AY 2019

Fluency in Real-Time Video Streaming by Learning Human Perceptive Traits to Reveal the Expected Section in Outstanding Quality

A Thesis Submitted to the

Department of Computer Science and Communication Engineering,

The Graduate School of Fundamental Science and Engineering of

Waseda University

In Partial Fulfillment of the Requirement for the

Degree of Master of Engineering

January 29th, 2020

By:

Chiang shu chiao

(511FG09-5)

ADVISOR: PROF. TATSUO NAKAJIMA

Table of Contents

ABSTRACT

Currently, the quality of digital media and the quantity of contents are both increasing rapidly. For instance, watching e-sport competitions often suffers from unstable bandwidth, which causes the video to stutter or have a low resolution. In this situation, users will have a negative experience. Many situations can cause problems of congestion in real-time applications or 3D displays. To solve this kind of problem, we attempt to determine an inverse solution according to the path. This project adopts a reverse operation that reduces necessary data but maintains the same quality perception of user experience by utilizing the characteristics of the human vision and brain. To explore our approach, we develop a prototype that changes the resolution of the image according to a user's habit and shows the part in focus clearly while leaving the resolution of the background lower. To select optimized sub-image in pictures with higher quality and achieve a lower transmission requirement, full quality is reduced. This will allow the user experience smoother streaming when there is congestion or unstable situations. Then, we conduct a preliminary user study to investigate some future directions and explore some potential flaws.

ACKNOWLEDGEMENT

I would first like to thank my thesis advisor Prof. Tatsuo Nakajima of the Department of Computer Science and Engineering at Waseda University. The door to Prof. Nakajima's office was always open whenever I had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to acknowledge Mr. Zebo of the Department of Computer Science and Engineering at Waseda University as the first counselor of this thesis to help me start this research, and I am gratefully indebted to him for his very valuable assist on this thesis.

The last, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and though the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you all above.

Author

Chiang shu chiao

CHAPTER 1 INTERDUCTION

Regardless of whether virtual or real, required resolution is steadily increasing. Moreover, many applications tend to develop the 3D aspect, which requires many times the data flow of 2D [7]. Therefore, extracting significant areas will create an efficient method to reduce congestion and instantly increase demand. For example, in real-time sport races, there may be many people linking at the same time despite it not being a busy period. In that case, the user may experience intermittent loading or a low resolution screen, as shown in Figure 1. It just compares the both movie screen shoot at 4K and 480 pixels. We can be easy to distinguish the difference from them. Even when communication equipment provides greater bandwidth, the data requirement will also increase with the bandwidth due to more devices linking or a higher data consumption cost. For metaphor, building more roads is not the solution to traffic jams, and changing the usage habit of transportation is necessary [1]. And this concept also inspires me to change the method of accessing video.

a) The screenshot at 1080p

b) The screenshot at 360p

Figure 1. When users view these on large size monitors, they obviously note the difference between their sharpness[9]

If we can provide a more fluent model in various conditions, it will relieve pressure for route and improve performance. In addition, in 3D, we create a hypothetical scenario. In the future, car windows may have augmented reality services, as shown in Figure 2, which can show the information about landmarks, stores, or traffic warnings. However, drivers may not want to display all things all the time, so we can predict the driver's interest and show what he/she needs at that moment.

In general, we also want to use it in movies, but it will be very difficult to recognize, identify, and filter images because every scene will have underlying connotations that are open to interpretation. And some elements accompany by other object would be indispensable. That will be the flaw that we only reference the attributes of image. Therefore, if we want to make this project work on multiple media, it still need more research in diverse fields to have better understand in connotations.

Figure 2. Showing out the concept map of car window equips the AR device

This project presents a system that will show everything in best quality that the bandwidth allows. It develops a new storage architecture to improve the problem at current version and let this project have better performance. When the network transmission volume decreases, it will turn on a system to sort the hierarchy of sub-objects, collect a user's eye gaze on the screen, find the area where the user is looking and train the personal interest database by collecting gaze information. For the purpose of assuring the video can run smoothly in any situations with the same perception to its user, it can also record the effect on the user after using this system. Based on the above approach, we can determine how to enhance this prototype.

CHAPTER 2. RELATED WORK

Section 2.1 STREAMMING & RESOLUTION

Streaming and resolution are the basic elements in this project. Pixel is the fundamental units which descripted in channel 3 in video frame and the overall evaluation level of each frame is resolution. Then, media streaming is the way to transfer those data. Therefore, we need to be clear with their definitions.

Streaming media technology enables the real time or on demand distribution of audio, video and multimedia on the Internet. Streaming media is the simultaneous transfer of digital media, so that it is received as a continuous real-time stream. Streamed data is transmitted by a server application and received and displayed in real-time by client applications. These applications can start displaying video or playing back audio as soon as enough data has been received and stored in the receiving station's buffer.

Resolution refers to the number of pixels in an image. Resolution is sometimes identified by the width and height of the image as well as the total number of pixels in the image. For example, an image that is 2048 pixels wide and 1536 pixels high (2048 x 1536) contains 3,145,728 pixels. You could call it a 2048 x 1536 or a 3.1 Megapixel image. As the megapixels in the pickup device in your camera increase so does the possible maximum size image you can produce.

Section 2.2 CODEC

A video codec is a software or sometimes a piece of hardware that compresses and decompresses digital video. In other words, a codec processes raw digital video and stores it in a stream of bytes. It converts uncompressed

video to a compressed format to take up less space on your computer and vice versa. A video codec is usually identified by MPEG, DivX, HEVC, etc.

In fact, there is a huge list of codecs that you can find. A majority of videos are encoded with the help of most popular codecs mentioned above and can be played with almost any multimedia player. However, there are rare codecs from special video cameras that can be viewed with VLC or a similar player with a proper codec library.

A video codec isn't the same as a video format or container. It much close to be like an algorithm. And a container is a bundle of files. Inside it, you can find data that has been compressed by using a particular codec. For example, an AVI file can contain video compressed by XviD, or DivX, or MPEG-2 codecs. Usually, a container comprises a video and audio codecs, plus it can also contain other files like subtitles and chapters. Popular video formats or containers are AVI, MP4, WMV, MKV, MOV, FLV, etc.

Section 2.3 ADAPTIVE BITRATE STREAMING (AUTOMATICALLY ADJUST VIDEO QUALITY)

Adaptive bitrate streaming is a technique for dynamically adjusting the compression level and video quality of a stream to match bandwidth availability. In order to provide a better viewing experience based on internet bandwidth.

Older video streaming approaches relied on distributing a fixed bitrate video stream. If your network connection could not support that bitrate, you could not watch the video without dramatic buffering, if at all. With ABR, you can stream video across the Internet, with both point to point streaming and OTT services to multiple devices.

For point to point streaming, ABR can mean adapting a single RTMP or SRT stream to fit the available bandwidth between two devices such as an encoder and decoder. For point to point video streaming, an encoder needs to be able to adapt the compression level of a stream in real-time, as available bandwidth is constantly changing. This is also known as network adaptive encoding.

For OTT services, ABR will usually rely on an ABR packaging protocol such as HLS or MPEG-DASH where multiple streams are defined by profiles such as low, medium, and high quality. The ABR streams are divided into chunks of video, between 1 – 15 seconds, so that individual viewing devices can dynamically pick and choose the video chunk that best fits available bandwidth at a given time. ABR streaming for OTT requires the use of an encoder or transcoder which can encode a single video source at multiple bitrates.

Section 2.4 IMAGE RECOGNITION

Image recognition is a computer vision task that works to identify and categorize various elements of images and/or videos. Image recognition models are trained to take an image as input and output one or more labels describing the image. The set of possible output labels are referred to as target classes. Along with a predicted class, image recognition models may also output a confidence score related to how certain the model is that an image belongs to a class.

Section 2.5 COMPREHENSIVE NEURAL NETWORK

A Convolutional Neural Network is a deep learning algorithm which can take in an input image, assign importance to various aspects or objects in the

image and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms.

While in primitive methods filters are hand-engineered, with enough training, CNN have the ability to learn these filters and characteristics. The architecture of a CNN is analogous to that of the connectivity pattern of neurons in the human brain and was inspired by the organization of the visual cortex. Individual neurons respond to stimulate only in a restricted region of the visual field known as the receptive field. A collection of such fields overlap to cover the entire visual area.

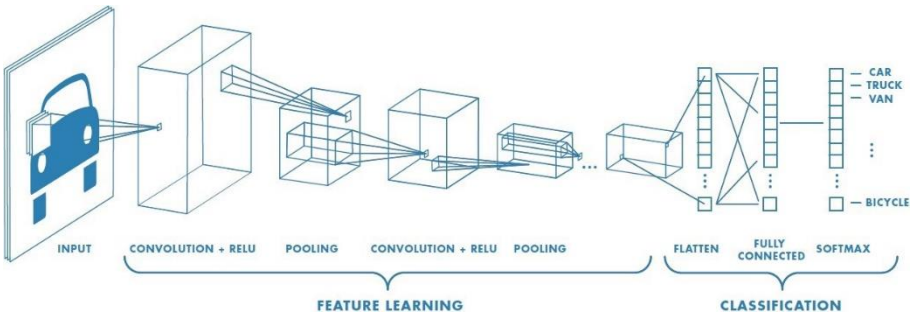


Figure 3. The algorithm that converting the image to be the series data. And these series data could be the deep learning training source[10]

Section 2.6 IMAGE PROCESS IN VIDEO

Scientists have studied human visual attention for decades. Some of them focus on understanding the image data. Others tried to understand the temporal effect of eye motion in videos. Others have attempted to understand the behavior within the shot and build a high-level theory. Since then, considerable progress in image saliency has been proposed, but less work has been performed on video saliency. Some researchers working on video saliency have built

methods by narrowing the thought focus to a small frame of candidate gaze location or having a higher result by transitioning over time in the video field [2]. In addition to the interaction of viewing and video display, considerable information needs to be considered; visual attention is not limited to analyzing pictures or image processing. The aim of an objective image quality assessment is used to evaluate the quality of pictures or videos as a human observer. Previous studies have investigated the content of pictures related to human behavior [3]. However, machine learning technologies have flourished recently. This shows better performance in processing human traits [4]. In previous researches on this issue, most studies determine users' interests by analyzing their habits or studying which image or region attracts people's attention. We now change to the deep learning method to find the target.

Image resolution also plays an important role, and the resolution of streaming media shown is dependent on the network speed. Image resolution finds the best fit pixels of the frame for the client, so there are many different thresholds for increasing or decreasing the storage database [5]. We propose a modification to this structure such that the new unit will not be the frame but a sub-image of the frame with position information [6].

This work will integrate the above benefits; leading to an application evolved to another level that has more interaction with humans.

CHAPTER 3 APPROACH AND METHODS

Section 3.1 OVERVIEW

The basic flow in the proposed method is shown in Figure 3. The process requires obtaining the video at the beginning. First, video is placed in filters to segment and crop the image into several sub slices from a complete frame. Then, the eye gaze is obtained to indicate areas that are interesting to the user. Third, those images are saved into a data pool. Next, we compare the trained deep learning database and record the user's private interest orientation. Finally, a client part saves the information, and another server part is established to save those sliced images. Finally, the processed video is shown according to the above steps.

Section 3.2 STRUCTURE

Here lists out the elements in this prototype. The details in each step will be described in next section. As follows:

1. Segmenting and Cropping
2. Getting gaze information
3. Contrast data set
4. Building database
5. Recording traces

Those are not fully ordered, some of elements are parallel or divided operations. They are just the key points of this application and the main idea of algorithm.

Section 3.3 STEP DETAILS

A. Segment / crop image

In the first step, matrix operations need to be conducted, such as 1) shifting to reduce the gradient and trivial pixels in a single picture, 2) blurring the original image to make it simpler to capture, 3) fuzzing color blocks to create a boundary and 4) making preliminarily analyzing the image. Then, the raw information is roughly combined with the results to crop the main objects of video. This makes it easier to distinguish each item in the frame. Then, those target areas are defined in boxes, and their location information is noted (Figure 4a). These details are recorded in a temporary list for comparison with eye gaze location. The next step describes how to obtain the eye gaze location.

a) Finding the potential objects and segmenting them as the sub-image

B. Catch the gaze

The proposed design will learn what part is suitable for a user. Therefore, this step is combined with the first step. After we conduct the initial process that

segments out sub-objects in the screen, it obtains the user's gaze to select where the main interesting object is, which has to be packaged in each frame (Figure 4b). Then, only an interested area is showed in high quality, but comprehensive perception is still close to a full high quality picture (Figure 4c). So, the total required data is significantly less than the original full high quality picture. The selected object is stored in the database and will have priority if the same object appears again. This work helps us train the database to recognize the same object at a later time. The information also promotes the efficiency of segmenting and cropping images. Therefore, the first and second steps are mutually optimized.

b) Getting the gaze information

C. Match / classify label

We match those selected sub-images to the real underlying meanings; just as many things in the real world are rich in meaning, it will change according to cultural practices [8]. Thus, we should define some similar items in the set and

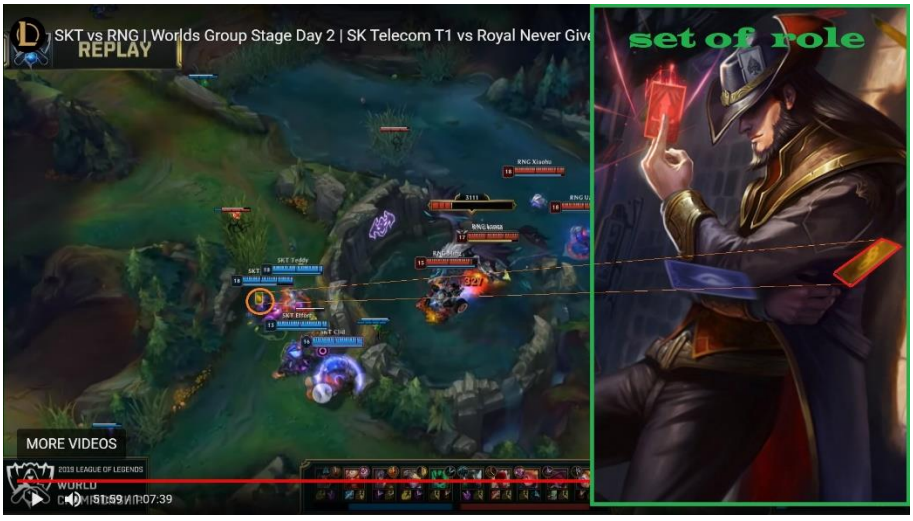
then teaching the machine to know which meanings are the same will be another challenge. There are two main works that need to be completed here. One is defining a new cluster when we identify an object that cannot be classified in the existing set. The other is assigning the object to the correct set (Figure 4d). These studies will require deep learning technology, as mentioned above, so that the system will be able to more quickly and efficiently recognize and analyze underlying messages in the real world. We want to find a positive method for labeling them. The next section explains these steps in more detail.

c) Showing the eye location in high resolution

D. Build database

This is the main part of the previous section because building the label set is the principal challenge. The difficult part is that many things have the same meaning but not the same appearance. We need to teach the machine to recognize them, create a new set from the data pool, and find the characteristics

of each set for classification. When we match new slices, it needs to mark those features because we use “feature matching” to compare the selected object and the sets in the database. At high speeds, which are often necessary, it is compared with only the common features of each set. Therefore, feature extraction will be trained by deep learning, which imitates how humans recognize an item, similar to how humans can rapidly determine implications from small clues.



d) Matching with database to build the user’s private pool

Figure 4. Descripting the detail of each section

E. Storage strategy

First, some basic concepts should be understood. We should learn how videos are saved in current streaming platforms. Currently, general frames are saved as complete pictures in different hierarchies of pixel levels. Frames are defined as the basic unit of a video. However, this system will require a slight

change so that it mitigates the unit to the sub-object, so we need more space to record them or store those sub-objects separately at the beginning. Both plans need to increase one dimension to link the data to others. When the sensor detects a fluctuation in bandwidth, it changes the resolution to gradually decrease from the object of interest.

Second, we also need to set up a private pool for recoding the image weight for each person. In detail of image weight, there would fetch out several sub images and refer the gaze information, we could track the interested area for each user. Therefore, we create a sequence to sort the image weight and this sequence is attached on personal account like recommended system when user view similar videos it would apply same image weight conesquence to select interested area for user.

The system presented in the paper needs a complex structure to store data and labels because it has considerable information in each frame. Then, we determine which type of strategy is suitable for this idea.

CHAPTER 4 IMPLEMENTATIONS

Section 4.1 ENVIRONMENT

This work was developed in the environment as follow:

1. Ubuntu 18.4 (VirtualBox)
2. OpenCV 3
3. Python 3
4. Keras in Tensflow 2
5. R-CNN (mask-rcnn-coco) [11]

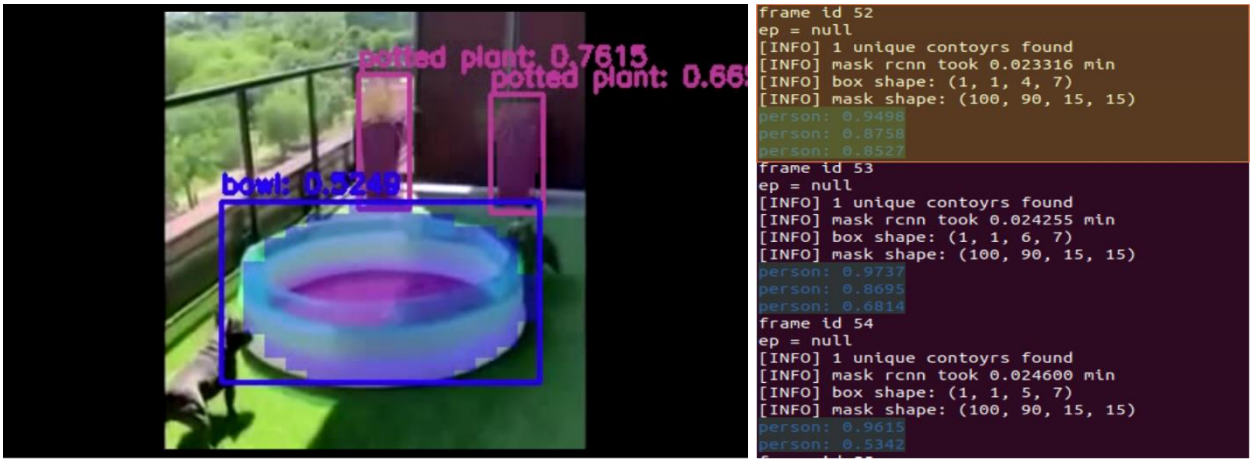
And devices as follow:

1. Dell laptop with GTX 1060
2. Tobii eye tracker 4c

And the source that I train for testing database is from free source bank "kaggle"

Section 4.2 INITIAL PLAN

Currently, the basic function of the experimental work has completed. This includes distinguishing out different things that include the features of filtering out the trivial color of images, segmenting out sub images by bounding the features of objects and cropping them out within a minimum frame box. Next, it matches the gaze locate to find the focus location and record the information back to a server for the purpose of iterative data updating. Then, we use the prepared deep learning database which is built by the Keras library in TensorFlow2. It would analyze pre-processed target pictures with the image sets in the database and list out results. However, auto-expansion of the image set is necessary for future work. This prototype can also track and collect users' gaze information, in order to make more effective predictions according to the users' habit. It shows the process result of raw video, as follow:



L) Showing out the processed raw video with framed sub objects, forecast labels and similarity

R) Printing out the advance process result and the forecast of objects in frame with raw video at figure 7L

Figure 5. Tasks on sever port

Section 4.3 PRODUCTION

We built the version for a user study. We set up the foremost stable database to represent the deep learning set. In other words, its iteration of recognition ability only trained when we set up the database (Figure 6).

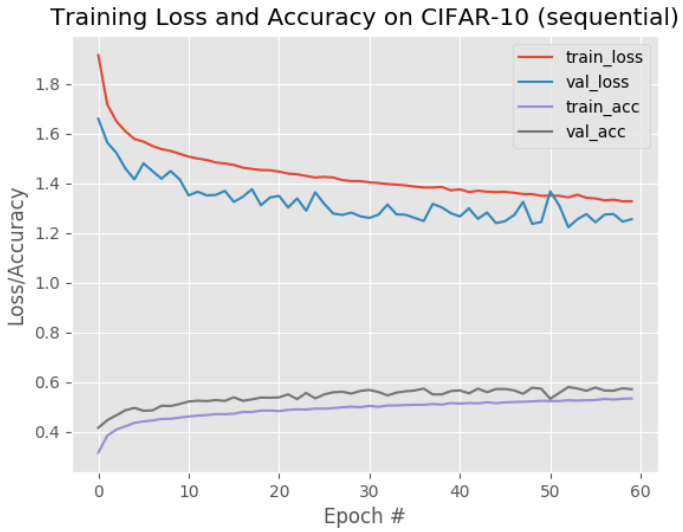


Figure 6. Recognition rate of own source (Loss is the penalty for a bad prediction, on the contrary Accuracy is the good prediction)

And used the mouse to replace a gaze tracker because it has better accuracy and a direct read source. What's more Tobii eye tracker 4c has not publish Linux SDK yet, so this part will leave to future work. Others have used the same structure and operation as we described above. This version can record all information, including the sub-image contain, split coordinates, object of gaze located and the caption of the object prediction. After the detecting procedures, it will show the processed video with high quality at vital parts. As follow:

L) Showing the gaze locaed section with high quality

R) Merging the gaze information into procedure and executing works

Figure 7. Procedure of selecting the vital sub image and it also be the performance of user study version

In the left picture, that is the main object in red box which segmented by machine and it parses there has a character inside. On the right picture, it prints the gaze address in red word and it is an effective operation, so this sub image's weight will be increase and system records its information. Combining both facts, it will show the vital section in high quality with detial descriptions like left one (Figure 7a). In contract, it will not display sub images distinct.

CHAPTER 5 EVALUATION

Section 5.1 PRELIMINARY RESULT

When working on this project, there are some problems while setting up environment such as Window OS is not friendly to build Opencv and Keras, so I develop on Ubuntu. Owing to this reason, I can't equip the eye tracker on this application at this moment. Combining the factors above, it should develop a more widely design since there are much various devices on the market. And its graphics operation need strong power, so we slight adjustment the structure. The consumption of this prototype is really high. Thus, it needed to be optimized at some features. However, its saving data transmission is considerable. I should promote its efficacy, but maintain its achievement.

Section 5.2 USER STUDY

In this section, we present some simple questions to the participants in the user study, as below:

1) How does the display compare with general videos?

2) Does this application help you?

3) Which part should be promoted to increase user interest?

4) what does the participant respond?

After we show the demo version, which includes an introduction of the concept and a trial of the prototype, depend on their career (programmer, video editor and common user).

First, we provide a survey to the programmers. After they used the system, they described two main concerns. They considered the speed to not be truly immediate. If this technology is going to work on real-time video, this problem needs to be solved. Therefore, we consider the main part where work on analyzing, segmenting and cropping should be built on the server part because those operations require very advanced devices to achieve real-time streaming. Only gaze tracking and data collection should be embedded in the client part. Then, we push this system to mobile devices because mobile devices have lower efficiency CPUs. Reaching the real-time goal will be the most significant challenge. Additionally, this system may sometimes ignore the supporting cast in favor of the main characters. This results in the related matters being ignored as well. The link between objects is too weak, which creates this defect. It is also a serious problem in general videos. Movies, music and videos all contain many meanings within every sub-image. However, the shooting technique is a topic for another paper. There are some scattered doubts, such as reducing segmentation in each frame, enlarging the minimum segment size, using a decreasing method to show the resolution around the main object from high to low, or implementing this technology in the gaming field.

We investigated a group of video creators in second. For these creators, the content of their videos is of utmost priority. Because media is the method through which they express their thoughts, they want to completely convey their ideas to the viewer. Therefore, when those potential users consider how the system works, they assess whether this technology would affect their product. Thus, they concentrate on object weight calculation and the weight of interactors in the feedback. For example, there is a scene of a competition in which we want

to know the main object and the competitor. There are some comments that indicate that the system should provide a function for the creator to set the weight when they edit the video. Thus, the creator can have better control of the connotations that they want to display to the viewer instead of ranking the weight by users' preferences, as it may cause communication errors. They were also interested in whether this system would create benefits for the video editor. For example, rendering the video to a normal format requires considerable time and storage. If this new video storage method can speed up the process, it would be welcomed by video creators. This potential application may lead to some innovations. How it combines with video editing or optimizing rendering functions would be another use for this system.

Finally, we surveyed some general users. For those participants, we described using scenarios in real-time, live shows, and dynamic videos in social network services. This feedback contained more varied opinions. The most common question was whether 5G will solve this problem. Of course, 5G offers more bandwidth to the user, then it will enable full high quality video transmission more smoothly, but the users suspect that the hypothesis would be achieved because new services will easily exhaust its bandwidth as mentioned before. So, we consider that our approach can be used as better countermeasure. Other comments were about psychological issues, such as a live sporting race, which can also be viewed on a TV using cable to obtain data. If the network is not running well, the users have another choice, or the video can be pre-downloaded in high quality so that they do not need to watch it in real-time; thus, only live video would unavoidably fall into that case. Even in the case in which clients want to save data, telecommunications providers also provide unlimited data options if

they do not truly care about the fee. Therefore, only the user who wants to watch live when many others are watching, or the network is being intensively used would require this system. Otherwise, this system will be more beneficial for mobile users or if it can provide a function that allows the user to select the size of the high-quality area. However, this system also has interesting uses in 3D space. For instance, in the scenario we mentioned before, one application would be on car windows and could determine where the driver is looking.

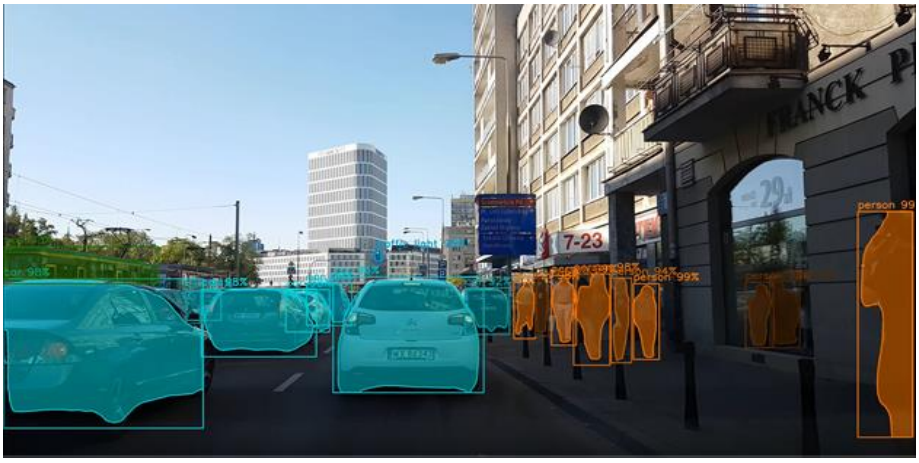


Figure 8. The preview of idea, if we equip it on car window (It was tested on driving record video)[11]

General users were surprised at this idea. It can filter the important information to the driver or passengers. This fresh idea earned more interest from the general user. These plans may become a future blueprint.

CHAPTER 6 LIMITATIONS AND FUTURE WORK

Section 6.1 RELATED TECHNOLOGIES

There is not only one way to serve the problem of media transmission. It can be solved in distinct policies such as enlarging transmission amount, enhancing compression efficiency or modifying media container protocol. They extract better information from media or let devices be powerful. Relatively, my project is focusing on the human perception that use software technologies to help servers provide more accurate data to user. In other word, this project does not optimize any algorithm of image data but enhance the usage of resource, so it may be more close to favorite recommend system.

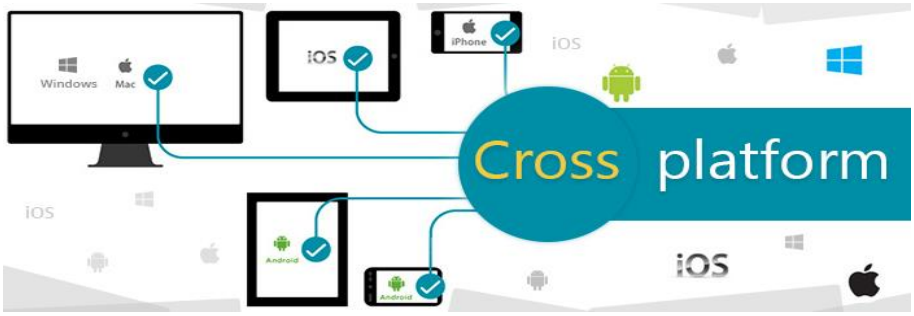
Section 6.2 LIMITATION OF DEVICE

In the process to parse raw material, it need very advanced GPU, otherwise it will cost several times time to finish it. However, if it falls into this case, it would not match the plan that we set up. We want to make it promote the user experience at real time video. Thus, they would not be call as live, if it has too long lag. We must to move most work to sever port, owing to equipping more powerful efficacy than general consumer electronics. It will easier to reduce the process time in this strategy.

Section 6.3 COUNTERMEASURES

Nowadays, browsers have already comprehensive developed there would be the perfect place to deal with cross-platform. Replanting this work on there will be the best way. What's more, media data are also stored at server port. In

this situation not only light the burden from client port, but also confront that decreasing transmission is an important direction of development.



Section 6.4 FUTURE WORK

In this section, I will talk about outlook. 3D virtual world still be the undeveloped environment. There a lot of products are in experimental stage. Therefore, these technologies could be published or not still be doubt. We also want to take part in this field. I consider 3D application will be next step of media. For instance, argument reality will let media information be more intuitive to users. And it also be the kind of image, which need several times data than 2D image. Thus, I want to push this technology to 3D field.

Figure 10. Anticipation working environment of future work [12]

CHAPTER 7 CONCLUSION

Our goal is going to propose a new architecture for streaming video that can play media more fluently in any situation. Based on this plan, we develop a prototype with several features to achieve it. This prototype equips functions to process the raw material which include parsing objects inside each frame, segmenting out those items and storing their information with crops. Then, we also make it can detect gaze positon to collect and track users' traits. Based on both elements, we have sources to do some approximate predictions for increasing users' perception. And, we build the deep learning database to practice it by Keras. In the experimental drill, we improve some small flaws to make it have better performance. Following from that, we will research how to execute this application with low efficacy consumption and transplant it into 3D environment. Next paragraph shows the achievement and comment of our idea.

After this prototype is finished and a more complete user study is completed, this system will have considerable potential to be utilized in different aspects. There are still many sub-features that are needed to make it complete. For example, the training of deep learning database still needs to be considered because this prototype is using the prepared data source. However, the training source needs to come from multiple usage habits for making the system practical. So, how to integrate the data from users will be another problem. We should perform some studies to better understand how human perception detects objects on the screen so that we can offer more effective applications that can also run on mobile devices. There is still a long way to go before mobile hardware can match the performance of the high-end computers. Therefore, determining

which part is the most helpful and transplanting this system will be a significant procedure for increasing the usage of this system. It is both a challenge and opportunity if we can simplify its operation such that it does not need to rely on advanced GPUs. It will be an innovation in the image processing field.

REFERENCES

[1] Jan Gehl, “Cities for People”, Island Press, First Edition, 2010

[2] D. Rudoy, D. B. Goldman, El.Shechtman and L. Zelnik-Manor, “Learning Video Saliency from Human Gaze Using Candidate Selection”, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1147-1154

[3] A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, “Does where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric”, IEEE International Conference on Image Processing, 12 November 2007

[4] M, Stewart, “The Actual Difference Between Statistics and Machine Learning”, Medium, 2018

[5] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, M. Ouaret, “The DISCOVER codec: Architecture, Techniques and Evaluation”, In Proceedings of the Picture Coding Symposium (PCS'07), Lisbon, November 2007

[6] W. B. Boyle, “Method and apparatus for storing a stream of video data on a storage medium”, US7657149B2, United States, 2000R.

[7] Oh-yun Kwon, Hye-Hyun Heo “Apparatus and method for 3d image conversion and a storage medium, US8977036B2, United States, 2011

[8] M. Bang, “Picture This – how pictures work”, In Perception and composition, Chronicle Books, 20

[9] https://www.youtube.com/watch?v=PQTjyKtsOuc

[10] Sumit Saha, “A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way”, December 16, 2018

[11] https://github.com/matterport/Mask_RCNN

[12] https://wayray.com/sdk/challenge