Mallesham Dasari - Research Statement

Multimedia has played a significant role in driving Internet usage and has led to a range of technological advancements such as content delivery networks, compression algorithms, and streaming protocols. With the rise of augmented and virtual reality (AR/VR), multimedia is undergoing a fundamental shift in sharing experiences online, and continues to drive the future of the Internet. The shift towards these *immersive experiences* (3D video and beyond) has put a strain on current generation computing and networking infrastructure, with stringent requirements in terms of low latency, high bandwidth links and cost effective resource provisioning solutions.

My research vision is to enable future immersive experiences by solving core problems in multimedia systems. Specifically, I focus on three important problems in multimedia content delivery: 1) The sender-side content delivery pipeline to capture 3D scenes in real-time is compute and GPU memory intensive, resulting in scalability issues; 2) Unlike traditional videos, immersive 3D content requires orders of magnitude more bandwidth for the same perceived quality which requires efficient field-of-view guided streaming solutions; 3) Most of today's content delivery mechanisms are fundamentally limited by the use of traditional monolithic compression methods that are not optimal for adapting to variability in network capacity.

Addressing these challenges, I have built three multimedia systems, namely Mosaic [1], Parsec [2], and Swift [3]. My key insight in solving these problems is designing context-aware multimedia data representations. For example, compression algorithms can be significantly content-aware. DNN compressors might be good for natural scenes, but these same approaches make much less sense for synthetic 3D rendered content. This work naturally falls at the intersection of several areas in CS & ECE including Computer Systems, Networks, Computer Vision, Graphics, and HCI. I believe this inter-disciplinary approach is necessary to sustain the constant growth of the multimedia ecosystem on the Internet.

Contributions. I published my work in several premier conferences in the areas of Systems and Networking (NSDI & SIGCOMM), Internet Measurements (IMC & PAM), Multimedia and HCI (ISM & IMWUT/UbiComp). I have built several open source multimedia tools such as `sr360` [4] and `vqa-deep` [5] that are widely used by the multimedia community [6]. I developed vqa-deep [5] during my internship at HP Labs, which has been tested out on Aruba (an HP enterprise company) WiFi access points for practical deployment. My research was recognized for best paper award at IEEE Symposium on Multimedia, best demo award at the CONIX annual review (an SRC center), a best presentation award at ACM MobiCom wireless students workshop, and my PhD thesis received a departmental nomination for ACM SIGMM outstanding thesis award.

Immersive Multimedia Content Delivery

Thus far, my research has focused on discovering efficient data representations for multimedia content delivery, addressing the above fundamental challenges— high computational complexity, limited bandwidth and its fluctuations. This led to three notable contributions described below.

MOSAIC [1] tackles the first problem: significantly decreasing the computational requirement of 3D scene capture. 3D content delivery enables immersive telepresence applications that allow remote viewers to watch from any location or angle within the 3D scene (often refered to as 6DOF video). A key finding of this study is that existing 3D content delivery solutions adopt inefficient native source representations such as point clouds and RGB-D and therefore suffer from high bandwidth requirements. Instead, I propose using an intermediate representation based on texture meshes that is more compact and requires significantly less bandwidth for the same quality. Creating textured mesh representations from raw camera sources poses nontrivial compute and memory challenges and often leads to high latency that is not tolerable for real-time interactive applications. Mosaic tackles this challenges by using a split-merge-based distributed architecture, to partition the scene capture pipeline into many compute nodes with each node reconstructing a per-camera scene and merging them into one scene on a central server. Mosaic achieves significantly lower latency and requires orders of magnitude less bandwidth for the same quality compared to state-of-the-art Microsoft Holoportation 3D streaming system.

PARSEC [2] addresses the second problem: reducing the high bandwidth requirement of panoramic 360° videos. Parsec introduces a viewport adaptive streaming system which downloads only portions of a 360° scene that a viewer perceives. Viewport adaptation significantly reduces the bandwidth requirement by culling content outside of the field of view. However, it is difficult to predict user head movement with a high accuracy and at a low latency. Consequently, viewport adaptation is not used in current streaming platforms that simply transmit the complete 360° scene on each frame. Instead, Parsec introduces an easy path to adopt viewport adaptation by proposing neural data representations for compressing the content. Parsec streams the predicted viewport in high quality and the rest of the non-viewport region in a significantly downscaled and compressed version. In case of an incorrect prediction, Parsec uses a deep learning based super-resolution model to upscale the mispredicted low quality regions to the original resolution. Parsec introduced a concept called 'micro-models' to reduce neural network overhead in terms of inference latency and model download overhead.

SWIFT [3] focuses on the third problem of adapting streaming to bandwidth fluctuations. While my earlier systems Mosaic and Parsec propose compact data representations to minimize the bandwidth requirements of high quality content, Swift designs new more agile data representations. Rate adaptation is a fundamental problem in Internet content delivery to compensate for the uncertainty in network capacity. However, despite the years of effort in new adaptation algorithms, existing content delivery is shown to be sub-optimal because of monolithic encoding. Swift takes an alternative approach called layered coding that optimally solves the bandwidth variability problem. However, devising an algorithm for layered data representations is nontrivial due to cross-layer compression overheads and high coding latencies. Instead, Swift takes a clean-slate approach using learning based codecs to realize layered coding. Swift has three components: i) a layered encoder that learns to encode a video frame into layered codes by purely encoding residuals from previous layers without introducing any cross-layer compression overheads, ii) a decoder that can fuse together a subset of these codes (based on availability) and decode them all in one go, and, iii) an adaptive bit rate (ABR) protocol that synergistically adapts video quality based on available network and client-side compute capacity. Swift's layered neural codecs enable software defined coding, agile codec upgrades, and fine-grained bitrate adaptation capabilities.

Wireless Support for Immersive Systems

In addition to multimedia content delivery, I worked on several wireless systems for multimedia described below.

Wireless and Vision Fusion. Localization and tracking is a classical problem for AR/VR applications. Today, most of the industry solutions adopt visual algorithms (e.g., SLAM) for tracking. However, visual tracking is highly sensitive to environmental conditions, limiting these applications to specialized environments. Instead, I have built a system called RoVAR, which augments visual tracking with wireless positioning (e.g., UWB) and fuse them together using algorithmic and machine learning based techniques to achieve robustness, high accuracy, and scalability for multi-user scenarios. I believe wireless positioning will be ubiquitous in the near future and mixed reality will greatly benefit from RoVAR style fusion of multi-modal tracking solutions.

High Frequency Wireless Links. In the near-term, ultra-thin AR/VR glasses will likely require remote rendering and streaming to achieve high fidelity graphics. For highly latency sensitive applications, video encoding/decoding could be too costly leading to multi-gigabit bandwidth requirements. I worked on a system called Cyclops [7] that uses free-space optical (FSO) technology to enable extremely high throughput and low-latency links. Cyclops is designed with an FSO transmitter placed on a nearby rendering computer (or ceiling mounted access point), streaming high quality VR content to an FSO receiver placed on a VR headset. The key challenge here is to maintain the narrow wireless link to tolerate users' movement. Cyclops develops accurate tracking and pointing methods to steer the FSO beam as the user moves. Cyclops style of solutions have significant potential for other applications such as reconfigurable datacenters and high speed inter-satellite communication.

Future Research

The growth of immersive technologies is now ushering in a new wave of shared experiences and applications. At a high-level, my future work will bridge the gap between emerging multimedia technologies and the underlying systems required to support them. In this section, I discuss two specific threads of research.

For Better or Metaverse. There has been a surge of interest in platforms that tightly couple digital content with the physical world (e.g. Digital Twins, Metaverse, AR Cloud, Spatial Web, etc). The success of these platforms hinges not only on strong technical components, but also well executed social and policy strategies. On the technical front, I plan to work on three critical components: First, I would like to explore end-to-end solutions for interactive real-time 3D streaming in the context of remote assist applications (e.g. industrial factory floors, medical, defense). As part of this, I am interested in how to scale 3D capture and streaming systems to wide-areas where you need to fuse many camera sources and multiple users. Second, I want to look at how network and nearby compute can better support resource constrained headset platforms where size and power are currently deal-breakers for mass adoption. This can be addressed with energy-efficient decoding, rendering and communication pipelines that partition between headsets, body-worn computers (like phones), and nearby edge computers. Third, we are seeing a fascinating shift in graphics pipelines with technologies like light field cameras/displays and learning-based model representation and rendering (e.g. Neural Radiance Fields, etc). As these approaches mature, there will be a plethora of new systems challenges around how to efficiently represent, store, stream, and process these new media. I am well positioned to bridge emerging concepts from the graphics and vision community into networked systems. Though not my research focus, I am also interested in tracking activities related to Metaverse policy and regulation within the broader community. In the same way that academia's involvement in designing the Internet was critical in shaping its evolution, the Metaverse will also benefit from incubating in a research environment. To this end, I hope to aid in the democratization of many of the key driving XR technologies needed to enable an open and sustainable vision of the Metaverse.

Convergence of Sensing and Networking. The past decade has witnessed significant advances in sensing and communication technologies to optimize localization, multi-user tracking, and adaptive streaming solutions. Moving forward, I plan to explore how we can leverage sensing and networking codesign to improve performance in more tightly integrated platforms. Intelligent sensing can help scope what information needs to be transmitted and when. My Parsec work was a simple example of how head tracking can reduce streaming overhead. This concept can be generalized across a wide body of perception tasks in XR and even robotic systems. Looking at the problem from the other direction, communication and edge computing can be used to offload complex sensing tasks. This often boils down to a problem of how to partition sensing and processing pipelines in networked systems. One early example I would like to explore is how offloading could help multi-user, multi-sensor, spatial mapping platforms. Most current multi-user SLAM systems perform monolithic updates on a single shared model of the environment, but in the future this needs to be a more continuous and dynamic process.

References

[1] Mallesham Dasari, Jin Tao, Smith Connor, Apicharttrisorn Patrick, Rowe Anthony, and Seshan Srinivasan. Meshreduce: A distributed scene capture architecture for 3d telepresence. In *Submission, ACM SIGCOMM*, 2023.

[2] Mallesham Dasari, Arani Bhattacharya, Santiago Vargas, Pranjal Sahu, Aruna Balasubramanian, and Samir R Das. Streaming 360-degree videos using super-resolution. In *IEEE INFOCOM 2020*, pages 1977–1986. IEEE, 2020.

[3] Mallesham Dasari, Kumara Kahatapitiya, Samir R Das, Aruna Balasubramanian, and Dimitris Samaras. Swift: Adaptive video streaming with layered neural codecs. In *USENIX NSDI 2022*, pages 103–118, 2022.

[4] Super Resolution for 360° Videos. `https://github.com/VideoForage/SR360`. Accessed: 2022-11-30.

[5] DeepVQA. `https://github.com/VideoForage/VQA-Deep-Learning`. Accessed: 2022-11-30.

[6] Multimedia Libraries. `https://github.com/VideoForage`. Accessed: 2022-11-30.

[7] Himanshu Gupta, Max Curran, Jon Longtin, Torin Rockwell, Kai Zheng, and Mallesham Dasari. Cyclops: an fso-based wireless link for vr headsets. In *ACM SIGCOMM 2022*, pages 601–614, 2022.