Textual summarization system of human personalities from long video

Kumo B4 Yubin

Supervisor: Osamu Nakamura, Shigeya Suzuki, and Ahmad Thamrin Husni

Contents

Background

These days, a lot of information can be obtained through video. When reading textual data, we can quickly find a page with the information we need as fast as we read it. However, we must watch the video to obtain the desired information when watching videos. In order to increase time efficiency, this project intends to develop a system that organizes the main contents of the video into the textual form so that one can get information easily.

Idea

➢ Interested in studying human focused video…

- What can be analyzed from human-content video?
- What method can be used to collect the data?
- What method is needed to get meaningful results from the data?

This project starts with the above questions.

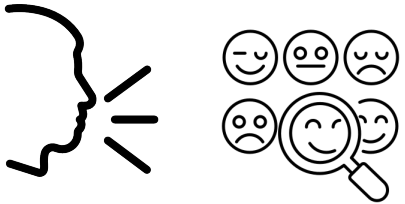
Specific Goal

- **How** to summarize the video?

  > Textual Summary of Long Video

- **Which Features** are analyzed in the video?

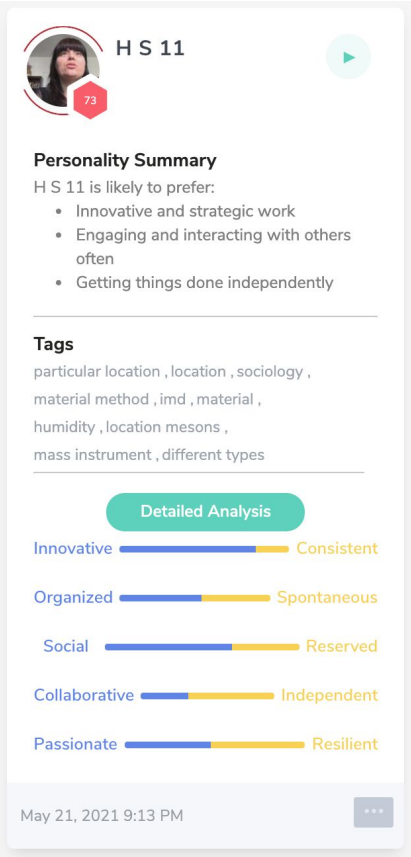
  > Human Action, Facial Expression



- **What** will be the results?

  > Reasoning a person's personality

Current Technology and Limitations



The company ‘Myinterview’

>> judge the interviewee's personality through their **intonation**.

- **Whether intonation would be enough as an indicator of human characteristics**

But intonation isn’t a reliable indicator of personality traits, says Fred Oswald, a professor of industrial organizational psychology at Rice University. “We really can’t use intonation as data for hiring,” he says. “That just doesn’t seem fair or reliable or valid.”

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

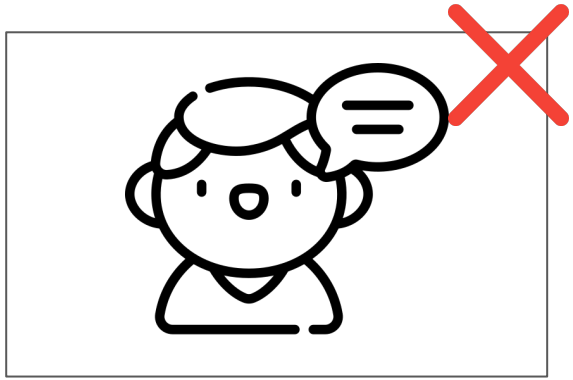
Problem Definition

In a video of a person talking **alone**, it is difficult to analyze a person's personality. The questions of the Big Five personality test are also based on **interactions with others**.

- ➢ Then, 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?

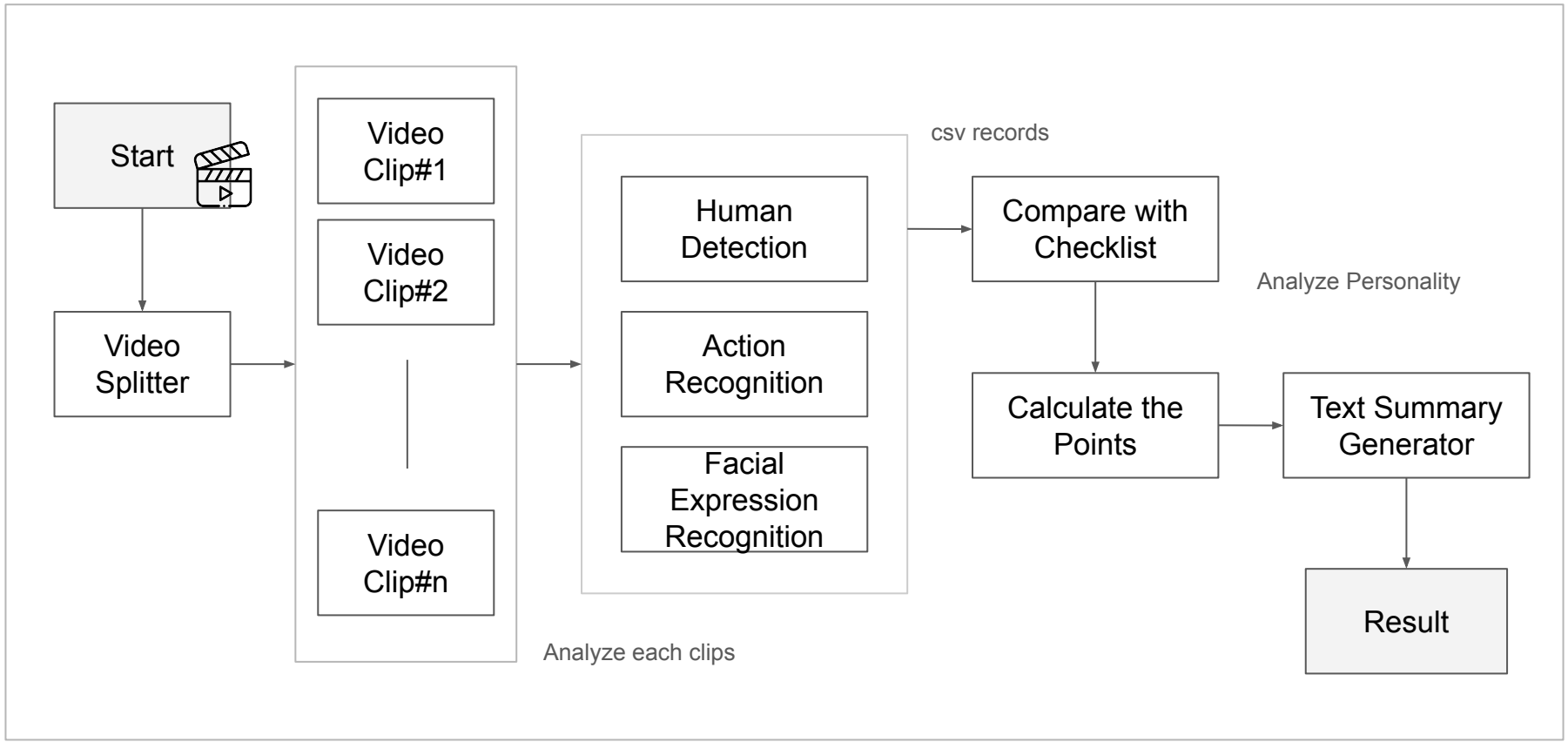
Proposed Methodology

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 will be organized. The personality will be assessed with using the collected data and questionnaires.

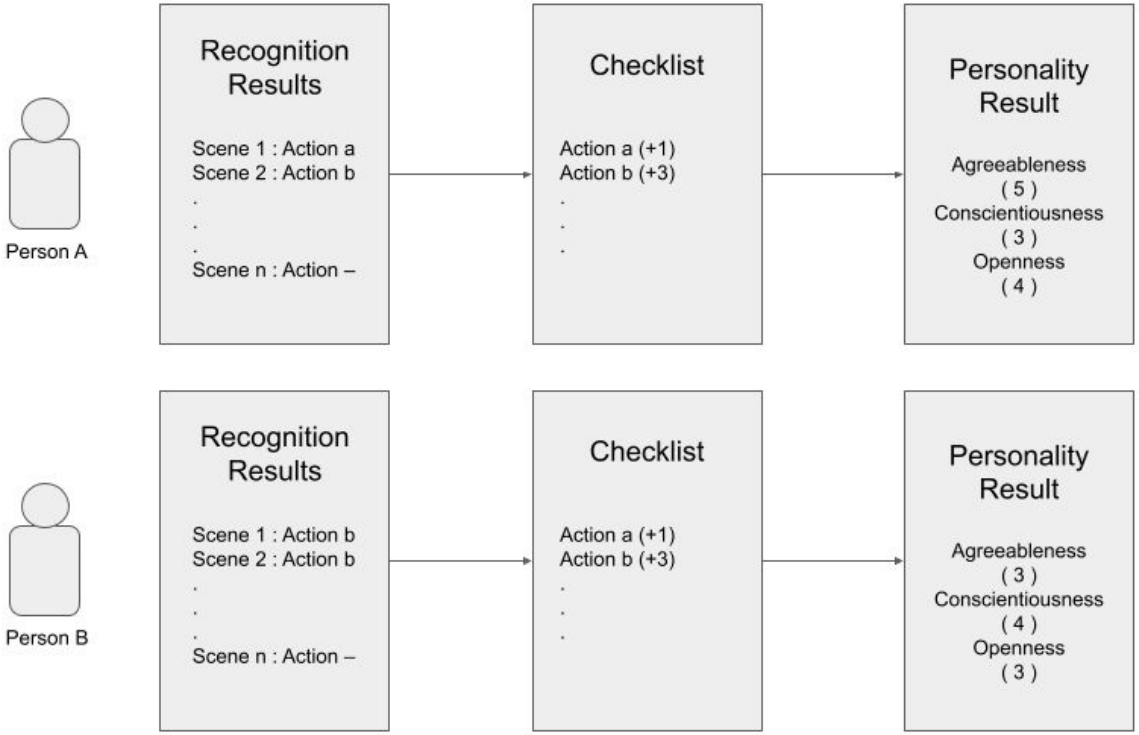


System Design

System Architecture



Expected Flow



Environments

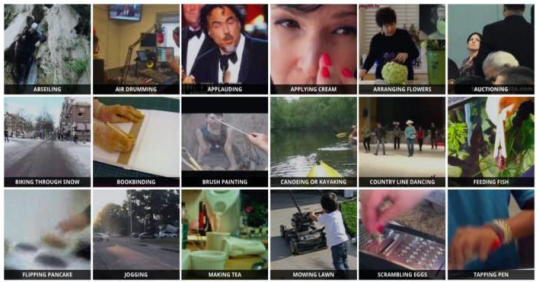
1) my laptop
Ubuntu 20.04 / GPU GeForce GTX 1660 Ti / CUDA 10.1

2) Training : on remote server - anaconda jupyter notebook
Cuda compilation tools, release 11.5, V11.5.119

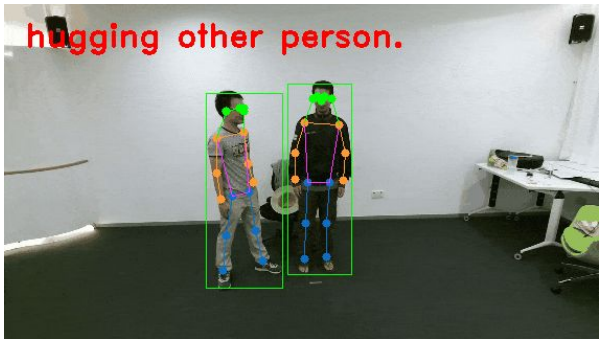
3) Google Colab
Cuda compilation tools, release 11.1, V11.1.105
gcc (Ubuntu 7.5.0-3ubuntu1~18.04) 7.5.0

opencv
pytorch
torch==1.9.0+cu111
torchvision==0.10.0+cu111
.
.
.

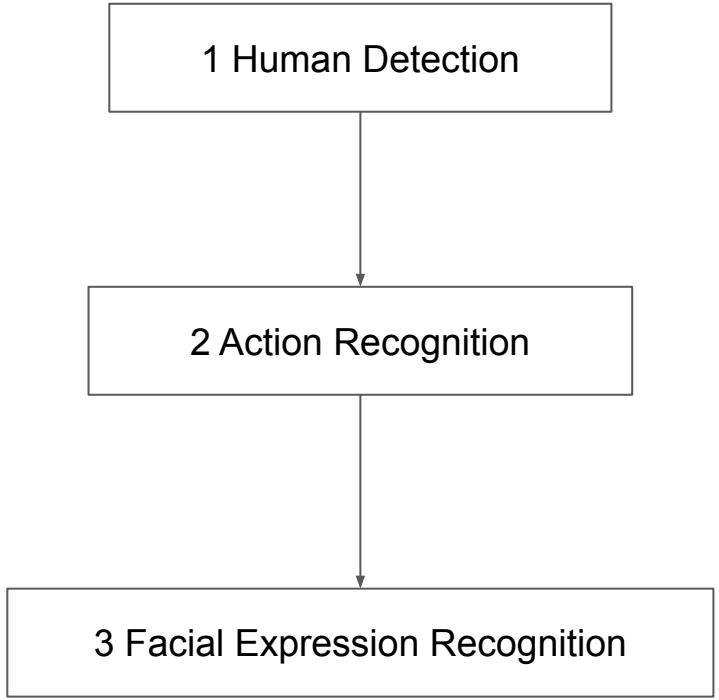
Datasets Kinetics400 https://paperswithcode.com/dataset/kinetics

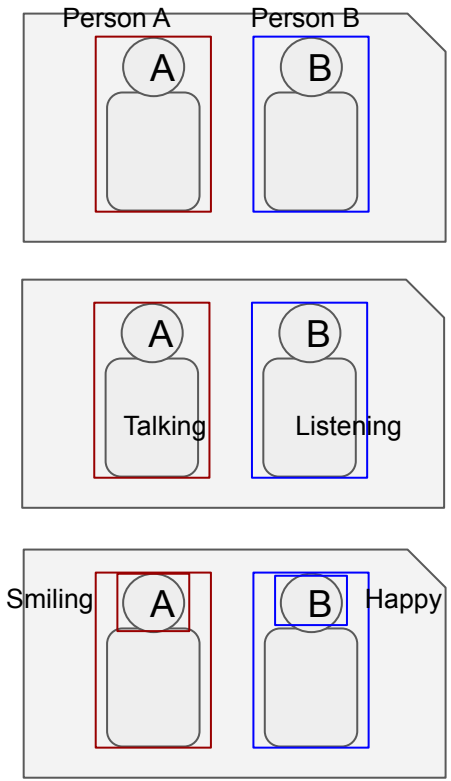


Models mmaction2 https://mmaction2.readthedocs.io/en/latest/



Step 1: Design Recognition Model





Step 2: Personality Questionnaire

Big Five Personality Test

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

The selection criteria are whether the question can be checked with yes (1) / no (0) as binary classification with visual evidence, among **50 questions**.

Section excluded: No.5 Stress Tolerance, No. 3 Extroversions

This project aims to analyze recognizable features

to assess **1 Agreeableness, 2 Conscientiousness, and 4 Openness**

Step 2: Personality Questionnaire

	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.	○	○	○	○	○

Step 2: Personality Questionnaire

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 (1-) - facial, activity
26. Have little to say(1-) - similar to no.6
46. Am quiet around strangers (1-)
(36. Don't like to draw attention to myself(1-)

(1+)
11. Feel comfortable around people(1+) - facial
21. Start conversations(1+) - activity
31. Talk to a lot of different people at parties(1+) - activity(asking to others)
(41. Don't mind being the center of attention(1+)

(2-)
12. Insult people(2-) - facial
22. Am not interested in other peoples' problems(2-) - facial, activity(eye contact)
32. Am not really interested in others(2-) - similar to 22

(2+)
7. Am interested in people(2+) - activity(asking to others)
(17. Sympathize with others' feelings(2+) - facial reaction
(42. Feel others' emotions (2+)
47. Make people feel at ease (2+)

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

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

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

(5. Have a rich vocabulary(5+) - tone fluency)

Implementation

Checklist

400 labels in kinetics400, 8 labels in facial ex.
> don't need them all

Check only labels that can be used for determining personality test questions

<< Checklist >>

Activity (kinetics label)

asking / answering questions

talk to / listen to

watch (eye contact)

applauding, clapping

laughing

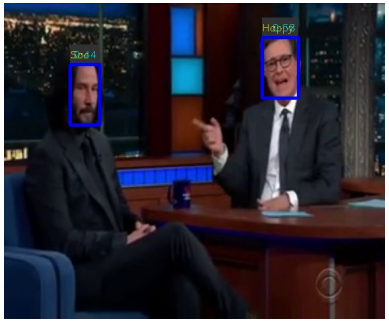
Facial expressions

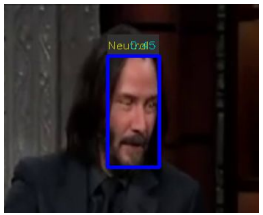
Happy

Recognition Models

Action Recognition

Facial Expression Recognition





Method to Calculate the Points

Agreeableness: Asking questions or Talking

Max counts: Divide the whole clip numbers(n) with the number of people in the video(2)

Divide into 5 points (Very Low, Low, Neutral, High, Very High)

Point1 0 - n/5*2 | **Point2** (n/5*2)+1 - (n/5*2)*2 | **Point3** (n/5*2)*2+1 - (n/5*2)*3 |

Point4 (n/5*2)*3+1 - (n/5*2)*4 | **Point5** (n/5*2)*4 +1 - n |

Check both 'Asking Questions' and 'Talking' >> Calculate the average.

Ex. Total 60 clips

Point 1: 0 - 6 clips
Point 2: 7 - 12
Point 3: 13 - 18
Point 4: 19 - 24
Point 5: 25 - **30**(or more)

Conscientiousness: Listening, Watching / Applauding, Clapping or Laughing

Max counts: Same(n/2) for Listening & Watching

Just add 1 point when other reactions are detected (n/20) times

(Nothing happens when already MAX 5 point)

Ex. Total 60 clips

Point 1: 0 - 6 clips
Point 2: 7 - 12
Point 3: 13 - 18
Point 4: 19 - 24
Point 5: 25 - **30**(or more)

Add 1 point if detected **3 times**

Method to Calculate the Points

Openness: Happy facial expression

Max counts: The whole clip numbers (n)

Point1 0 - n/5 | **Point2** n/5+1 - (n/5)*2 | **Point3** (n/5)*2+1 - (n/5)*3 |

Point4 (n/5)*3+1 - (n/5)*4 | **Point5** (n/5)*4+1 - n |

Ex. Total 60 clips

Point 1: 0 - 12 clips
Point 2: 13 - 24
Point 3: 25 - 36
Point 4: 37 - 48
Point 5: 49 - **60**

Result Example

10 min video >> 60 clips with 10 sec

Textual Summary

Person A talked in 26/30 rate, reacted in 8/15 rate. Also, Person A was in good mood for 55/60 rate. Therefore, Person A's Agreeableness is High(4), Conscientiousness is Middle(3), and Openness is Very High(5).

Person B talked in 28/30 rate, reacted in 5/15 rate. Also, Person B was in good mood for 50/60 rate. Therefore, Person B's Agreeableness is High(4), Conscientiousness is Low(2), and Openness is High(4).

Evaluation

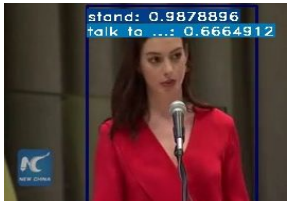
Evaluation Method 1

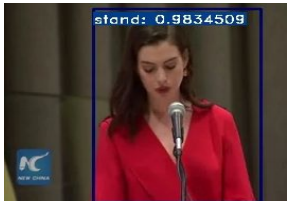
Comparison of label distributions measured in one-person and two-person videos

Test Samples: 5 video pairs of celebrities (speech vs 1 on 1 interview)

Evaluation Method 1

Speech Video

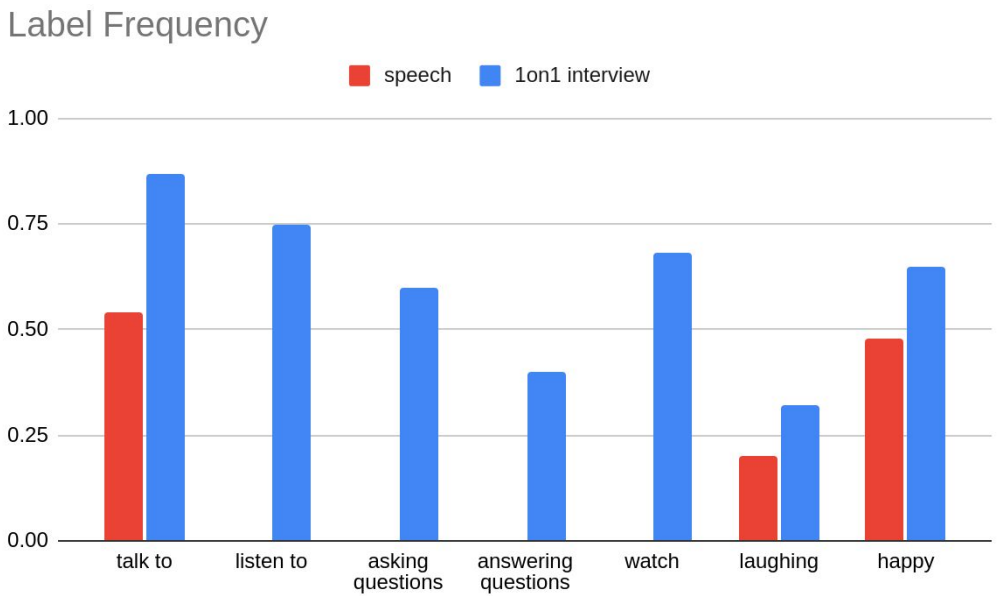




1 on 1 Interview Video

Video with 2 people having conversations shows more labels (Indicators for assessing Personality)

Evaluation Method 1



1. Most of the action labels were measured more frequently in 1 on 1 interview videos than speech videos.

2. For some of the labels (such as ‘listen to’, ‘asking questions’, ‘answering questions’, ‘watch’) , it was not possible to recognize without opponent person.

Evaluation Method 2

Human Evaluation

Test Subjects : 4 people age ranging from 20 - 25 years old

Provide 4 Sample Videos (10min) 

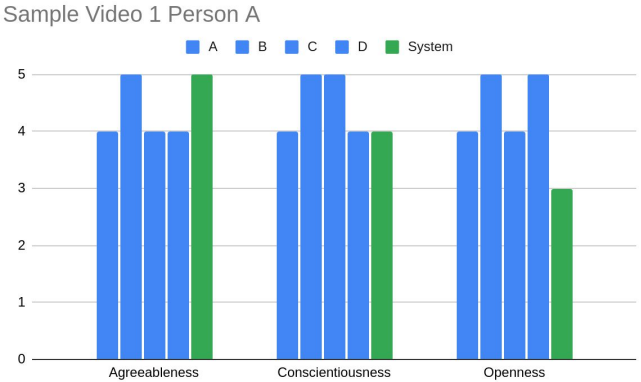
Check who is more applicable when assessing

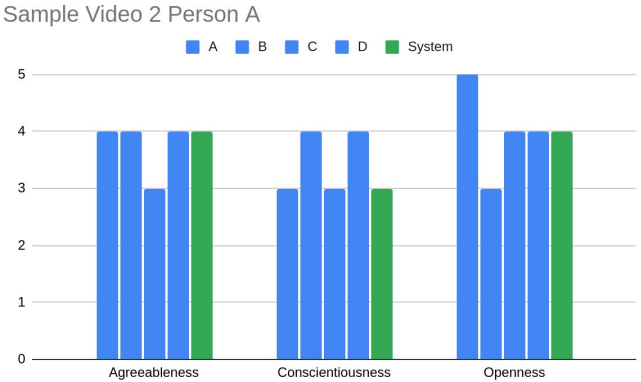
1 Agreeableness, 2 Conscientiousness, and 3 Openness.

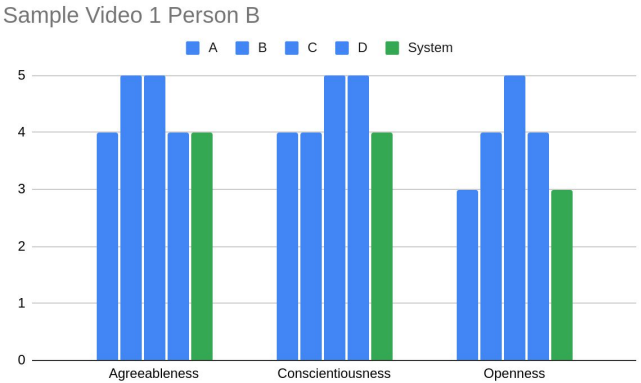
Collect the human-assessed personality via Google Form

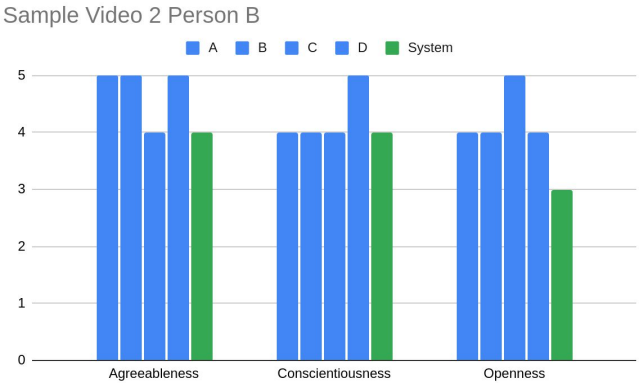
Compare the inferred result from the system

Evaluation Method 2



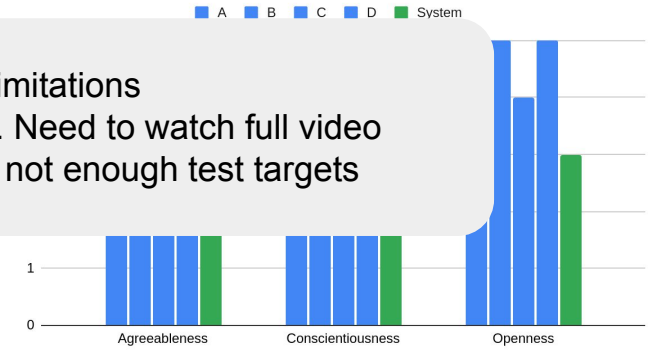






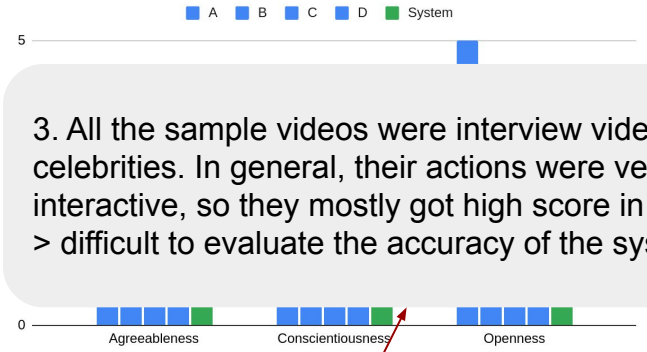
Evaluation Method 2

Sample Video 1 Person A



Limitations
1. Need to watch full video
> not enough test targets

Sample Video 2 Person A

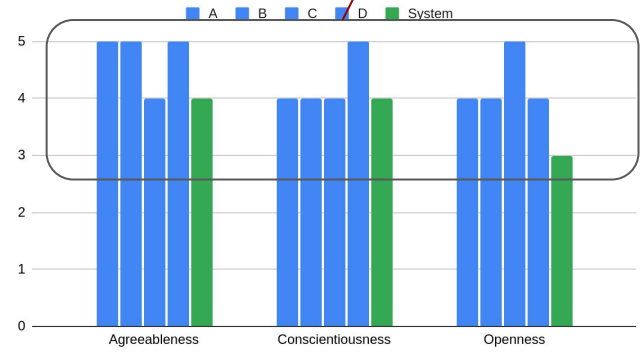


3. All the sample videos were interview videos of celebrities. In general, their actions were very interactive, so they mostly got high score in all.
> difficult to evaluate the accuracy of the system

Sample Video 1 Person B

2. Each person assesses differently. The number of Test Targets is too low to cover errors.

Sample Video 2 Person B



Conclusion

- Created 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 & quantity of indicators(labels) for assessing personality when the person is having conversation
- Accuracy assessed by Evaluation 2 is hard to trust due to low number of the test targets / not appropriate sample data

Future Work

Limitations

- Just for **2** people video
- Using visual data only
- Unreliable **Human Evaluation** Result

Features that I wanted to add more in future work…

- **Voice** Recognition Model
- More **Labels** by Training (Other reactions such as …)
- Application working with webcam (Shoot video and Get result immediately)
- Another Evaluation Method
  - take videos of the test targets, (not celebrities) to collect more diverse personalities & also check their own personalities by themselves with the Big Five Test.

References

[1] Apostolidis, Evlampios & Adamantidou, Eleni & Metsai, Alexandros & Mezaris, Vasileios & Patras, Ioannis. (2021). Video Summarization Using Deep Neural Networks: A Survey. https://arxiv.org/pdf/2101.06072.pdf

[2] Haq, Hafiz Burhan & Asif, M & Bin, Maaz. (2021). Video Summarization Techniques: A Review. International Journal of Scientific & Technology Research. 9. 146-153. https://www.ijstr.org/final-print/nov2020/Video-Summarization-Techniques-A-Review.pdf

[3] 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 https://www.sciencedirect.com/science/article/abs/pii/S0045790612002364

[4] Z. Yue Temporal Action Detection with Structured Segment Networks, arXiv:1704.06228 https://arxiv.org/pdf/1704.06228v2.pdf

[5] mmaction2 https://github.com/open-mmlab/mmaction2 | Benchmarks https://mmaction2.readthedocs.io/en/latest/benchmark.html | Supported Datasets https://mmaction2.readthedocs.io/en/latest/supported_datasets.html | Model Zoo https://mmaction2.readthedocs.io/en/latest/modelzoo.html

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

[7] Facial Expressions Recognition - Real-time Convolutional Neural Networks for Emotion and Gender Classification https://arxiv.org/pdf/1710.07557.pdf

[8] Sample Videos https://www.youtube.com/watch?v=8u6zR4rIF5Y https://www.youtube.com/watch?v=L71glMiI83k https://www.youtube.com/watch?v=fl0x9Swgfgw