第 22 回日本バーチャルリアリティ学会大会論文集（ 2017 年 9 月）

Study on Telexistence LXXXVIII
Foveated Streaming: Optimizing video streaming for Telexistence systems using eye-gaze based foveation”

MHD Yamen Saraiji[1)]，Kouta Minamizawa[1)]，Susumu Tachi[2)]

1) 慶應義塾大学大学院メディアデザイン研究科
(〒 223-8521 神奈川県横浜市港北区日吉 4-1-1, {yamen,kouta}@kmd.keio.ac.jp)
2) 東京大学 高齢社会総合研究機構 (〒 113-8656 東京都文京区本郷 7-3-1, tachi@tachilab.org)

Abstract: Current Telexistence systems use full resolution image stream that is sent over the network to the user side. However, due to performance and network limitations, the resolution is significantly limited and thus can not provide sufficient details for the human eyes. This paper addresses the topic of increasing spatial visual acuity in telexistence related applications by taking the advantage of human eyes retina properties. Instead of streaming full resolution images over the network, hierarchical, multi-resolution image regions are sent to the user based on his eyes gaze position. Using this method, it’s possible to increase the spatial visual information perceived by the user, while significantly reducing network bandwidth and increasing processing performance of the sender/receiver sides. Here we present the overall system including the challenges in maintaining per-frame eye gaze synchronization over the network, and quantitative and qualitative studies verifying the effectiveness of foveated streaming are described.

Keywords: Telexistenc, Network Transmission, Perception

1. Introduction

Recent advances in virtual reality display systems began to offer end users immersive and high fidelity experiences in a rather accelerating manner. Current Head Mounted Displays (HMD) such as Oculus, Vive, and Samsung GearVR offer wide field of view (FoV) displays ($100^\circ \sim 120^\circ$) with a pixel density of $100 \sim 175$ per degree at refresh rate ranging from $60 \sim 90$ frame per second (FPS). In future, expected HMDs are to offer much higher resolution and refresh rate than now in order to enhance the visual experience and immersion towards human eye. However, such visual experiences would require high processing bandwidth to deliver such information stream to the HMD, and performance would decrease exponentially based on the target pixel density and linearly based on the target frame rate.

On one hand, virtual reality applications that run locally would only require a system capable to operate at such high bandwidth, which is mostly addressed by hardware manufacturers. On the other, Telexistence and Telepresence systems would suffer more at addressing this problem due to network bandwidth limitations. Even with the current standard compression codecs such as MJPEG, H264/MPEG-4, and VP8 will have limited bitrate to accommodate for network streaming. Also, in this type of systems, encoding/decoding are required before streaming and after receiving the image stream, which performance is also dependent on the pixel resolution of the images. Extending the problem for omnidirectional type cameras which sends the entire 360° view, the resultant pixel resolution of the equirectangular image (the stitched omnidirectional 2D image) would increase 5 to 8 times more than the limited field of view pinhole cameras image type (at $90 - 110^\circ$). As a consequence, the performance of encoding and streaming would increase the latency and decrease the overall experience for Teleoperation.

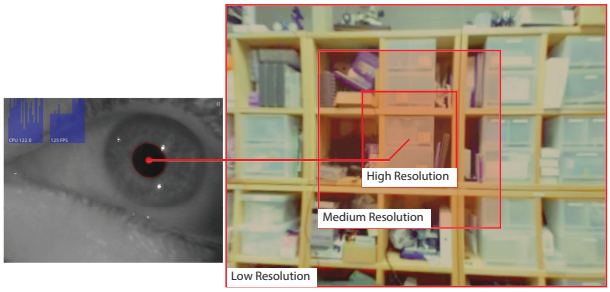

Figure.1: Foveated Streaming overview: multi-resolution regions based on eye gaze position.

To address the previous issues related to Telexistence and Telepresence systems, a perceptual driven approach was used. Human eyes have the highest visual acuity at the fovea in the retina, which is only about 2° of the visual field, and this visual acuity decreases the further it is from the center of the fovea towards the peripheral vision. By taking the advantages of this property, a multi-resolution image stream is generated from the original high-resolution images of the Teleoperated system. This multi-resolution stream maintains the high visual quality at the center of the eye gaze, while lower spatial resolution images are used for the peripheral area as shown in Figure 1. Using this method, the image stream would result in a higher compression ratio based on the selected fovea size, and the encoding/decoding performance would increase according to that too. This paper is based on the previous research area of foveated rendering, and it contributes as follows:

- Providing real-time network streaming synchronization for eye-gaze using RTP stream.
- Quantitative and qualitative studies showing the effect of foveated streaming for performance and human perception.

2. Related Work

The idea of using eye fovea as a driver to optimize the image spatial resolution has developed an important body of research in the area of multimedia and computer graphics optimization. In image compression, Wang et al. [1] proposed image coding system by taking into consideration the nonlinear decrease of spatial resolution in the human eye, and removing the high-frequency information from the peripheral area, thus improving the overall compression performance. Itti [2] used a different approach to address image compression by using saliency information rather than just the eye gaze position. And Ryoo et al. [5] proposed a video streaming service based on eye gaze foveation to enhance network bandwidth. In computer graphics, Guenter et al. [3] used foveated rendering in order to enhance rendering performance by 5-6 times than direct rendering. Patney et al. [4] used foveated rendering in virtual reality applications to reduce the rendering cost, the work showed that foveated rendering did not introduce a significant difference in visual quality compared to direct rendering.

3. System Design & Implementation

3.1 Encoding & Streaming

Foveated streaming pipeline is shown in Figure 2. This pipeline runs on the robot side, and the eye-gaze data is

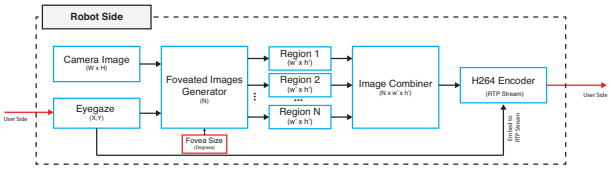

Figure.2: Streaming pipeline.

received from the user side. The pipeline generates a number of regions (N) which can be set in using system parameters. Region size in degrees is also determined by the Fovea Size parameter, in the conducted experiments, a value of 15 deg produced a good balance between streaming compression and perceptual awareness. The generated regions are combined into a single image frame as in Figure 3 (Streamed Regions). Then the images are compressed using the H264 encoder and converted to RTP stream to be sent to the user side. One important thing is to synchronize the eye gaze information in order to reproduce the same arrangements when combining the images on the user side. To do that, eye gaze information used for foveation is embedded into the first RTP packet of the image frame before being sent to the user. The main advantage of embedding this information into the RTP stream is to ensure exact synchronization of the data and the image stream when rebuilding the original image at the user side.

3.2 Unpacking & Presentation

User side receives the packed regions as an image stream, each frame contains foveation levels as was described in the previous section. When decoding the stream, first the eye gaze position is extracted from the first RTP packet of the packed frame and is used to arrange the regions and positioning them exactly as they were sent. After unpacking the images, the regions are presented starting from the last region. Region number N is rendered into an offscreen image covering the entire field of view, next Region number N-1 is rendered at the original scale, and so on until Region 1 is rendered on top of them as shown in Figure 3 (Foveated Image). To mitigate the effect of change in resolution between two consecutive regions, a mask is applied to each region which helps to fade the edges and to reproduce similar effect of eye retina decrease of visual acuity, while maintaining high resolution at the center of the region as shown in Figure 3 (Region Masking).

4. Evaluation

To evaluate foveated streaming method, two types of evaluation were conducted:

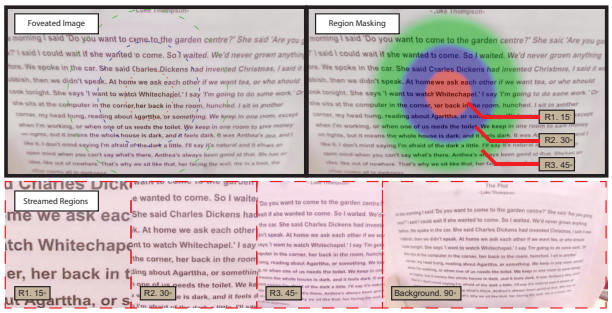

Figure.3: Foveation streaming decomposition (1) Foveated Image: final reproduced image at the user side, (2) Region Masking: masks used for combining the regions, and (3) Streamed Regions: streamed frame containing regions of foveation.

- Performance Evaluation: to evaluate the overall performance of the system compared with full resolution image streaming, the bandwidth usage, and CPU consumption are tested.
- Qualitative Evaluation: a user study confirming the effectiveness of foveated streaming in maintaining the perceptual consistency of visual information.

Same system and hardware setup were used in both evaluations. USB3.0 camera module (Model no. See3CAM CU130 by e-consystems) were used to capture image stream, and connected to Intel NUC for video encoding and streaming over an ethernet LAN to the user side. For media encoding and streaming, GStreamer library was used. The video stream is compressed using H.264 video codec provided by the library.

4.1 Performance Evaluation

The goal of performance evaluation is to measure the effectiveness of using foveated streaming network bandwidth, and CPU performance at the encoder side compared with full resolution streaming. In this evaluation, three different streaming modes were used: 640x480 pixels at 60 FPS, 1280x720 pixels at 60 FPS, and 1920x1080 pixels at 30 FPS. To maintain image quality among the images in both cases, adaptive bitrate quantizer set at 20% was used for the video codec. Foveated streaming parameters used were: 15° Fovea Size, and 3 levels of foveation. Figure 4 shows the quantitative results the proposed system. The results show the performance effectiveness of foveated streaming compared to full resolution streaming for both the bandwidth (compression rate varies from 5 to 8 depending on the streaming resolution) and CPU performance (about half CPU usage for foveated streaming).

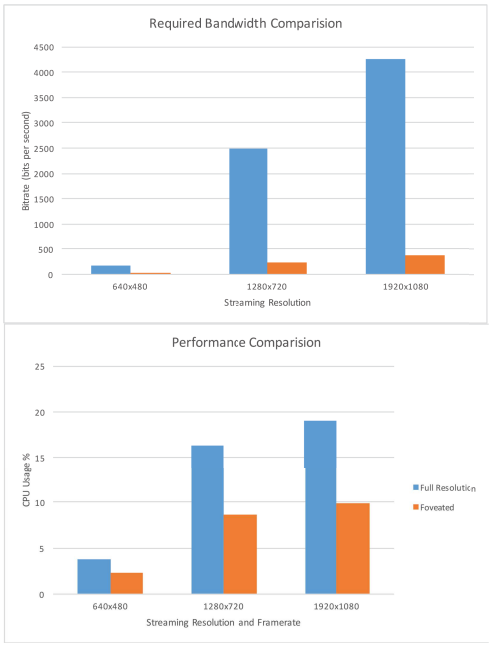

Figure.4: Performance evaluation of foveated streaming compared with full resolution streaming.

4.2 Qualitative Evaluation

A user study was conducted to evaluate the visual perceptual effectiveness when using foveated streaming compared to full resolution streaming. The hypothesis is when using foveated streaming, there is no significant difference in reading speed compared when using full resolution streaming.

4.2.1 Study Setup

The study setup consisted of a three-axis telexistence robot head (providing pitch, yaw, roll motion) equipped with 90° cameras for video streaming. The cameras were set to capture 1920x1080 pixels at 30 FPS when streaming. For foveated streaming settings, fovea size used was 15° with 3 levels of foveation resulting a stream of 1528x320 pixels per image. An A3 paper containing one page of text (font used bold Arial at a font size of 24points, and 1.5 line spacing) placed in front of the robot at a distance of 30cm. Page was fixed that the first line of text is placed at the eye level of the robot. The total number of pages is 5, with an average of 225 words per page. The text used was a short English essay (The Plot by Luke Thompson). The user is connected to the robot over the network and uses an HMD to control the head motion of the robot and to perceive the visual information from it. The HMD is equipped with an eye gaze tracker (Pupil-lab for Oculus DK2) to measure eye movement when foveated streaming mode is enabled. In this setup, the user will be required to use both his head mo-

tion and eye motion to read the text from the beginning to the end of the page along yaw and pitch axises.

4.2.2 Procedure

Prior to the study, participants are asked to fill a questionnaire containing basic information such as the age, eyes visual acuity for the left and right eye, and frequency of using HMDs in general. The participants are only informed of the task of reading and essay at one page a time and were asked to inform the experimenter once they finished reading the active page. After wearing the HMD, the participants are asked to calibrate their eye gaze using 9-points eye gaze calibration procedure, then a waiting gray screen is shown before starting reading. Once they inform their readiness to start reading, the user gets connected to the robot, and they start reading the text from top to bottom, left to right. After reading of the page is finished, the waiting screen is shown again and the procedure continues for the next page until the 5 pages are over. The streaming mode for the pages is switched between full streaming/foveated streaming when changing the pages, and the first page's mode is randomly selected between both modes. Reading time is measured per page and stored along with the streaming mode for performance evaluation. After the study is over, qualitative questions are asked about the reading experience and whether anything has been noticed while reading and switching the pages.

4.2.3 Results and Discussion

In this evaluation, 7 participants joined the study (6 males and 1 female with an average age of 24 ± 3) from different ethnicities. Overall feedback did not include any significant difference in the reading experience of the pages. Two participants reported seeing blurry words around the edges of gaze point (participant 1 and 4), this is due to the eye gaze tracker which could have been moved during the experiments (HMD was moved from the calibration position), the performance results show the impact of the calibration mismatch to their reading speed in foveated streaming. For performance results of the study, Figure 5 shows the average time per word for both full resolution and foveated streaming per participant. Overall reading speed can of both modes shows no significant difference between both modes. Reading speed varies per page, and overall the speed slightly increases when proceeding in the study, this can be due to the adaptation time of reading over a Telexistence robot. From these results, it can be shown that foveated streaming provided similar results to the full streaming under the condition of consistent eye gaze tracking.

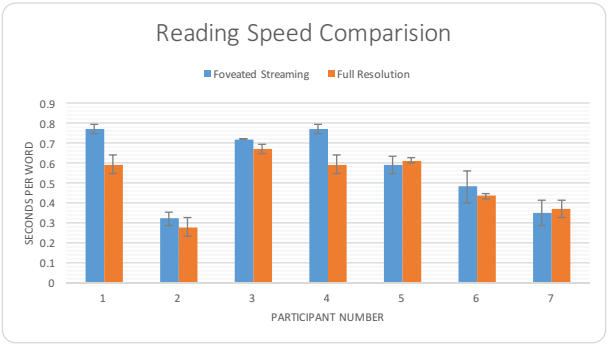

Figure.5: Subjective evaluation of reading speed for foveated streaming and full resolution streaming.

5. Conclusion

This paper shows a foveated streaming system for Telexistence applications to reduce the network bandwidth and processing requirements for encoding and decoding video images. Performance evaluation shows a significant decrease in bandwidth size, 5-8 times less than full resolution streaming, and twice the performance increase. A user study showed no significant difference when using foveated streaming compared to full resolution streaming.

Acknowledgment

This work is supported by JST ACCEL Embodied Media Project Grant Number JPMJAC1404.

References

[1] Wang, Zhou, and Alan C. Bovik. "Embedded foveation image coding." IEEE Transactions on image processing 10, no. 10 (2001): 1397-1410.

[2] Itti, Laurent. "Automatic foveation for video compression using a neurobiological model of visual attention." IEEE Transactions on Image Processing 13, no. 10 (2004): 1304-1318.

[3] Guenter, Brian, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. "Foveated 3D graphics." ACM Transactions on Graphics (TOG) 31, no. 6 (2012): 164.

[4] Patney, Anjul, Joohwan Kim, Marco Salvi, Anton Kaplanyan, Chris Wyman, Nir Benty, Aaron Lefohn, and David Luebke. "Perceptually-based foveated virtual reality." In ACM SIGGRAPH 2016 Emerging Technologies, p. 17. ACM, 2016.

[5] Ryoo, Jihoon, Kiwon Yun, Dimitris Samaras, Samir R. Das, and Gregory Zelinsky. "Design and evaluation of a foveated video streaming service for commodity client devices." In Proceedings of the 7th International Conference on Multimedia Systems, p. 6. ACM, 2016.