Graduate School of Advanced Science and Engineering
Waseda University

博士論文概要
Doctoral Dissertation Synopsis

論文題目
Dissertation Title

Improving Mixed Reality with Multi-task Scene Understanding and Data Augmentation

マルチタスク学習を用いたシーン理解とデータ拡張による複合現実感の向上

申請者
(Applicant Name)
Qi FENG
馮起

Department of Pure and Applied Physics, Research on Image Processing

May, 2022

Mixed reality as a broad concept describes a process of blending the physical world and the digital environment computed and rendered in real-time by computers, and it has been experiencing increased popularity with more affordable hardware and polished software. Mixed reality extends both ways from a completely virtual world (virtual reality) and our perceived reality (augmented reality), enabling numerous potential interactions and applications. When putting more emphasis on the virtual environment, virtual reality allows users to enter an entirely computer-rendered environment and interact intuitively. This sparks studies of human-computer interactions in training, entertainment, education, and many other topics. While the disconnection from reality is the nature of virtual reality, augmented reality has drawn more attention with its capability to present computer-generated visualizations in addition to the real world. However, achieving a more sophisticated and immersive mixed reality experience is a challenging task. Due to the difficulty of accurately understanding the complex relationship between the virtual environment and the physical world, seamlessly merging the two is only possible with an effective and efficient understanding of the scene.

As one of the fundamental components of mixed reality, scene understanding has seen significant progress over recent years thanks to more advanced deep learning algorithms. By sharing learned knowledge between several related tasks implicitly and explicitly, multi-task learning generalizes better for original tasks and has gained great popularity with its great capability and accuracy. Despite the great capabilities of the multi-task scene understanding, obtaining high-quality paired training data in different representations for training mixed reality tasks is critical and challenging. While there are abundant samples for traditional computer vision tasks such as object recognition, in the context of scene understanding in mixed reality, many tasks suffer from low-quality databases. For instance, captured photos that are intended for mixed reality usage are usually stored with an entirely different projection with strong distortions, rendering traditional perspective training databases less effective. As a result, while possessing great potential, using multi-task scene understanding to solve challenges in mixed reality is a less explored field.

In this thesis, we investigate using multi-task learning-based scene understanding algorithms to improve immersive mixed reality. We will focus on several topics of scene understanding tasks that are particularly crucial in mixed reality: semantic segmentation, depth prediction, and pose estimation. Semantic segmentation can provide each pixel with correct labels to satisfy the necessity of high-level understanding in the mixed reality environment. Depth prediction helps understand the size, shape, and distance of a physical object. It directly enables occlusions to correctly render the virtual augmentation which is a key aspect of an immersive experience. Finally, pose estimation allows more intuitive interactions and high degree-of-freedom designs which were unfeasible using controllers. It is well-known that labeled training data are crucial for learning-based approaches, and great performance high-capacity deep learning models need an adequate number of annotated samples. To accommodate the proposed multi-task scene understanding methods, we further combine different data augmentation techniques to obtain high-quality large-scale databases that are currently scarce or not available.

The structure of the thesis follows the order of different spatial scopes, local, regional, and global, to understand and improve mixed reality. After a brief review of the literature, we start from a smaller scale of understanding users' hand-object interactions to resolve occlusions by exploring the relationship between two tasks: pose tracking and semantic segmentation. We then focus on the foreground objects of mixed reality scenes. We choose humans as an example to demonstrate the capability of simultaneously predicting depth and semantic segmentation with different network designs. Later, we propose to comprehend the global environment with multi-task learning of two representations, equirectangular projection, and cubemap projection, and show its usage in mixed reality. Finally, we demonstrate practical mixed reality applications of multi-task scene understanding by conveying the knowledge of depth estimation and semantic segmentation to an image inpainting algorithm. In the remainder of the thesis, we will discuss limitations and give potential directions for future computer vision for mixed reality. The dissertation consists of the following 6 chapters.

Chapter 1 introduces this study with research background of mixed reality and scene understanding. We start from the definition and processing principles of existing mixed reality to highlight the importance and functions of underlying scene understanding capabilities including object recognition, semantic segmentation, depth prediction, and pose estimation. After reviewing relevant state-of-the-art algorithms, we define the research problems and objectives of this study. An explanation of our methodology to approach the problem and a brief overview of each chapter conclude this chapter.

Grasping the Local: Solving Hand-object Occlusion in Mixed Reality (Chapter 2). The hand is one of the key components in mixed reality, and hand-object interactions are critical to a wide range of MR applications such as surgery simulations. However, their practicality and immersive experiences are severely limited by occlusions. In Chapter 2, we first revisit existing occlusion solutions, followed by explaining the proposed RGBD database generated with data augmentation, and then a novel joint learning process to predict hand postures and masks. We finally present our novel two-step approach to resolving the occlusions in mixed reality with implementation details, evaluations, and a user study. This research can be applied to egocentric mixed reality applications that include hand-object interactions such as apparatus-involved training.

Observing the Regional: Foreground-aware 360-degree Depth Prediction (Chapter 3). Although the ability to predict depth from a single 360-degree image can benefit plentiful applications, existing approaches produce sub-optimal results for foreground objects. In this chapter, we propose to augment databases with realistic foregrounds with an image-based approach and design a novel auxiliary deep neural network to predict depth and semantic segmentation simultaneously. We further design a bi-projection-based network to improve the capability of understanding the foreground object. We demonstrate the system using humans as the foreground due to its complexity and contextual importance and show consistent and accurate local estimations compared with state-of-the-arts.

Comprehending the Global: 360-degree Depth Prediction in the Wild (Chapter 4). Although data-driven learning-based methods demonstrate significant potential in understanding the entirety of 360-degree images, scarce training data and ineffective 360-degree estimation algorithms are still two key limitations hindering accurate estimation across diverse domains. In this chapter, we first establish a large-scale database by exploring the use of a plenteous source of data, 360-degree videos from the internet, using a test-time training method. We then propose an end-to-end two-branch multi-task learning network, SegFuse, that mimics the human eye to effectively learn from the dataset and estimate high-quality depth maps from diverse monocular RGB images. We showcase that our method has a great understanding of the global mixed reality scene under arbitrary conditions.

Employing Scene Understanding in Immersive Mixed Reality (Chapter 5). With the established understanding of different scales of scenes, we explore the practical applications of mixed reality in Chapter 5. We propose to convey the knowledge of depth estimation and semantic segmentation to an image inpainting algorithm to solve a practical mixed reality problem. We expect this application-focused chapter can shed more light on more practical employments of newer scene understanding algorithms in the modern virtual/augmented reality era.

Conclusion (Chapter 6). In Chapter 6, we start with a summary of the work. We then discuss the limitations of current scene understanding in upcoming mixed reality and attempt to explore some promising directions for alike future research that focus on solving the computer vision aspect of mixed reality technologies.

List of research achievements for application of Doctor of Engineering, Waseda University

Full Name： 馮　起　　seal or signature

Date Submitted(yyyy/mm/dd): 2022/4/25

種類別 (By Type)	題名、　発表・発行掲載誌名、　　発表・発行年月、　　連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
学術誌 原著論文	1. Nozawa Naoki, Shum P. H. Hubert, Feng Qi, Ho S. L. Edmond, Morishima Shigeo, "3D car shape reconstruction from a contour sketch using GAN and lazy learning", The Visual Computer, 1-14, April 2021. (DOI: https://doi.org/10.1007/s00371-020-02024-y)
学術誌 原著論文	○ 2. Feng Qi, Hubert P. H. Shum, Shigeo Morishima, "Resolving hand-object occlusion for mixed reality with joint deep learning and model optimization", Computer Animation and Virtual Worlds, 31(4-5), e1956, September 2020. (DOI: https://doi.org/10.1002/cav.1956)
学術誌 原著論文	3. Shimamura Ryo, Feng Qi, Koyama Yuki, Nakatsuka Takayuki, Fukayama Satoru, Hamasaki Masahiro, Goto Masataka, Morishima Shigeo, "Audio–visual object removal in 360-degree videos", The Visual Computer, 36(10), 2117-2128, October 2020. (DOI: https://doi.org/10.1007/s00371-020-01918-1)
学術誌 原著論文	○ 4. Feng Qi, Hubert P. H. Shum, Shigeo Morishima, "Foreground-aware Dense Depth Estimation for 360 Images", Journal of WSCG, 28(1-2), 79-88, June 2020. (DOI: https://doi.org/10.24132/JWSCG.2020.28.10)
査読のある 国際会議論文	○ 1. Feng Qi, Hubert P. H. Shum, Shigeo Morishima, "360 Depth Estimation in the Wild - The Depth360 Dataset and the SegFuse Network", IEEE conference on virtual reality and 3D user interfaces (VR), Pages 664-673, New Zealand (online), March 2022. (DOI: https://doi.org/10.1109/VR51125.2022.00087)
査読のある 国内会議論文	○ 2. Feng Qi, Hubert P. H. Shum, Shigeo Morishima, "Bi-projection-based Foreground-aware Omnidirectional Depth Prediction", Visual Computing + VC Communications, Pages 1-6, Tokyo (online), September 2021.
査読のある 国際会議論文	○ 3. Feng Qi, Hubert P. H. Shum, Shigeo Morishima, "Foreground-aware Dense Depth Estimation for 360 Images", International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, Pages 79-88, The Czech Republic (online), May 2020.
査読のある 国際会議論文	4. Shimamura Ryo, Feng Qi, Koyama Yuki, Nakatsuka Takayuki, Fukayama Satoru, Hamasaki Masahiro, Goto Masataka, Morishima Shigeo, "Audio–visual object removal in 360-degree videos", Computer Graphics International 2020, Pages 1-8, Geneva (online), October 2020.

List of research achievements for application of Doctor of Engineering, Waseda University

Full Name： 馮　起 seal or signature

Date Submitted(yyyy/mm/dd): 2022/4/25

種類別 (By Type)	題名、　発表・発行掲載誌名、　　発表・発行年月、　　連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
査読のある 国際会議論文	○ 5. Feng Qi, Hubert P. H. Shum, Shigeo Morishima, "Resolving occlusion for 3D object manipulation with hands in mixed reality", Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology, Pages 1-2, Tokyo, November 2018.
査読のない 学会発表論文	1. Feng Qi, Hubert P. H. Shum, Shigeo Morishima, "Occlusion for 3D Object Manipulation with Hands in Augmented Reality", In Proceedings of The 21st Meeting on Image Recognition and Understanding, Pages 1-4, Sapporo, August 2018.