Promoting Fluency of Streaming Video by Learning Human Perceptive Traits to Reveal the Vital Section in Outstanding Quality

Chiang Shu Chiao
Department of Computer Science and Engineering
Waseda University
Tokyo, Japan
e-mail: csc19950207@dcl.cs.waseda.ac.jp

Tatsuo Nakajima
Department of Computer Science and Engineering
Waseda University
Tokyo, Japan
e-mail: tatsuo@dcl.cs.waseda.ac.jp

***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. It selects interested sub-image in pictures and only displays them with higher quality to achieve a lower transmission requirement. This optimization 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.**

Keywords -- Image processing; deep learning; accelerated streaming; media data structure.

I. INTRODUCTION

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. Even when communication equipment provides greater bandwith, the data requirement will also increase with the bandwidth due to more devices linking or a larger data consumption cost. As an analogy, building more roads is not the solution to traffic jams, and changing the usage habit of transportation is necessary [1].

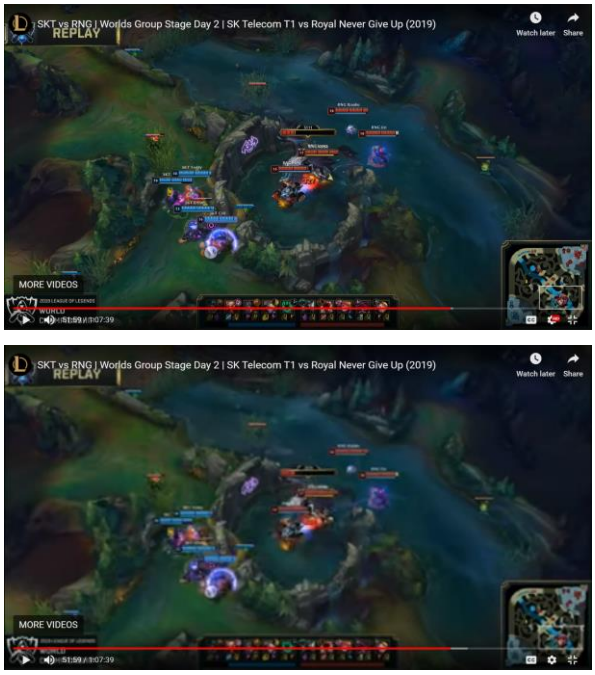


Figure 1. The top is a screenshot at 1080p, and the buttom is a screenshot at 240p. When users view these on large size monitors, they obviously note the differences between their sharpness [9]

If we can provide a more fluent model, it relieves pressure for route and improves 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.

 ISBN: 978-1-61208-772-6

Figure 2. Showing out the concept map of car window equips the AR device [10]

This paper presents a system that will show everything in best quality that the bandwidth allows. When the network transmission volume decreases, it will turn on a system to sort the hierarchy of subobjects, 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.

The remainder of the paper is structured as follows. Section 2 presents the information about researches that are related to the human's usage of multiple media and papers which study the human perceptive influence from media. In addition, we compare it with current model to find out the adventage of this idea. Section 3 expresses the method how we process this problem and proposes a new structure for this anticipation. This includes the work in the integral architecture and individual steps. Section 4 describes the implementation. In section 5, we list the main feedback collected from investigations. Last, we make a brief summary and indicate the feasible future work.

II. RELATED WORK

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.

Recent models almost use "adaptive bitrate streaming" technology to solve the problem of automatically adjusting video quality. Thus, image resolution also plays an important role here, 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 units 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.

III. METHOD

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.

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.

 ISBN: 978-1-61208-772-6

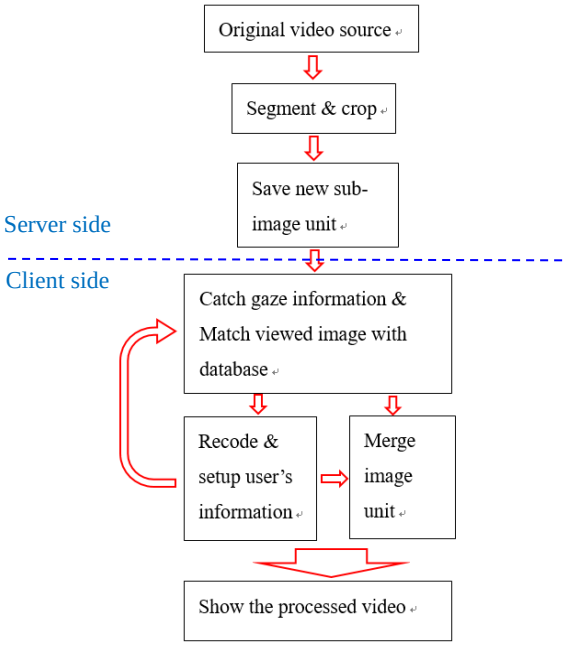


Figure 3. This is the basic system flow that demostrates how to process raw video and the design of data structure arrangement

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.

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 teach 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.

a) Finding the potential objects and segmenting them as the sub-image
b) Getting the gaze information
c) Showing the eye location in high resolution
d) Matching with database to build the user's interest pool

Figure 4. Descripting the detail of each section [11][12]

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.

E. *Storage strategy*

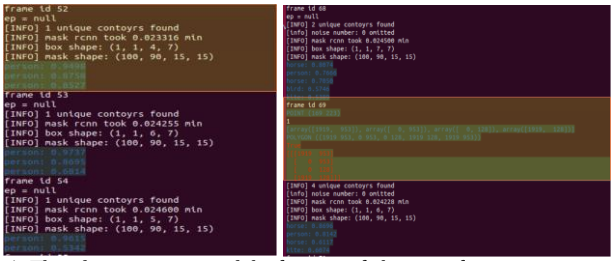
First, some basic concepts should be understood. We should learn how videos are saved in current streaming platforms. Currently, general frames are saved as complete pictures in different hierarchies of pixel levels. Frames are defined as the basic unit of a video. However, this system will require a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

 ISBN: 978-1-61208-772-6

IV. EXPERIMENT

This work was developed in the environment of Ubuntu 18.4, OpenCV3 and Python3. 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 (figure 5b). However, auto-expansion of the image set is necessary for future work. This prototype can also track and collect users' gaze information (Figure 5a), in order to make more effective predictions according to the users' habit.

We built the version for a user study. We set up the foremost stable database to represent the deep learning set and used the mouse to replace a gaze tracker because it has better accuracy and a direct read media source. Others have used the same structure and operation as we described above.



a) The advance process and the forecast of objects in frame
b) Merging the gaze information into procedure

Figure 5. The print out of surveying vision

V. PRELIMINARY USER STUDY

In this section, we present some simple questions to the participants in the user study. 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) How 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. 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

 ISBN: 978-1-61208-772-6

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. It has positive benefits to the driver or passengers,

In sum, general users were surprised at this idea. It can filter the important information for them. This fresh idea earned more interest from the general user. Therefore, these plans may become a future blueprint.

VI. 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] J. 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 and 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] O-Y. Kwon and H-H. 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=JAKQAAxNvvc, 2019/10/19

[10] https://www.blippar.com/blog, 2019/10/07

[11] https://www.youtube.com/watch?v=JAKQAAxNvvc, 2019/10/19

[12] https://leagueoflegends.fandom.com/wiki/Twisted_Fate, 2019/10/21

 ISBN: 978-1-61208-772-6