Immersivemote: Combining Foveated AI and Streaming for Immersive Remote Operations

Pietro Lungaro
Department of Commucation System
KTH Royal Institute of Technology
pietro@kth.se

Konrad Tollmar
Department of Commucation System
KTH Royal Institute of Technology
konrad@kth.se

ABSTRACT

Immersivemote is a novel technology combining our former foveated streaming solution with our novel foveated AI concept. While we have previously shown that foveated streaming can achieve 90% bandwidth savings, as compared to existing streaming solutions, foveated AI is designed to enable real-time video augmentations that are controlled through eye-gaze. The combined solution is therefore capable of effectively interfacing remote operators with mission critical information obtained, in real time, from task-aware machine understanding of the scene and IoT data.

CCS CONCEPTS

• **Computing methodologies → Mixed / augmented reality**;

KEYWORDS

Eye-tracking, Foveated Streaming, Remote Operations

ACM Reference Format:
Pietro Lungaro and Konrad Tollmar. 2019. Immersivemote: Combining Foveated AI and Streaming for Immersive Remote Operations. In *Proceedings of SIGGRAPH '19 Talks.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3306307.3329895

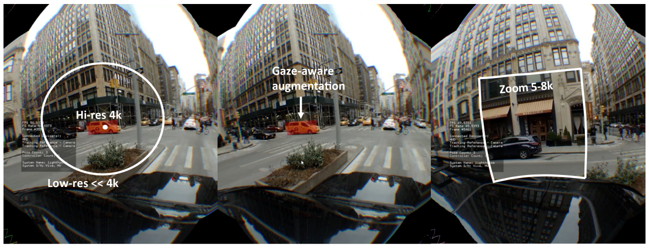

Figure 1: From left to right, a) foveated composition and b) foveated augmentation, and c) new display modalities for increased immersivity, e.g. 5-8K Zooming

1 INTRODUCTION

Autonomous machines and vehicles are highly celebrated as potential drivers of a new industrial revolution, bringing unprecedented benefits for the society at large and improved production capabilities and safety in industrial settings. In many application domains, however, the achievement of complete autonomy is currently challenged by a number of unsolved problems and even if the growth in processing power and the increasing scalability and maturity of AI algorithms bring a positive outlook, the time horizon for such an achievement is most likely on the order of a decade (or more). This is for example the case of level 5 autonomy in self-driving vehicles [Casner et al. 2016]. Thus, in many practical settings, the actual control in these upcoming systems will be at a level somewhere in between "fully manual" and "fully automated", leaving a clear and central role to the human operators [Pretlove and Skourup 2007].

Understanding the spectrum of future roles for the human operators and enabling them with novel technologies to accomplish their new tasks is therefore crucial to fulfill this vision of "automated society". In fact, the human-in-the-loop represents the most vulnerable system element and the one most easily overlooked [Pretlove and Skourup 2007].

Even if the concept of telepresence has been around for quite some time [Minksy 1980], with different embodiments and areas of potential applications proposed and explored over the years, it has clearly not reached its anticipated potential. The motivation is simply found in the steep technological constraints to achieve a satisfactory user experience in the targeted use cases. On the other hand, we are now facing for the first time intelligent connectivity [GSMA 2018], in which 5G networks with high bandwidth and low latency, AI advancements and affordable processing power have all reached a level of maturity needed for the emergency of telepresence services as feasible and potentially disruptive technologies.

2 HUMAN-AND TASK-AWARE INTERFACES

To maximize human and task awareness in the decision making of remotely controlled/monitored autonomous systems, we have developed an innovative end-to-end approach, featuring both an advanced content delivery mechanism and a novel user interface designed to effectively interface humans to different services leveraging the intelligent connectivity paradigm.

By exploiting the information from connected eye-trackers, we have shown that foveated streaming [Lungaro et al. 2018] can lead to up to 90% bandwidth savings, compared to state-of-the-art solutions. This approach effectively extends foveated rendering [Guenter et al. 2012] to become an end-to-end content provision paradigm, achieving in the process also similar computational savings. While our initial efforts focused on "on-demand" pre-encoded video types, in this work we instead showcase, for the first time, an embodiment adapting foveated streaming to supporting "live" video in real-time.

SIGGRAPH '19 Talks, July 28 - August 01, 2019, Los Angeles, CA, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6317-4/19/07.
https://doi.org/10.1145/3306307.3329895

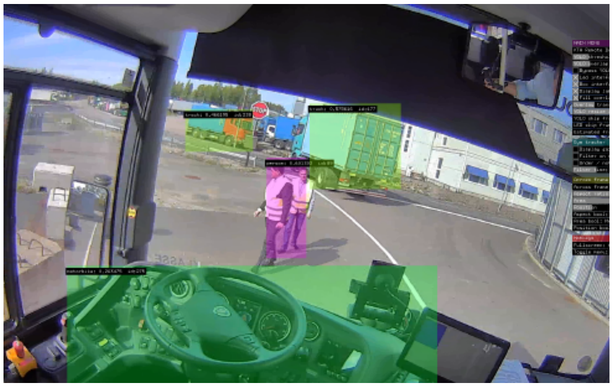

Figure 2: Example of task-aware augmentations inserted for supporting a remote driving task.

Extremely high levels of bandwidth optimization have been confirmed also in this case. Such savings are of paramount importance for enabling remote operations, as they can be "transformed" into increased safety and scalability:

- being able to support much higher video resolution leads to increased immersivity, safety and control precision,
- these savings allow supporting many more vehicles or machines for a given mobile infrastructure configuration
- they provides a more flexible approach, as available networking solutions and performances may vary substantially in different locations, e.g. not all featuring 5G.

An example of Foveated AI is illustrated in figure: 2, where the live video feed from a Scania remotely controlled bus is delivered to an office several kilometers away [Ericsson 2017]. Since the performed task is remote driving, we have implemented a solution utilizing eye-gaze to understand whether the remote driver has not looked at specific obstacles along the current vehicle’s motion path. In that case these are augmented with different colors, depending on their class (e.g. vehicles, pedestrians). Another available embodiment also features the possibility of triggering a "zoom" in the content. This provides higher-resolution in a portion of the frame, centered around operators gaze, or based on their viewing behavior, e.g. automatically triggered when looking at an object for more than a time threshold. Currently we are providing 5K-8K resolution in zoomed area (figure: 1c), and this is controlled in real time by the remote users.

3 TECHNOLOGY DEMONSTRATOR

Our current testbed embodiment features a remote operator side equipped with a Tobii HTC Vive HMD with in-built eye-trackers and powered by VR gaming PC. The foveated streaming and task-aware augmentation is done with a similar machine, acting as the unit on-board of a remotely monitored vehicle (server). Before transmission, each frame is processed and composed, according to our foveated streaming workflow [Lungaro et al. 2018], into a foreground region at high resolution and a low quality background (figure: 1b). This is done utilizing the latest eye-gaze estimates available at the server side. Currently the foreground resolution is limited to 4K by the headset display.

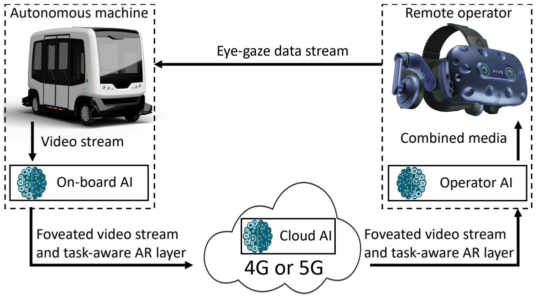

Figure 3: Service architecture for live foveated streaming with Foveated AI inserted at vehicle, cloud or operator sides.

Both the size of the foreground and the background resolution can be tailored to specific individual preferences. These depends on visual characteristics of the users’ eyes: in specific the size and shape of their fovea and their peripheral color perception. In parallel to this, the same frame is also processed to extract all relevant metadata regarding objects present in the frame and their categories. In the current proof-of-concept embodiment we are utilizing YOLOv3 for object detection [Redmon and Farhadi 2018], but in the near future we plan to access and process more complex information, e.g. that from the actual self-driving units on board of autonomous busses or the one obtained by industrial IoT devices. While object detection has been currently implemented only at the remote driver side, we envision a complete architecture where machine intelligence metadata can be added to the video stream by different entities. The possibility for different entities to simultaneously cooperate in enriching the video stream with metadata obtained by machine intelligence processes or sensor data is illustrated in (figure: 1a). There AI agents can be located in the autonomous vehicles ("on-board AI"), intermediate cloud nodes ("cloud AI") and/or software agents at the human-user side ("operator AI").

REFERENCES

Stephen M. Casner, Edwin L. Hutchins, and Don Norman. 2016. The Challenges of Partially Automated Driving. *Commun. ACM* 59, 5 (April 2016), 70–77. https://doi.org/10.1145/2830565

Ericsson. 2017. Remote monitoring and control of vehicles – Ericsson. https://www.ericsson.com/en/mobility-report/remote-monitoring-and-control-of-vehicles

GSMA. 2018. Intelligent Connectivity: the Fusion of 5G, AI and IoT. https://www.gsma.com/iot/news/intelligent-connectivity-5g-ai-iot/

Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D Graphics. *ACM Trans. Graph.* 31, 6 (Nov. 2012), 164:1–164:10. https://doi.org/10.1145/2366145.2366183

Pietro Lungaro, Rickard Sjoberg, Alfredo Jose Fanghella Valero, Ashutosh Mittal, and Konrad Tollmar. 2018. Gaze-Aware Streaming Solutions for the Next Generation of Mobile VR Experiences. *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (April 2018), 1535–1544. https://doi.org/10.1109/TVCG.2018.2794119

Marvin Minksy. 1980. Telepresence. https://web.media.mit.edu/~minsky/papers/Telepresence.html

John Pretlove and Charlotte Skourup. 2007. Human in the loop. *ABB Review* 1 (2007), 6–10.

Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *arXiv* (2018).