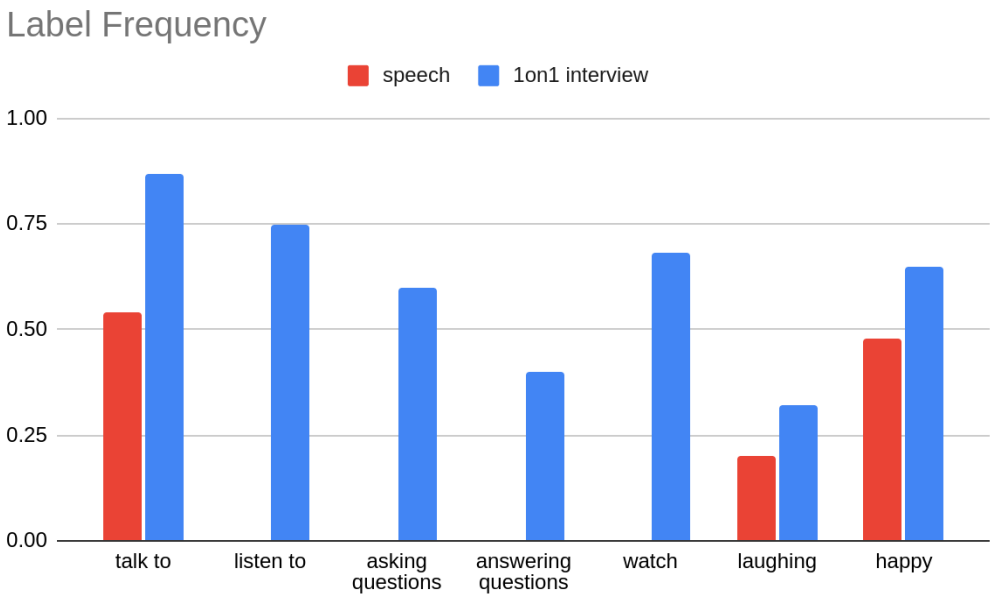

Figure 6.1.2 Comparison for label frequency of the videos

In addition, the system recognizes more diverse labels in the video of the two people talking. Labels such as 'listen to', 'asking questions', 'answering questions', and 'watch' do not

seem to be detected if there is no person present as an opponent. Looking at Figure 6.1.2, the following label was not recognized at all in the single-person speech video. However, in the video where the two are talking, these labels appear at a fairly high frequency. Therefore, through these results, it will be proved that various indicators for personality diagnosis can be extracted from the video of two people talking rather than the video of talking alone.

6.2 Evaluation 2

6.2.1 Evaluation Design

This evaluation is human evaluation. Four 10-minute videos were designated as samples. The test subjects were four people between the ages of 20 and 25. These 4 videos are first run through the system, and the system infers the personalities of each of the two persons in the video. In addition, test subjects also watch the video and check the assumed answer to diagnose the person's personality through the questionnaire. The questionnaire brought questions from the Big Five Test, which could diagnose agreeableness, consensus, and openness, respectively, to evaluate the person in the video.

The answers of the questionnaire are in 5 steps, from very low(1), low(2), neutral(3), high(4), very high(5).

6.2.2 Results

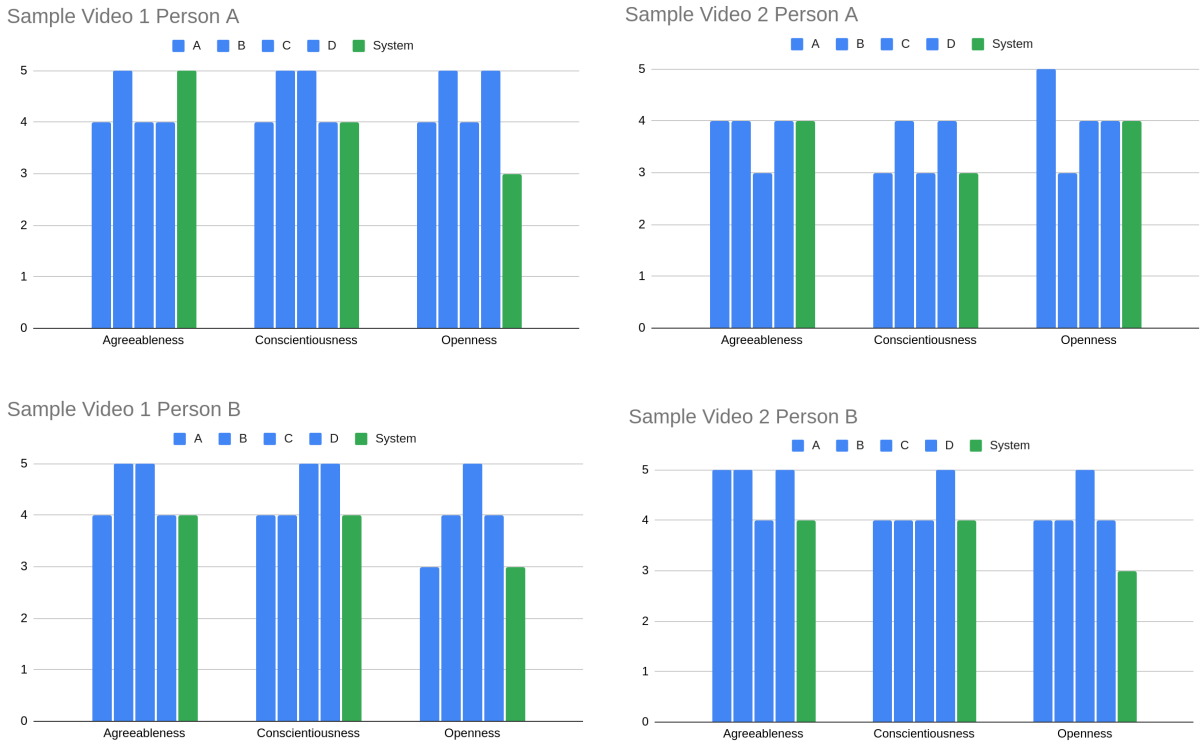

Figure 6.2.2 Result of human evaluation

Figure 6.2.2 is a graph showing the comparison between the personality evaluated by people and the personality evaluated by the system. However, there are three problems with this assessment. First, since a single video is 10 minutes long, it takes a lot of time for the test takers, and thus does not have many test target personnel. Second, there are cases where each person evaluates the same person differently. This is evident in Sample Video 1 Person B's evaluation of Openness personality. For one person, a very different result came out. Considering these factors, the question arises as to whether human evaluation is appropriate to evaluate this system. If more test targets have been secured, there is a possibility that these errors have been supplemented.

The third is that every video is a TV show video of celebrities. In general, their actions were very interactive, so they mostly got high scores. So it resulted in the difficulty in evaluating the accuracy of the system.

Chapter 7

7 Conclusion

This Chapter will show the conclusion, limitations and future work of this project

7.1 Limitations

What we can use effectively in video data is also sound information, not only visual information. In this project, the audio recognition model was not added to the system due to time problems.

In addition, there is a problem with the sample of videos used in evaluation. As for the sample video for evaluation, interview videos of celebrities released were selected and used due to copyright issues and good accessibility. However, problems arise in this selection method. The behavior of celebrities in the videos is quite contrived. Whatever the actual characteristics of celebrities, their actions and facial expressions on TV shows are generally similar. They will laugh more and actively engage in conversation. Therefore, it was difficult to confirm whether the personalities of various people were properly distinguished.

7.2 Conclusion

The meaning of this project is to create a system to infer human personality with textual explanation from a long video. Through Evaluation 1, it is assumed that we can get better quality

and quantity of indicators (labels) for assessing personality when the person is having a conversation. Accuracy assessed by Evaluation 2 is hard to trust due to low number of the test targets and not appropriate sample data. Although, it was meaningful in a point that the system could derive its own results and compared them with human results.

7.3 Future Work

Sound is not yet used in this system. In the future, it will be a more accurate system if it is upgraded to a system that uses all the various data available from videos, including sound information. Also, if the system can distinguish more diverse indicators, this will also increase accuracy. Currently, the system is made up of labels that already exist. However, the reliability in the personality infer system will increase if the recognition model is upgraded to configure other action labels through training with the new dataset.

Considerable technical development factors include the development of applications. If an application with recording real-time video with the camera becomes possible, it will be possible to shoot a video of the people having a conversation and get the result of inferred personalities after a few minutes. In addition, implementing a new evaluation method that would compensate for the shortcomings of the existing human evaluation. First, a video in which test targets communicate with each other is taken. Next, let them infer their personalities through the Big Five Test on their own. Then the video will be also analyzed through the system. The purpose of this evaluation is to study how similar the test targets' personality assessed by themselves is to the personality inferred through the system. This procedure is expected to

somewhat compensate for the problems that appeared when using celebrity videos. Unlike TV show videos of celebrities, test targets' videos will be able to show a variety of personalities.

Lastly, there is a possibility of enhancement with the number of people having a conversation in the video. The system is for analyzing conversation videos of only two people. It would not work properly if there are more than two people. However, it is assumed that the analysis will be appropriate for the purpose of the system only when there are more than two people. The reason lies in the nature of the conversation. When two people have a conversation, either side must speak. Therefore, it is difficult to find out who is active in interacting with others. However, if the number of people who participate in the conversation increases, there will be a clear distinction between those who lead the conversation, those who answer more, or those who cannot ask questions and do not participate in the conversation. Therefore, the system needs to be upgraded to analyze conversations of three or more people later.

Acknowledgements

First of all, I would like to thank Professor Osamu Nakamura, Professor Shigeya Suzuki and Professor Achmad Husni Thamrin for their great support and guidance throughout this project. I would also like to thank all the people who participated in my experiment.

I want to thank all members of the SFC RG Joint Research Group and KUMO for their support in the past years I have spent in Keio University. The experiences that I have had through the research group have become a huge asset in my life. Which made me grow further, and gave me the foundation for my next journey. I hope to get the opportunity to meet you again someday.

Thank you.

References

[1] Adnan Veysel Ertemel, Ahmed Ammoura, “Is YouTube a Search Engine or a Social Network? Analyzing Evaluative Inconsistencies,” in Business and Economics Research Journal, 2021 https://www.ceeol.com/search/article-detail?id=994260

[2] Apostolidis, Evlampios & Adamantidou, Eleni & Metsai, Alexandros & Mezaris, Vasileios & Patras, Ioannis. (2021). Video Summarization Using Deep Neural Networks: A Survey.

[3] Haq, Hafiz Burhan & Asif, M & Bin, Maaz. (2021). Video Summarization Techniques: A Review. International Journal of Scientific & Technology Research. 9. 146-153.

[4] Engin Mendi, Hélio B. Clemente, Coskun Bayrak, Sports video summarization based on motion analysis, Computers & Electrical Engineering, Volume 39, Issue 3, 2013, Pages 790-796, ISSN 0045-7906

[5] Z. Yue et al. Temporal Action Detection with Structured Segment Networks, arXiv:1704.06228

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, pages 1725–1732, 2014.

[7] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, pages 568–576, 2014.

[8] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In ICCV, pages 4489–4497, 2015.

[9] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In CVPR, pages 4305–4314, 2015

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In CVPR, pages 2625–2634, 2015.

[11] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In CVPR, pages 4694–4702, 2015.

[12] MMAction2 Contributors, (2020) OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark https://github.com/open-mmlab/mmaction2

[13] mmaction2
Benchmarks https://mmaction2.readthedocs.io/en/latest/benchmark.html
Model Zoo https://mmaction2.readthedocs.io/en/latest/modelzoo.html

[14] Christoph Feichtenhofer et al. SlowFast Networks for Video Recognition arXiv:1812.03982 https://github.com/facebookresearch/SlowFast

[15] A. Ryan et al., "Automated Facial Expression Recognition System," 43rd Annual 2009 International Carnahan Conference on Security Technology, 2009, pp. 172-177, doi: 10.1109/CCST.2009.5335546.

[16] The Big Five Personality Test site https://openpsychometrics.org/tests/IPIP-BFFM/

[17] Goldberg, Lewis R. "The development of markers for the Big-Five factor structure." Psychological assessment 4.1 (1992): 26.

[18] https://www.youtube.com/watch?v=bVXPnP8k6yo

[19] https://www.technologyreview.com/2021/07/07/1027916/we-tested-ai-interview-tools/

[20] The Kinetics Human Action Video Dataset arXiv:1705.06950

[21] Kaggle dataset
https://www.kaggle.com/datasets/selfishgene/youtube-faces-with-facial-keypoints

[22] Sharif et al. EURASIP Journal on Image and Video Processing (2017) 2017:89 DOI 10.1186/s13640-017-0236-8 “A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection“

[23] Du Tran et al. “Learning Spatiotemporal Features with 3D Convolutional Networks” arXiv:1412.0767

[24] SPATIO TEMPORAL ACTION DETECTION MODELS - mmaction2 documents https://mmaction2.readthedocs.io/en/latest/detection_models.html

[25] U. Gogate, A. Parate, S. Sah and S. Narayanan, "Real Time Emotion Recognition and Gender Classification," 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), 2020, pp. 138-143, doi: 10.1109/ICSIDEMPC49020.2020.9299633.

[26] Video Samples from: https://www.youtube.com/c/ColbertLateShow, https://www.youtube.com/watch?v=rOqUiXhECos, https://www.youtube.com/watch?v=cXZ7GBZrG8k, https://www.youtube.com/watch?v=PeLNRMrAEUA