EIC MESSAGE

Multimedia Research Toward the Metaverse

Shu-Ching Chen, *Florida International University, Miami, FL, 33199, USA*

Metaverse, a set of fully immersive digital worlds where people can interact with one another using three-dimensional (3-D) avatars, is poised to become an increasingly popular modality to communicate with others. Many in the space consider it an evolution of the Internet, with augmented and virtual reality headsets providing the hardware to both visualize and interact with it. However, creating custom 3-D content and environments for a digital world can be extremely time-consuming and costly compared to typical web pages. To this end, research in procedural content generation is required to empower the efficient creation of high-quality multimedia contents for the metaverse. One other core concern when creating a fully decentralized metaverse is the need to download relevant environments on demand as people traverse across different worlds. Further research by the multimedia community on efficient multimedia networking, storage, and compression is necessary to ensure that digital worlds remain easily accessible to others.

Metaverse aims to build an immersive virtual world that allows the users to interact with the digital environment and other users in real time. Such concept has recently attracted much attention from both the academic and industry, but there remain many fundamental problems to be resolved to realize the metaverse. Among them, multimedia also contributes to an important portion and faces many emerging but challenging technical problems.[1] Specifically, the primary hurdles toward the creation of a full-fledged metaverse include efficient creation and generation of three-dimensional (3-D) immersive contents, efficient multimedia content transmission, and so on.

In terms of the 3-D multimedia content generation, works on architecting immersive 3-D experiences, known as the problem of procedural content generation (PCG), involve the integration of large amounts of multimedia contents, including 3-D models, 2-D images, and 360 video.[2] Without in-depth knowledge of the tools and processes necessary to create such content, many creators in the metaverse will be extremely limited in what they can create. Traditionally, 3-D PCG has involved the manual creation of small building-block 3-D assets that can be reconfigured together to create new objects.[3,4] These methods do not allow for fine-grained control over the environment, and still require the creation of 3-D assets.

More recently, AI systems have proven to be a powerful tool for 3-D PCG but are still in their infancy. One of the seminal works in AI toward reducing the need for such equipment is the introduction of NeRFs.[5] NeRFs allow for novel-view synthesis based on a fixed set of input images. Many variants have been developed to allow for fast training,[6] compression for low-latency viewing,[7] and not requiring camera parameters.[8] However, further research is needed to understand the viability of neural rendering specifically for 3-D immersive environments, where users may look at assets from extreme angles. Another recent approach of PCG is based on cross-modal generative models using the natural language, which utilizes the user-generated language description of the content to directly create metaverse environments. Meta has recently announced BuilderBot, using natural language to help users generate 3-D environments using voice prompts. Chen *et al.*[9] and Howard-Jenkins *et al.*[10] have provided ways to leverage natural language to generate 3-D rooms, which can help create 3-D environments with assets already laid out. Large-scale multimodal models such as OpenAI's DALL-E aim to generate 2-D images from arbitrary natural language prompts.[11]

1070-986X © 2022 IEEE
Digital Object Identifier 10.1109/MMUL.2022.3156185
Date of current version 4 May 2022.

While many of the aforementioned works rely on only one type of input for their PCG algorithms, it may be of interest to the multimedia community to build systems that may integrate different input modalities together. Such work is similar in concept to multimodal deep generative models.[12] Using different kinds of input could allow users to have additional flexibility and control over the content they wish to create, and lead to more robust systems for PCG in the metaverse.

Another fundamental problem toward the metaverse is to efficiently transmit multimedia contents to users' devices. Due to the extremely large volume of multimedia contents contained and displayed in the metaverse, a hybrid communication strategy based on both local storage and data streaming is explored.[13] However, such methods would introduce intensive amounts of multimedia data to be transmitted. To improve the efficiency of 3-D multimedia content transmission, various strategies have been developed. Multimedia coding and compression methods have been utilized to encode the data with less disk storage.[14] Especially with the help of deep neural network, large-volume multimedia contents in the metaverse can be compressed into low-dimensional features with encoders, and then reconstructed by another decoder network in a local device without significant loss of content quality.[15] Meanwhile, since only partial data in metaverse are needed for the users due to limitations in their perception without the reduction in the quality of experience, methods have been investigated to control the quality of presented contents[16] and transmit only the data expected to be displayed based on the prediction of saliency, viewport, and user behaviors.[17] While many multimedia techniques have been developed to facilitate network communication for the applications of the metaverse, the latency of the network transmission remains a challenging problem. Meanwhile, the data streaming, network-level optimization on the 5G networks, and other further research on multimedia coding and subjective quality optimization remain important to enable broader and more efficient development of the metaverse.

Metaverse integrates a collection of multimedia contents and requires various techniques to enable efficient and high-quality creation, generation, transmission, and visualization of the contents. In addition, various domains such as network communication, artificial intelligence, virtual reality hardware, etc., need to work together to provide the comprehensive solution. While major advancements have been made in the past decade, continuous research is still required to bring the immersive digital world to a broader range of end users and enable interactive and immersive experiences for them.

REFERENCES

1. S. -M. Park and Y. -G. Kim, "A metaverse: Taxonomy, components, applications, and open challenges," *IEEE Access*, vol. 10, pp. 4209–4251, 2022, doi: 10.1109/ACCESS.2021.3140175.
2. E. Coltey, Y. Tao, T. Wang, S. Vassigh, S. -C. Chen, and M. -L. Shyu, "Generalized structure for adaptable immersive learning environments," in *Proc. IEEE 22nd Int. Conf. Inf. Reuse Integr. Data Sci.*, 2021, pp. 294–301, doi: 10.1109/IRI51335.2021.00047.
3. W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," in *Proc. 14th Annu. Conf. Comput. Graph. Interactive Techn.*, 1987, pp. 163–169, doi: 10.1145/37402.37422.
4. H. Kim, S. Lee, H. Lee, T. Hahn, and S. Kang, "Automatic generation of game content using a Graph-based wave function collapse algorithm," in *Proc. IEEE Conf. Games*, 2019, pp. 1–4, doi: 10.1109/CIG.2019.8848019.
5. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 405–421, doi: 10.1007/978-3-030-58452-8_24.
6. A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "Plenoctrees for real-time rendering of neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5752–5761.
7. P. Hedman, P. P. Srinivasan, B. Mildenhall, J. T. Barron, and P. Debevec, "Baking neural radiance fields for real-time view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5875–5884.
8. Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "NeRF–: Neural radiance fields without known camera parameters," 2021, *arXiv:2102.07064*.
9. Q. Chen, Q. Wu, R. Tang, Y. Wang, S. Wang, and M. Tan, "Intelligent home 3D: Automatic 3D-house design from linguistic descriptions only," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12625–12634.
10. H. Howard-Jenkins, S. Li, and V. Prisacariu, "Thinking outside the box: Generation of unconstrained 3D room layouts," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 432–448.
11. A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.
12. M. Suzuki and Y. Matsuo, "A survey of multimodal deep generative models," *Adv. Robot.*, vol. 36, no. 5/6, pp. 261–278, 2022.

13. V. Verdot and A. Saidi, "Virtual hybrid communications–a telecom infrastructure for the metaverse," *J. Virtual Worlds Res.*, vol. 4, no. 3, pp. 1–10, 2011.
14. K. Yoon, S.-K. Kim, S. P. Jeong, and J. -H. Choi, "Interfacing cyber and physical worlds: Introduction to IEEE 2888 standards," in *Proc. IEEE Int. Conf. Intell. Reality*, 2021, pp. 49–50, doi: 10.1109/ICIR51845.2021.00016.
15. D. Liu, Y. Li, J. Lin, H. Li, and F. Wu, "Deep learning-based video coding: A review and a case study," *ACM Comput. Surv.*, vol. 53, no. 1, pp. 1–35, 2020.
16. M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3516–3530, Dec. 2019, doi: 10.1109/TCSVT.2018.2886277.
17. M. N. Akcay, B. Kara, S. Ahsan, A. C. Begen, I. Curcio, and E. Aksu, "Head-motion-aware viewport margins for improving user experience in immersive video," *ACM Multimedia Asia, Article*, vol. 43, pp. 1–5, 2021, doi: 10.1145/3469877.3490573.

SHU-CHING CHEN is currently a Professor with Florida International University, Miami, FL, USA. Contact him at chens@cs.fiu.edu.