Differences in Visual Context with Near-Identical Textual Taggings in COVID-19 Videos from China and the US

Chen, Yu-Shih
Department of Computer Science
Columbia University
yc4001@columbia.edu

Directed Research Report - Fall 2021
Advisor: Prof. John R. Kender

January 4, 2022

1 Abstract

This work is part of the ongoing project ”Tagging and Browsing Videos According to the Preferences of Differing Affinity Groups”[1] where the project is directed by Professor John R. Kender and aims at analyzing differences in textual tags preferred by different affinity groups through news and related videos. Specifically, past work done within this project revolved around searching for near-identical frames between the videos from different affinity groups and comparing the difference between the textual tags associated with them. However, in this study, we propose to gauge at the differences in visual context, instead of textual ones. We reverse the pipeline to finding frames with similar textual tags and trying to find visual differences. We looked at COVID-related news videos from China and US, specifically, in this study.

2 Introduction

From the invention of VCRs to the YouTube channel, videos have since become ubiquitous elements in our lives. Videos come in as seamless streams of sequential frames (which are essentially still images) that provide a rich experience of visual movement on a screen to the viewers. As of 2021, there are around 149 Billion videos on YouTube, and around 80% of both adult male and female have access to and watch YouTube – and this is just YouTube alone. There are many other platforms (e.g., cinema, Netflix, etc.) that target various kinds of videos (e.g., TV shows movies). Videos have not only become an indispensable part of modern humans’ entertainment, but they are also one of our main segues of expressing our cultures.

News reports and videos are one of the most straight forward ways different countries use to communicate events going on in the world to their citizens. Furthermore, since news videos are usually administered by the authorities of the representing country, we hypothesize that news videos from different affinity groups should contain cultural differences. News reports consist not only of the visual content but also the textual component (from what the reporter reports orally). Past studies in this project has been revolving around comparing between the textual context between near-identical frames between two videos from different affinity groups. However, in this study, we propose to go the other way around – instead of comparing textual context between near identical frames, we are interested in comparing the visual context between near identical texts. We believe the results from this approach would add yet another layer to the overall analysis of the project ”Tagging and Browsing Videos According to the Preferences of Differing Affinity Groups.”

[1] See http://www.cs.columbia.edu/ jrk/NSFgrants/videoaffinity/ for more details

In particular, we have focused on comparison between news videos related to COVID-19. Furthermore, we have focused on comparing between news reports from China and from the US. For the Chinese news videos, we referred to CCTV, and for US news videos, we referred to ABC news.

3 Methods

3.1 Data

We simply resolved to news videos on YouTube. For English news videos, we used the ABC news channel, and for the Chinese news videos, we used the CCTV channel on YouTube. We collected 30 videos from both the Enlgish and Chinese source (60 videos in total). On specific methods for processing and preparation for analysis, please refer to the next section.

3.2 Preparation for Analysis

In this study, we aim at finding differences between the visual context with a corresponding text in two different videos from different affinity groups. More concretely, we chose to use the keyword ”19” and ” 冠”[2], which serve representations for ”COVID-19” in English and Chinese.

In other words, we have to extract the frames that correspond to when the reporter mentions the keywords for both sources, and then we could conduct a comparison between the resulting frames from both sources. The following illustrates the steps for preparing the frames from videos from both sources.

English source (See Figure 1 as well):

1. Use YouTube API to download the news video and the corresponding caption (transcription)
2. Time-stamp the transcription
3. Run custom script to extract the frames that correspond to when the keyword ”19” was mentioned in the time-stamped transcription

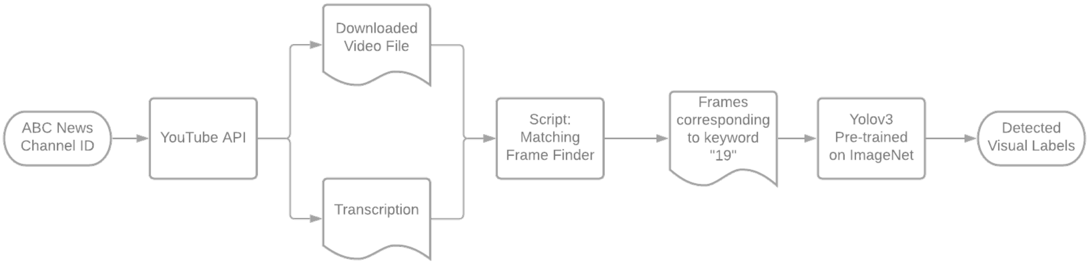
ABC News Channel ID
YouTube API
Downloaded Video File
Transcription
Script: Matching Frame Finder
Frames corresponding to keyword "19"
Yolov3 Pre-trained on ImageNet
Detected Visual Labels

Figure 1: **English News Videos Frame-Keyword Extraction Preparation Pipeline.** Prediction with Yolov3 will be explained in the next section.

Chinese source (see Figure 2 as well):

1. Use YouTube API to download the news videos
2. Use Google Speech-to-Text API to perform transcription on the Chinese news videos
3. Time-stamp the transcription
4. Run custom script to extract frames that correspond to when the keyword “冠” was mentioned in the time-stamped transcription

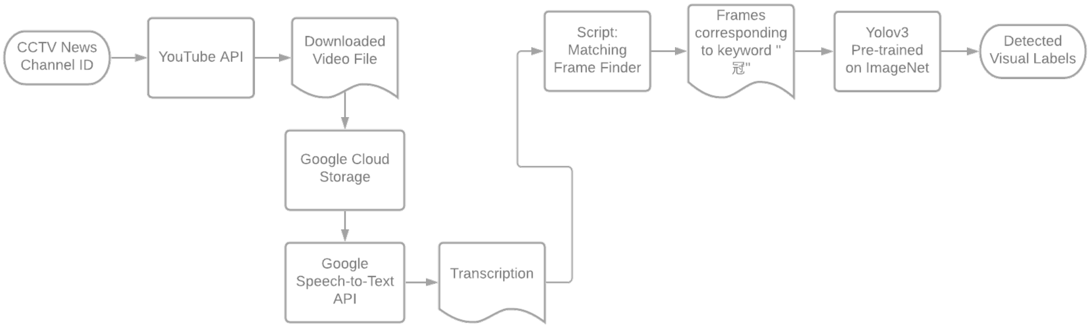
CCTV News Channel ID
YouTube API
Downloaded Video File
Google Cloud Storage
Google Speech-to-Text API
Transcription
Script: Matching Frame Finder
Frames corresponding to keyword " 冠"
Yolov3 Pre-trained on ImageNet
Detected Visual Labels

Figure 2: **Chinese News Videos Frame-Keyword Extraction Preparation Pipeline.** Prediction with Yolov3 will be explained in the next section.

Note that for Chinese source pipeline, there is an extra step of using the Google Speech-to-Text API for transcription. This is due to the news videos downloaded using YouTube API for the Chinese news videos not having captions that came with them. Hence, an extra step of automatic transcription was required (as shown in the diagram as well).

All of the pipeline implementation was achieved using Python and its libraries.

3.3 Approach for Comparison

After we extract the frames that correspond to the keywords (for COVID), we are interested in finding any visual differences in the frames from both English and Chinese sources. To approach this, we decided to compare the objects that appear in the frames. We chose the Yolov3 network[1] (see Figure 3 for architecture) pretrained on the ImageNet dataset. We chose Yolov3 because of its popularity for object detection (both real-time and offline) and ImageNet due to it's vast variety of classes. We simply ran the network on the frames extracted from both the Enlgish and Chinese sources (prepared as described in the ”Preparation for Analysis” section), and plotted a histogram for both for comparison.

4 Results & Discussion

After running Yolov3, we obtained the following counts for the objects that appeared in the frames (corresponding to the keywords) from English source and Chinese source – we show the top 10 appeared objects (according to the detection network). See Table below.

Object Name	Counts
Windsor Tie	240
Web site	227
Ski	210
Streetlight	204
Hospital bed	126
Screen	102
Bulletproof vest	87
Desktop computer	42
Television	36
Home theater	33

Object Name	Counts
Screen	267
Web site	257
Television	225
Windsor Tie	219
Desktop Computer	126
Bulletproof vest	81
Theater curtain	78
Home theater	69
Hospital bed	39
Hair spray	36

We can make some observations from the object counts. First of all, we see some very generic objects on the highest counts (i.e., windsor tie, web site, screen, etc.) which represents the talking heads (See example frames

[2] COVID-19 is “新冠肺炎” in Chinese, and “新” is too generic which could appear anywhere, so we chose “冠”

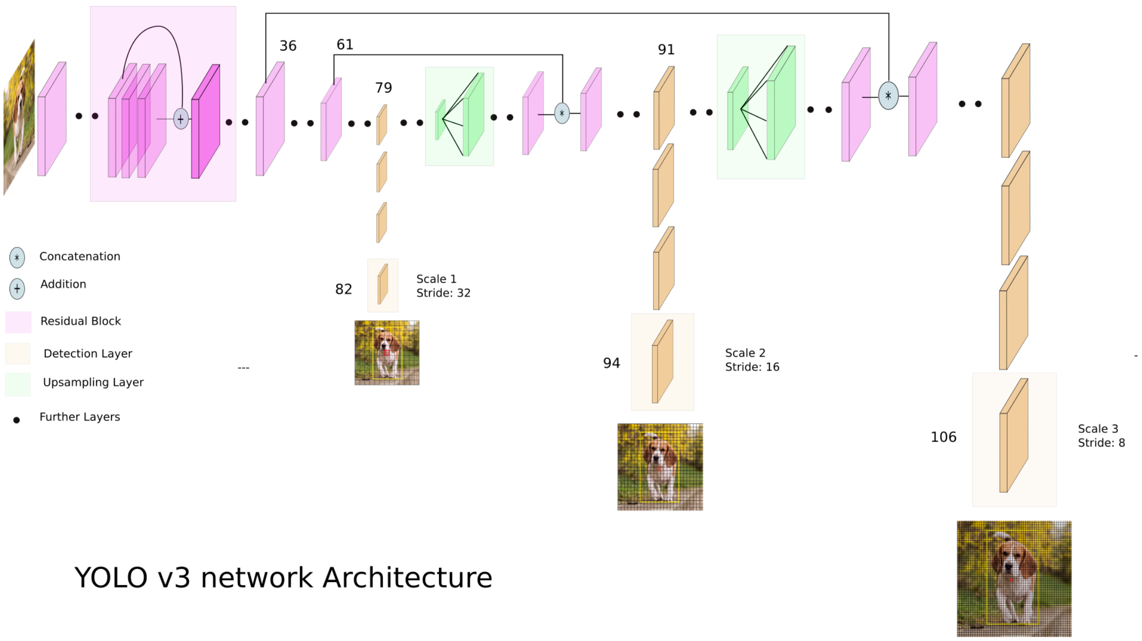

Figure 3: **Yolov3 Architecture Illustration**

for this below). They have the highest counts because we did not remove the talking heads, and presumably, they come out a lot in the videos. Hence, for analysis, we will focus on the items that come after those.

For the English source, we see some interesting objects appear such as Hospital bed, Bulletproof vest, and streetlight. Hospital bed makes sense, since the videos are about COVID. Streetlight might indicate that a lot of the scenes were taken outside of a building, on the streets. We can also observe that there are more Windsor tie detected in the English source. After some manual inspection, we figured that the English source videos corresponding to keyword has more scenes of talking heads and also showing single authorities talk (see Figure 4). Bulletproof vest is also very interesting. However, after some manual inspection, they were referring to the medical suits (including a shield-like face mask) that the medical personnel wore (see Figure 6). See example frames below.

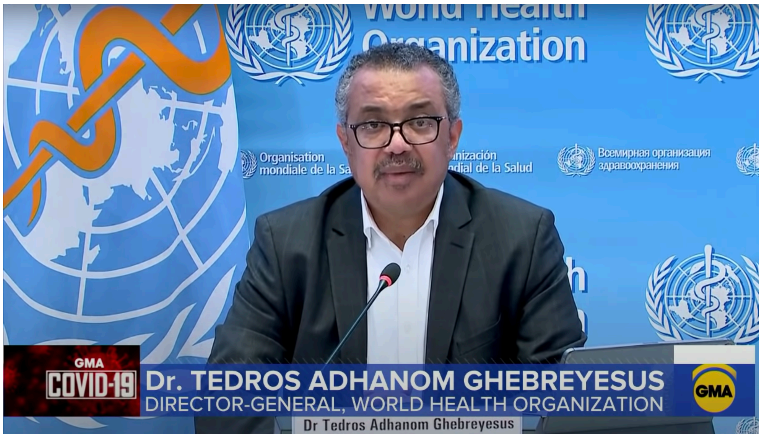
World Health Organization
GMA
COVID-19
Dr. TEDROS ADHANOM GHEBREYESUS
DIRECTOR-GENERAL, WORLD HEALTH ORGANIZATION
Dr Tedros Adhanom Ghebreyesus
GMA

Figure 4: **(Eng Source) Single authority speaking**

For the Chinese source, we see some common and also different objects appear (compared to the English source). For common objects, we can see Hospital bed and Bulletproof vest. However, something important to note here is that the Bulletproof vest is actually referring to both medical suit and actual police vests (see Figures 6 & 7). This indicates that the scenes that appear in the Chinese news videos corresponding to the keyword

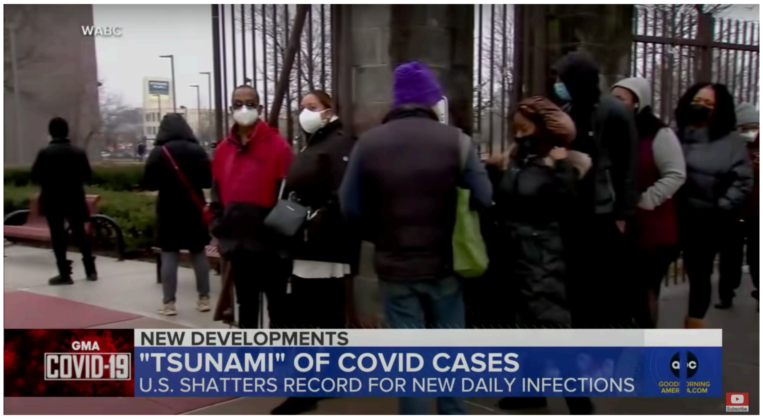

Figure 5: **(Eng Source) Outdoor scene**

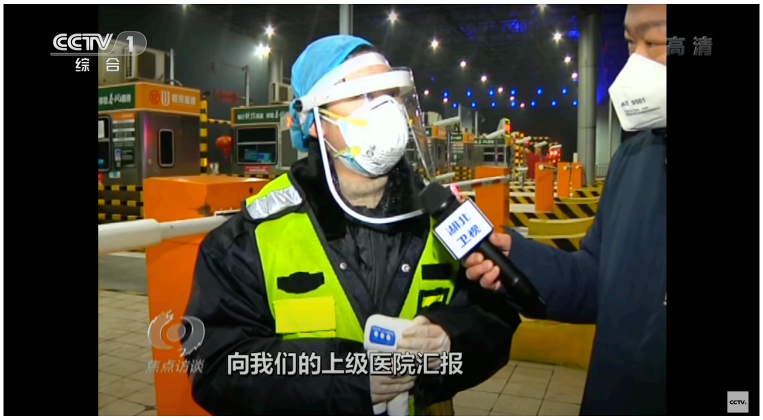

Figure 6: **(Chin Source) Medical Suit Misclassified as Bulletproof Vest**

could also show signs of police authorities coming out in the scene (might resemble dominance, to some extent). Additionally, we see theater curtain appear. Again, through manual inspection, theater curtain actually also refers to conference rooms, but in a greater scale. This is detected when a scene of many authorities appear at once in a large conference room (see Figure 8). See example frames below.

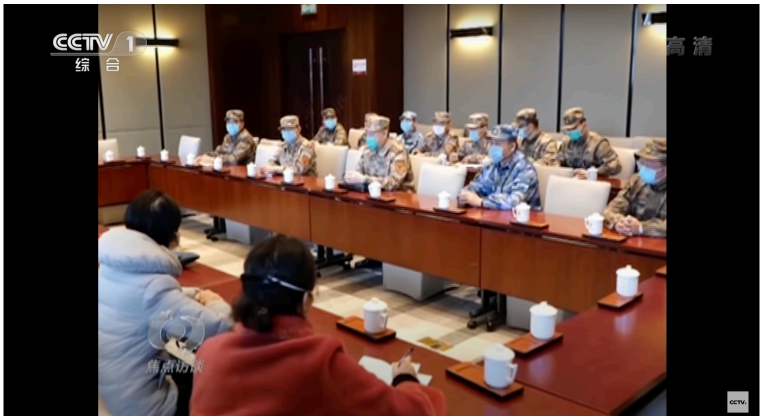

Figure 7: **(Chin Source) Example of Chinese Police Force in the Scene**

Comparing both sources, we can see some interesting (maybe cultural) differences. For the English source, the objects seem to revolve more on medical/hospital facilities, outdoor scenes, and small meetings. On the other

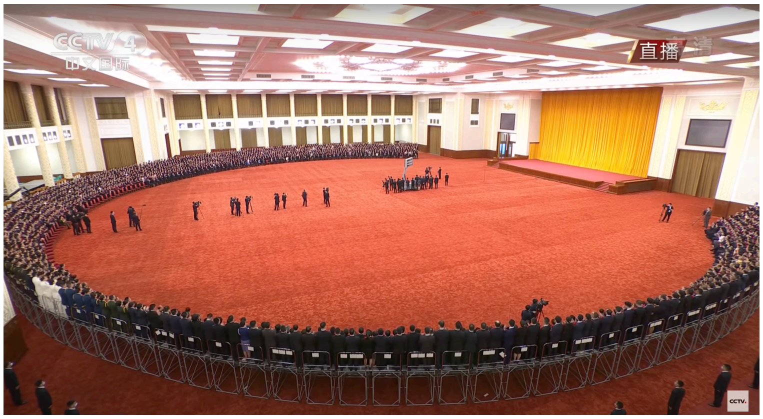

Figure 8: **(Chin Source) Large Conference Meetings**

hand, for the Chinese source, the objects seem to revolve more on greater scale meetings, less medical/hospital counts, and more indoor scenes. This might be an indication that the Chinese news videos tend to show authorities gathering in large rooms having meetings, and even some police-related scenes (which might, to some extent, resemble authority dominance), where in comparison, the English videos tend to focus more on the medical/hospital scenes, smaller meetings on a more individual level, and citizens.

5 Conclusions & Future Work

From the object counts and comparison we conducted, we have reached some interesting observations. It seems that videos from the English sources tend to focus more on hospital, medical, and individual-level scenes when videos from Chinese sources tend to focus more on large-scale meetings (authorities), some police forces, and less medical scenes (still present, though). We have to note, though, that this conclusion is not final and has a lot of room for improvement. It is rather preliminary.

As this is a new direction for the project, there are a lot of possible improvements for this approach. First of all, we could favor a larger quantity of videos for conducting a more meaningful comparison. Second, we could choose other object detection architectures for representing each frames. We could also try to use other keywords and compare between different frames from different keywords.

Furthermore, we could experiment with other ways of comparing between the frames. This could present to be more challenging, as the comparison has to make sense to humans as well. In other words, abstract comparisons (e.g., MSE between two frames) are not feasible as they fail to provide any meaningful message in the end. Nonetheless, exploring innovative ways of comparing visual context can also be another interesting branch to this project.

References

[1] J. Redmon and A. Farhadi. Yolov3: An incremental improvement, 2018. cite arxiv:1804.02767Comment: Tech Report.