Chapter 1
Introduction

Nadia Magnenat Thalmann, Jian Jun Zhang, Manoj Ramanathan, and Daniel Thalmann

Technological advances have a tremendous impact on everyone's life and are interwoven into everything we do. The computer was a significant invention in the last century, which has quietly transformed and morphed into many devices that have rapidly become indispensable to our daily life, such as phones, cameras, TVs, and autonomous vehicles. Many such devices and equipment have a computer engine embedded: cars, washing machines, ovens, microwaves, and digital books. Even lights and coffee machines can be operated from a smartphone with a Home Facilities app. There has been a tremendous increase in their capabilities due to the inbuilt computer systems. For instance, digital cameras can now recognize faces, perform auto focus, and enact numerous different functions, allowing us to capture the details of our lives with very realistic photos. However, the development of such functions relies more and more on the recognition and the creation of 3D scenes, making Intelligent Scene Modeling (ISM) a key field of research and development.

The smartphone is a key example of the fast development of the technology in recent years. The phone is now an essential device for everyone as it is able to perform

N. Magnenat Thalmann (✉) · M. Ramanathan
Nanyang Technological University, Singapore, Singapore
e-mail: thalmann@miralab.ch

M. Ramanathan
e-mail: ramanthan.manoj@gmail.com

N. Magnenat Thalmann
University of Geneva, Geneva, Switzerland

J. J. Zhang
Bournemouth University, Poole, UK
e-mail: Jzhang@bournemouth.ac.uk

D. Thalmann
EPFL, Lausanne, Switzerland
e-mail: Daniel.Thalmann@epfl.ch

N. Magnenat Thalmann et al. (eds.), *Intelligent Scene Modeling and Human-Computer Interaction*, Human–Computer Interaction Series,
https://doi.org/10.1007/978-3-030-71002-6_1

many tasks a person chooses. There is no longer a need to use a specific device such as a camera to take photos or a receptor to listen to the radio. The phone is no longer just used for voice communications. In fact, the smartphone is rather like a pocket-sized computer with inbuilt devices allowing us to communicate with our family and friends, to pay for our cup of coffee, or even to board a plane. It is now possible for any of these devices to recognize faces and objects through computer vision, to understand speech through natural language processing to name just a few. Recent progress on machine learning and deep learning methods has enabled the devices to operate not only more intelligently, but with less power and memory consumption.

The development of AI and deep learning methods has made these devices understand more about the 3D environment and the users involved. This has also led to the substantial advances in graphics (to show/render any output), virtual reality (to create immersive and interactive virtual environments) and augmented reality (to combine virtual entities with real-life scenes). These devices are now starting to emulate human behaviors and actions in many possible scenarios leading to the creation of Artificial Intelligence (AI)-based robots and virtual humans. AI has helped reduce the size of the devices and transform them into different products in addition to the significant elevation of the machine intelligence. AI-based technology, machine learning, and deep learning-based technology are becoming increasingly viable to every device. For instance, smart voice assistants such as Alexa from Amazon, Google Assistant, and Apple Siri have the ability to understand and respond to user speech and commands. Most of these devices allow interconnections, meaning that phones can be connected to voice assistants, and videos can be cast from phones or computers on to the TV screen making the user experience smooth and easy. The field of Human–Computer Interaction (HCI) has indeed dramatically changed and will continue these very fast transformations. As mentioned above, computers and phones can now interact in different ways to support our daily needs. Apart from the conventional methods of interaction such as keyboards and mouse, computers and devices can now interact through touch screens, voices and gestures, etc. For instance, face recognition or fingerprint recognition are more and more used to unlock computers and phones instead of keying in passwords. HCI is an ever-expanding research field with different modalities of interaction and different ways of understanding the various environments and user-related cues. With such a broad scope, this book aims to highlight the state of the art in HCI and its future. Specifically, we will look at different modalities of interaction and different algorithms in understanding the cues, and how they are applied in several devices ranging from PC, phones to robots, and virtual humans.

An easy and natural way of interaction with a computer is through a digital camera. It serves as an eye and provides visual cues to understand the environment and people within. The rise of computer vision (CV) algorithms has made the camera an integral part of any computer or smartphone. CV algorithms form the backend or backbone for several tasks performed by the devices. Object detection, action recognition, face recognition, emotion recognition, etc. are considered today as basic and core modules for any computer vision related interactions. Despite the wide use of CV techniques, computer vision itself as a fast-growing research discipline has seen a number of

open research challenges facing the research community implying a pressing need for intensive research efforts.

The book is organized into two sections: Intelligent Scene Modeling (ISM) and Human–Computer Interaction (HCI).

In the first section on ISM, after the introduction Chap. 1 of this book, for an optimum visual content understanding, we begin with simple 2D retrieval and recognition. In Chap. 2, Hanhui Li et al. present a comprehensive study on available object detection methods based on the recent development in deep learning. They introduce a general framework of utilizing deep learning techniques to detect object, which becomes a de facto solution for object detection, In Chap. 3, Weng Junwu et al. describe a non-parametric model for skeleton-based action/gesture recognition. In Chap. 4, Hui Liang et al. discuss about random forest-based Hough-voting techniques that form the basis of several computer vision algorithms such as pose estimation and gesture recognition. They propose further improvements to the algorithms and conduct experiments in several vision-related applications. This chapter is at the forefront of AI-based learning techniques as it combines two advanced methods: Random Forests and Hough Voting.

For a comprehensive understanding of environment, simple 2D recognition and retrieval are not sufficient. Understanding the 3D arrangement of a scene is essential for the computer or the device to conduct the tasks. Depth cameras or RGBD cameras such as Kinect, HoloLens devices have become increasingly important as they help perceive and understand 3D objects and environments. 3D scene modeling and 3D reconstruction of scenes are important research fields in computer vision.

In Chap. 5, Zeynep, Cipiloglu, Yildiz et al. present a survey of AI-based solutions for modeling human perception of 3D scenes which includes topics such as modeling human visual attention, 3D object shape and quality perception, and material recognition. The authors emphasize the impact of deep learning methods in several aspects of human perception, in particular for visual attention, visual quality perception, and visual material recognition.

In Chaps. 6, 7, 8, and 9, semantic scene modeling and rendering are being described. To understand a scene, it is necessary for the computer to analyze the 3D scene layout, the spatial functional and semantic relationship between the various 3D objects detected. Semantic modeling and rendering allow computers to recreate and render immersive graphics outputs for the end user. The ability to reconstruct scenes and render them plays an important role in virtual reality and augmented reality applications. It also opens a new way of interaction with the objects represented in the virtual world. To explain 3D scene modeling and reconstruction, several chapters in this book are dedicated to these topics. In particular, in Chap. 6, Yonghyun Jung et al. describe the reconstruction of 3D real-world objects using Microsoft HoloLens. The next three chapters specifically focus on semantic scene modeling and rendering starting from indoor environments to complete 3D semantics-based building models. In Chap. 7, Divya Udayan et al. provide examples of segmentation on façade components using deep learning methods. In Chap. 8, Yinyu Nie et al. describe a method for automatic indoor scene modeling based on deep convolutional features, In Chap. 9,

Pradeep Kumar Jayaraman et al. introduce a method for interactively grouping and labeling the faces of building models.

Three-dimensional reconstruction is not only essential for semantic scene understanding and modeling, it also opens research avenues and applications in various areas presented in Chaps. 10 and 11. In Chap. 10, Evropi Stefanidi et al. discuss a new tool called "TooltY", a 3D authoring platform to demonstrate simple operations of tools such as hammer, scissors, screwdriver, which are direct products of 3D reconstruction. The 3D reconstructed tools can be immersively experienced in Virtual environments. Additionally, they can be used to recreate 3D faces, their emotions, the expressions, and the subtleties involved. This ability to recreate faces has been a key aspect for many new applications. For instance, 3D Animoji in phones can representatively replicate facial expressions and voice messages of the user. In Chap. 11, Hyewon Seo et al. propose a recurrent neural network with a marker-based shape representation as the base to generate 3D facial expressions. The authors explain how the fast development of deep learning started to replace linear function approximators with deep neural networks to achieve drastically improved performance. In the complementary area of Big Data, visualization tools and user interfaces have become essential to analyze massive amounts of information and make data-driven decisions.

In the second section of the book, chapters deal with HCI-related topics. In Chap. 12, Yasin Findik et al. introduce an assistive analytical agent that can help with decision-making in exploratory and collaborative visual analytics sessions.

One of the newest and fast-developing topics is the embodiment of the computer through virtual humans or social robots. In order to create realistic and reliable human–computer interactions, virtual humans or humanoid robots can understand and mimic possible human behaviors in many situations. It opens different ways of communication including, for example, non-verbal interactions. They are an integral part of high-level Human–Computer Interaction.

A good example of new Human–Computer Interface can be found in Chap. 13. Evangelia Baka et al. review the currently available technologies in HCI with a focus on virtual humans and robots. In Chap. 14, Manoj Ramanathan et al. provide an insight into currently available non-verbal interaction methods with social robots. In Chap. 15, Nidhi Mishra et al. explore the possibility of using social robots in different roles such as customer guide, teacher, and companion, and study user acceptance of social robots in these roles using Human–Robot Interaction experiments.

Over the years, HCI has undergone a massive transformation with changes in hardware and software possibilities. The ability to shrink computing engines has led to the development of numerous devices that can be interconnected to each other. The development of machine learning and deep learning algorithms has contributed to a boom in artificial intelligence. Combining these developments has broadened the horizon of ISM and HCI, as there is a necessity to build complex scenes and to perceive, process, and interpret data from various sources such as cameras, microphones, and wearable sensors. Due to the fast and ever-changing nature of ISM and HCI, it is essential to understand and update the latest research methods available.

To summarize, this book aims to provide an insight into the most recent developments in ISM and HCI with new machine and deep learning methods, how these two complementary fields have expanded in modeling, perception, processing, and inference of multimodal data such as audio, video, speechless cues, and sensor data.

This book is specially dedicated to researchers and Masters/Ph.D. students who will enjoy a comprehensive overview of ISM and HCI methods, containing surveys, state-of-the-art reviews, novel research, and concrete results. In addition, the book provides new research methods in various global fields linked to both ISM and HCI and their integration including computer vision, 3D scene modeling, virtual humans, and social robots.

The contributors of this book are international researchers from diversified research backgrounds and universities working under the broad umbrellas of ISM and HCI or both. This allows the readers to understand the evolution of ISM and HCI over time and the current limitations.

Acknowledgements Part of the research described in this book is supported by the BeingTogether Centre, a collaboration between Nanyang Technological University (NTU) Singapore and University of North Carolina (UNC) at Chapel Hill. The BeingTogether Centre is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.