Advances in Computer Vision and Pattern Recognition

Paul L. Rosin
Yu-Kun Lai
Ling Shao
Yonghuai Liu *Editors*

RGB-D Image Analysis and Processing

Springer

Advances in Computer Vision and Pattern Recognition

More information about this series at http://www.springer.com/series/4205

Paul L. Rosin · Yu-Kun Lai ·
Ling Shao · Yonghuai Liu
Editors

RGB-D Image Analysis and Processing

Springer

Editors

Paul L. Rosin
School of Computer Science
and Informatics
Cardiff University
Cardiff, UK

Yu-Kun Lai
School of Computer Science
and Informatics
Cardiff University
Cardiff, UK

Ling Shao
IEEE
University of East Anglia
Norwich, UK

Yonghuai Liu
Department of Computer Science
Edge Hill University
Ormskirk, UK

ISSN 2191-6586 ISSN 2191-6594 (electronic)
Advances in Computer Vision and Pattern Recognition
ISBN 978-3-030-28602-6 ISBN 978-3-030-28603-3 (eBook)
https://doi.org/10.1007/978-3-030-28603-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Colours become weaker in proportion to their distance from the person who is looking at them.
Leonardo da Vinci, *Treatise on Painting*, 1651.

Leonardo da Vinci used *aerial perspective*, as defined above, to good effect. Nevertheless, for thousands of years, artists have had to struggle to capture depth in their images. It was only with the introduction of RGB-D sensors in recent years that capturing depth along with colour has become possible. Moreover, Microsoft's release of the immensely successful Kinect for the mass consumer market in 2010 was literally a game changer, making real-time RGB-D cameras more affordable, accessible, widespread, mainstream and more fun! While the Kinect was designed for home game controller applications, researchers and practitioners quickly realised that its ability as a natural user interface could be deployed in many other scenarios. And so today, RGB-D cameras are ubiquitous, ranging from expensive industrial scanners to webcams and smartphones.

Recent years have continued to see technical developments on RGB-D sensors, both in terms of hardware and software. Data capture is now relatively mature, but understanding and analysing the data remains challenging. Not surprisingly, giving its overwhelming success in many areas, deep learning has also been applied to RGB-D; not only is it effective at processing RGB-D images, but is increasingly used for the challenging task of monocular depth estimation, i.e. creating the -D directly from a single standard (i.e. passive) RGB image. However, despite all these advances, there remain many challenges, ensuring the continuation of active research and development in RGB-D. At the data acquisition stages (depending on which sensing technology is used), examples are coping with general scenes and unconstrained conditions, reflections, transparent surfaces and background light. Subsequent processing typically needs to be performed to remove noise, replace missing depth values and merge sequential RGB-D scans or multiple RGB-D camera outputs to reconstruct objects and scenes. Mid-level processing then consists of tasks such as segmentation and object detection, which remain active research topics both within the RGB-D community as well as the general computer vision community.

This book is structured to reflect such a breakdown into RGB-D data acquisition and processing followed by RGB-D data analysis, which then sets the scene for the final section on RGB-D applications. A set of chapters has been assembled to provide a thorough introduction to the area, with sufficient technical detail to prepare the reader for research and development with RGB-D imagery.

The future will continue to see increasing takeup of RGB-D. The wide availability of RGB-D sensors means that more data is becoming available, consequently facilitating improvements to be made via machine learning. In addition, further improvements in both the hardware and software will help extend the range of possible applications. As RGB-D sensors become smaller and reduce their power consumption, then emerging uses, that would have been impractical just a few years ago, are becoming more widespread and mainstream. Some examples are wearable RGB-D systems (e.g. providing navigation for the visually impaired), face recognition on mobile phones (biometrics), online shopping (e.g. virtual try-on for clothing), 3D mapping using drones and many more applications in health care, gaming, industry, etc. The improved capability to capture 3D environment and shapes also facilitates downstream applications, such as Augmented Reality and 3D printing.

In the future, RGB-D sensing can continue to draw from developments in the core technologies of image processing and computer vision. And just as Leonardo da Vinci's inventive mind was forever seeking out new ways of interpreting the world, we believe researchers will continue to be pushing RGB-D sensing forward to new approaches and applications in the future.

July 2019

Paul L. Rosin
Cardiff University, Cardiff, UK

Yu-Kun Lai
Cardiff University, Cardiff, UK

Ling Shao
Inception Institute of Artificial Intelligence, Abu Dhabi
United Arab Emirates

Yonghuai Liu
Edge Hill University, Ormskirk, UK

Contents

Part I RGB-D Data Acquisition and Processing

1 Commodity RGB-D Sensors: Data Acquisition 3
Michael Zollhöfer

2 Dealing with Missing Depth: Recent Advances in Depth Image Completion and Estimation 15
Amir Atapour-Abarghouei and Toby P. Breckon

3 Depth Super-Resolution with Color Guidance: A Review 51
Jingyu Yang, Zhongyu Jiang, Xinchen Ye and Kun Li

4 RGB-D Sensors Data Quality Assessment and Improvement for Advanced Applications 67
Pablo Rodríguez-Gonzálvez and Gabriele Guidi

5 3D Reconstruction from RGB-D Data 87
Charles Malleson, Jean-Yves Guillemaut and Adrian Hilton

6 RGB-D Odometry and SLAM 117
Javier Civera and Seong Hun Lee

7 Enhancing 3D Capture with Multiple Depth Camera Systems: A State-of-the-Art Report 145
Oscar Meruvia-Pastor

Part II RGB-D Data Analysis

8 RGB-D Image-Based Object Detection: From Traditional Methods to Deep Learning Techniques 169
Isaac Ronald Ward, Hamid Laga and Mohammed Bennamoun

9 RGB-D Salient Object Detection: A Review 203
Tongwei Ren and Ao Zhang

10 Foreground Detection and Segmentation in RGB-D Images 221
Runmin Cong, Hao Chen, Hongyuan Zhu and Huazhu Fu

11 Instance- and Category-Level 6D Object Pose Estimation 243
Caner Sahin, Guillermo Garcia-Hernando, Juil Sock and Tae-Kyun Kim

12 Geometric and Semantic Modeling from RGB-D Data 267
Song-Hai Zhang and Yu-Kun Lai

Part III RGB-D Applications

13 Semantic RGB-D Perception for Cognitive Service Robots 285
Max Schwarz and Sven Behnke

14 RGB-D Sensors and Signal Processing for Fall Detection 309
Susanna Spinsante

15 RGB-D Interactive Systems on Serious Games for Motor Rehabilitation Therapy and Therapeutic Measurements 335
Gabriel Moyà-Alcover, Ines Ayed, Javier Varona and Antoni Jaume-i-Capó

16 Real-Time Hand Pose Estimation Using Depth Camera 355
Liuhao Ge, Junsong Yuan and Nadia Magnenat Thalmann

17 RGB-D Object Classification for Autonomous Driving Perception 377
Cristiano Premebida, Gledson Melotti and Alireza Asvadi

18 People Counting in Crowded Environment and Re-identification 397
Emanuele Frontoni, Marina Paolanti and Rocco Pietrini

Appendix: References 427

Index 519

Contributors

Alireza Asvadi Laboratory of Medical Information Processing (LaTIM), University of Western Brittany, Brest, France

Amir Atapour-Abarghouei Department of Computer Science, Durham University, Durham, UK

Ines Ayed Departament de Ciències Matemàtiques i Informàtica, GresCom Lab, Ecole Supèrieure des Communications de Tunis, Universitè de Carthage, tunis, Tunisia

Sven Behnke Autonomous Intelligent Systems, Computer Science Institute VI University of Bonn, Bonn, Germany

Mohammed Bennamoun University of Western Australia, Crawley, WA, Australia

Toby P. Breckon Departments of Engineering & Computer Science, Durham University, Durham, UK

Hao Chen Department of Mechanical Engineering, City University of Hong Kong, Hong Kong SAR, China

Javier Civera I3A, Universidad de Zaragoza, Zaragoza, Spain

Runmin Cong Beijing Key Laboratory of Advanced Information Science and Network Technology, Institute of Information Science, Beijing Jiaotong University, Beijing, China

Emanuele Frontoni Department of Information Engineering, Universitá Politecnica delle Marche, Ancona, Italy

Huazhu Fu Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Guillermo Garcia-Hernando Department of Electrical-Electronic Engineering, Imperial Computer Vision and Learning Lab (ICVL), Imperial College, London, UK

Liuhao Ge Institute for Media Innovation, Nanyang Technological University Singapore, Singapore, Singapore

Gabriele Guidi Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy

Jean-Yves Guillemaut Centre for Vision, Speech and Signal Processing & University of Surrey, Guildford, UK

Adrian Hilton Centre for Vision, Speech and Signal Processing & University of Surrey, Guildford, UK

Antoni Jaume-i-Capó Departament de Ciències Matemàtiques i Informàtica, Universitat de les Illes Balears, Palma (Illes Balears), Spain

Zhongyu Jiang Tianjin University, Tianjin, China

Tae-Kyun Kim Department of Electrical-Electronic Engineering, Imperial Computer Vision and Learning Lab (ICVL), Imperial College, London, UK

Hamid Laga Murdoch University, Perth, WA, Australia;
The Phenomics and Bioinformatics Research Centre, University of South Australia, Adelaide, SA, Australia

Yu-Kun Lai School of Computer Science and Informatics, Cardiff University, Cardiff, UK

Seong Hun Lee I3A, Universidad de Zaragoza, Zaragoza, Spain

Kun Li Tianjin University, Tianjin, China

Nadia Magnenat Thalmann Institute for Media Innovation, Nanyang Technological University Singapore, Singapore, Singapore

Charles Malleson Centre for Vision, Speech and Signal Processing & University of Surrey, Guildford, UK

Gledson Melotti Department of Electrical and Computer Engineering (DEEC), University of Coimbra, Coimbra, Portugal

Oscar Meruvia-Pastor Department of Computer Science, Memorial University of Newfoundland, St. John’s, NL, Canada

Gabriel Moyà-Alcover Departament de Ciències Matemàtiques i Informàtica, Universitat de les Illes Balears, Palma (Illes Balears), Spain

Marina Paolanti Department of Information Engineering, Universitá Politecnica delle Marche, Ancona, Italy

Rocco Pietrini Department of Information Engineering, Universitá Politecnica delle Marche, Ancona, Italy

Cristiano Premebida Department of Aeronautical and Automotive Engineering (AAE), Loughborough University, Loughborough, UK

Tongwei Ren Software Institute, Nanjing University, Nanjing, China

Pablo Rodríguez-Gonzálvez Department of Mining Technology, Topography and Structures, Universidad de León, Ponferrada, Spain

Caner Sahin Department of Electrical-Electronic Engineering, Imperial Computer Vision and Learning Lab (ICVL), Imperial College, London, UK

Max Schwarz Autonomous Intelligent Systems, Computer Science Institute VI University of Bonn, Bonn, Germany

Juil Sock Department of Electrical-Electronic Engineering, Imperial Computer Vision and Learning Lab (ICVL), Imperial College, London, UK

Susanna Spinsante Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Ancona, Italy

Javier Varona Departament de Ciències Matemàtiques i Informàtica, Universitat de les Illes Balears, Palma (Illes Balears), Spain

Isaac Ronald Ward University of Western Australia, Crawley, WA, Australia

Jingyu Yang Tianjin University, Tianjin, China

Xinchen Ye Dalian University of Technology, Dalian, China

Junsong Yuan Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY, USA

Ao Zhang Software Institute, Nanjing University, Nanjing, China

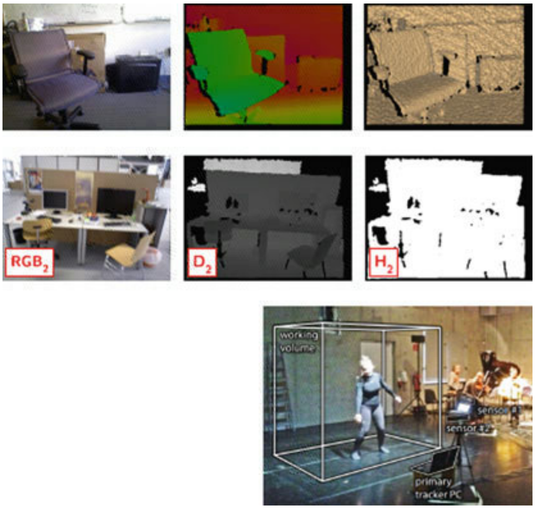
Song-Hai Zhang Department of Computer Science and Technology, Tsinghua University, Beijing, China

Hongyuan Zhu Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, Singapore

Michael Zollhöfer Stanford University, Stanford, CA, USA

Part I
RGB-D Data Acquisition and Processing

Part I of this book focuses on RGB-D data acquisition and processing. The two main approaches for capturing RGB-D images are passive and active sensing. In addition, with the rise of deep learning, monocular depth estimation has become possible, and is becoming increasingly popular. For the first two approaches, the images often have missing values (i.e. holes) which need to be filled, or low-resolution depth maps which need to be upsampled. RGB-D video enables active depth capture in which the sensor moves within a static scene, with the individual captures fused to produce a 3D reconstruction of the scene. Multiple RGB-D cameras can also be deployed, which facilitates reconstruction of dynamic scenes. Since low-cost RGB-D sensors will not have top quality data, it is important to consider a metrological analysis of their performance.



An RGB-D camera jointly captures colour and depth images, the latter describing the 3D geometry of the scene. The RGB-D acquisition process is described in Chap. 1. Along with the RGB-D image, the second row of images shows the hole mask indicating missing depth values, as described in Chap. 2. The third row shows a multiple camera setup using Kinects to capture a performer's motion (Chap. 7).

Chapter 1
Commodity RGB-D Sensors: Data Acquisition

Michael Zollhöfer

Abstract Over the past 10 years, we have seen a democratization of range sensing technology. While previously range sensors have been highly expensive and only accessible to a few domain experts, such sensors are nowadays ubiquitous and can even be found in the latest generation of mobile devices, e.g., current smartphones. This democratization of range sensing technology was started with the release of the Microsoft Kinect, and since then many different commodity range sensors followed its lead, such as the Primesense Carmine, Asus Xtion Pro, and the Structure Sensor from Occipital. The availability of cheap range sensing technology led to a big leap in research, especially in the context of more powerful static and dynamic reconstruction techniques, starting from 3D scanning applications, such as KinectFusion, to highly accurate face and body tracking approaches. In this chapter, we have a detailed look into the different types of existing range sensors. We discuss the two fundamental types of commodity range sensing techniques in detail, namely passive and active sensing, and we explore the principles these technologies are based on. Our focus is on modern active commodity range sensors based on time of flight and structured light. We conclude by discussing the noise characteristics, working ranges, and types of errors made by the different sensing modalities.

1.1 Introduction

Modern conventional color cameras are ubiquitous in our society and enable us to capture precious memories in a persistent and digital manner. These recordings are represented as millions of three channel pixels that encode the amount of red, green, and blue light that reached the sensor at a corresponding sensor location and time. Unfortunately, color images are an inherently flat 2D representation, since most of the 3D scene informations is lost during the process of image formation.

M. Zollhöfer (✉)
Stanford University, 353 Serra Mall, Stanford, CA 94305, USA
e-mail: zollhoefer@cs.stanford.edu

P. L. Rosin et al. (eds.), *RGB-D Image Analysis and Processing*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-28603-3_1

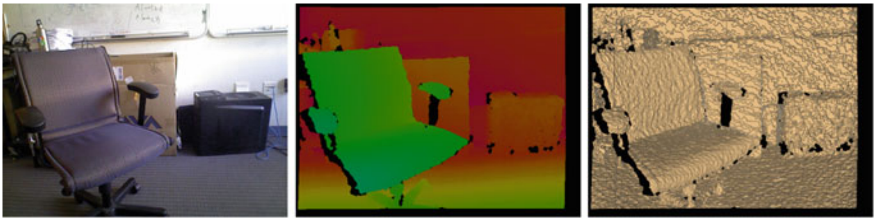

(a) Color (b) Depth (c) Phong

Fig. 1.1 An RGB-D camera jointly captures color (a) and depth (b) images. The depth image encodes the distance to the scene on a per-pixel basis. Green color means that this part of the scene is close to the camera and red means that it is far away. The Phong shaded image (c) is an alternative visualization of the 3D geometry

Over the past 10 years, we have seen a democratization of a new class of cameras that enables the dense measurement of the 3D geometry of the observed scene, thus overcoming the mentioned limitation of conventional color cameras. These so-called *range* or *depth sensors* perform a dense per-pixel measurement of scene depth, i.e., the distance to the observed points in the scene. These measured depth values are normally exposed to the user in the form of a *depth image*, which is a 2.5-dimensional representation of the visible parts of the scene. An *RGB-D sensor* is the combination of a conventional color camera (RGB) with such a depth sensor (D). It enables the joint capture of scene appearance and scene geometry at real-time frame rates based on a stream of color $\mathscr{C}$ and depth images $\mathscr{D}$. Figure 1.1 shows an example of such a color (a) and depth image pair (b). The phong-shaded image (c) is an alternative visualization of the captured 3D geometry that better illustrates the accuracy of the obtained depth measurements. Current RGB-D sensors provide a live stream of color and depth at over 30 Hz.

Starting with the *Microsoft Kinect*, over the past 10 years a large number of commodity RGB-D sensors have been developed, such as the *Primesense Carmine*, *Asus Xtion Pro*, *Creative Senz3D*, *Microsoft Kinect One*, *Intel Realsense*, and the *Structure Sensor*. While previous range sensors [8, 9, 19] were highly expensive and only accessible to a few domain experts, range sensors are nowadays ubiquitous and can even be found in the latest generation of mobile devices. Current sensors have a small form factor, are affordable, and accessible for everyday use to a broad audience. The availability of cheap range sensing technology led to a big leap in research [10], especially in the context of more powerful static and dynamic reconstruction techniques, starting from 3D scanning applications, such as KinectFusion, to highly accurate face and body tracking approaches. One very recent example is the current Apple iPhone X that employs the range data captured by an off-the-shelf depth sensor as part of its face identification system.

In the following, we review the technical foundations of such camera systems. We will start by reviewing the *Pinhole Camera* model and perspective projections. Afterward, we will introduce the ideas behind both *passive* as well as *active* depth sensing approaches and explain their fundamental working principles. More specifically, we will discuss how commodity RGB-D sensors based on *Stereo Vision* (SV), *Structured Light* (SL), and *Time of Flight* (ToF) technology work. We conclude by comparing the different depth sensing modalities and discussing their advantages and disadvantages.

1.2 Projective Camera Geometry

We start by reviewing the *Pinhole Camera* model, which is a simplified version of the projective geometry of real-world cameras, since it is a basic building block for many types of depth sensors. An illustration of the perspective projection defined by the *Pinhole Camera* model can be found in Fig. 1.2. A 3D point $\mathbf{v} = (\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z)^T \in \mathbb{R}^3$ in camera space is mapped to the sensor plane (green) based on a perspective projection [6]. The resulting point $\mathbf{p} = (\mathbf{p}_x, \mathbf{p}_y)^T \in \mathbb{R}^2$ on the sensor depends on the intrinsic properties of the camera, i.e., its focal length f and the principal point $\mathbf{c} = (\mathbf{c}_x, \mathbf{c}_y)^T$. Let us first assume that the principal point is at the center of the sensor plane, i.e., $\mathbf{c} = (0, 0)^T$. In the following, we show how to compute the 2D position $\mathbf{p}$ on the image plane given a 3D point $\mathbf{x}$ and the intrinsic camera parameters. By applying the geometric rule of equal triangles, the following relation can be obtained, see also Fig. 1.2 for an illustration:

$$\frac{\mathbf{p}_x}{f} = \frac{\mathbf{v}_x}{\mathbf{v}_z} \ . \tag{1.1}$$

With the same reasoning, a similar relation also holds for the y-component. Reordering and solving for $\mathbf{p}$ leads to the fundamental equations of perspective projection that describe how a 3D point $\mathbf{v}$ is projected to the sensor plane:

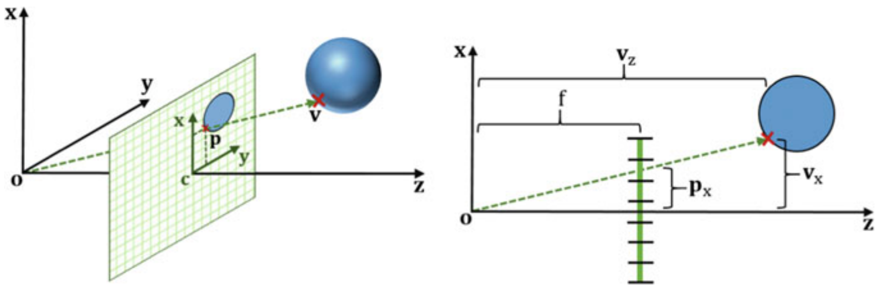
$$\mathbf{p}_x = \frac{f \cdot \mathbf{v}_x}{\mathbf{v}_z} \ , \tag{1.2}$$

$$\mathbf{p}_y = \frac{f \cdot \mathbf{v}_y}{\mathbf{v}_z} \ . \tag{1.3}$$

The same mapping can be more concisely represented in matrix-vector notation by using homogeneous coordinates. Let $\mathbf{K}$ be the intrinsic camera matrix:

$$\mathbf{K} = \begin{bmatrix} f & s & \mathbf{c}_x \\ 0 & f & \mathbf{c}_y \\ 0 & 0 & 1 \end{bmatrix} \ . \tag{1.4}$$

Here, s is an additional skew parameter [7] and $\mathbf{c}$ specifies the principal point, which we assumed to be zero so far. Given the definition of $\mathbf{K}$, the perspective projection can be represented as $\hat{\mathbf{p}} = \mathbf{K}\mathbf{v}$, where $\hat{\mathbf{p}} \in \mathbb{R}^3$ are the homogeneous coordinates of the 2D point $\mathbf{p}$. The intrinsic camera parameters can be obtained based on camera



(a) 3D View

(b) Side View

Fig. 1.2 Perspective camera geometry. The image sensor is shown in green. The *Pinhole Camera* model describes how a point $\mathbf{v} \in \mathbb{R}^3$ is mapped to a location $\mathbf{p} \in \mathbb{R}^2$ on the sensor. The **z**-axis is the cameras viewing direction and the **x**-axis is the up-vector. The perspective projection is defined by the camera's focal length f and the principal point **c**. The focal length f is the distance between the sensor plane and the origin **o** of the camera coordinate system

calibration routines [3, 18]. The *Pinhole Camera* model is one of the basic building blocks of range sensing approaches. It makes a few simplifying assumptions, such as that the lens is perfect, i.e., that there are no lens distortions. Lens distortion [16] can be tackled in a preprocessing step by calibrating the camera.

1.3 Passive Range Sensing

Similar to human 3D vision, passive range sensing is implemented based on the input of two or multiple [15] conventional monochrome or color cameras. Here, the term "passive" refers to the fact that passive sensors do not modify the scene to obtain the scene depth. The special case of obtaining depth measurements based on only two cameras [17] is known as *stereo* or *binocular reconstruction*. These systems are quite cheap and have a low-power consumption, since they are based on two normal color cameras. The basic setup of such a stereo camera system is illustrated in Fig. 1.3.

Scene depth can be estimated based on a computational process called *triangulation*. The first step in the estimation of scene depth is finding correspondences between the two camera views, i.e., pixels in the two images that observe the same 3D position in the scene. From these two corresponding points, the 3D position of the point that gave rise to these two observations can be computed via triangulation, i.e., by intersecting two rays cast through the detected point correspondences. Finding corresponding points between two different camera views is, in general, a highly challenging problem. Normally, the search is based on local color descriptor matching or on solving an optimization problem. One way to simplify this search is by exploiting the epipolar geometry between the two camera views. This reduces the 2D search problem to a 1D search along a line. Still, solving the correspondence problem requires sufficient local intensity and color variation in the recorded images,

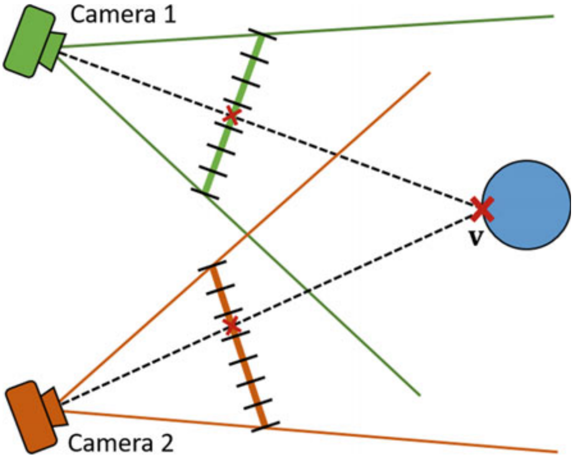


Fig. 1.3 Stereo reconstruction. Similar to human vision, stereo approaches employ two cameras to obtain observations of the scene from two slightly different viewpoints. In the first step of stereo reconstruction, the corresponding points in both images are computed, i.e., pixels of the images that observe the same 3D point in the scene. Based on these matches, the 3D position can be found via triangulation, i.e., by intersecting two rays cast through the detected point correspondences

i.e., enough features. Therefore, passive stereo reconstruction techniques work well in highly textured regions of the scene, but the search for correspondences might fail in featureless regions, which can result in missing depth information. Active depth sensing approaches aim at alleviating this problem.

1.4 Active Range Sensing

Besides passive range sensing approaches, such as the stereo cameras discussed in the last section, there are also active techniques for range sensing. Here, the term "active" refers to the fact that these sensors actively modify the scene to simplify the reconstruction problem. There are two classes of active approaches [13], which are based on different working principles, the so-called *Time of Flight* (ToF) and *Structured Light* (SL) cameras. Structured Light cameras project a unique pattern into the scene to add additional features for matching and thus simplify feature matching and depth computation. Therefore, they address the challenges passive reconstruction approaches face with featureless regions in the scene. On the other hand, Time of Flight cameras emit a (potentially modulated) light pulse and measure its round trip time or phase shift. Since Time-of-Flight cameras do not rely on color or texture to measure distance, they also do not struggle with texture-less scenes. In both of the cases, modern commodity sensors normally work in the *infrared* (IR) domain to not interfere with human vision and enable the simultaneous capture of scene appearance. In the following, we discuss both of these technologies in more detail and highlight their advantages and disadvantages.

1.4.1 Time-of-Flight Sensors

Besides passive binocular vision, many animals have implemented active range sensing approaches, e.g., the sonar used by whales is based on measuring the round trip time of a sound wave. As the name already suggests, the basic working principle of a Time-of-Flight camera is based on measuring the time of flight of an emitted light pulse [5]. More specifically, a light pulse is sent out from an emitter, it then traverses the scene until it hits an object and is reflected back to the Time-of-Flight camera, where a sensor records its arrival. In general, there are two different types of Time-of-Flight cameras.

The first class, *Pulsed Time-of-Flight cameras*, measures the round trip time of a light pulse based on rapid shutters and a clock. For Pulsed Time-of-Flight cameras, due to the constant known speed of light, the round trip distance can be computed by measuring the delay between sending and receiving the light pulse. The scene depth can then be computed as half of the measured round trip distance:

$$\text{Depth} = \frac{\text{Speed of Light} \times \text{Round Trip Time}}{2} \,. \tag{1.5}$$

There are two types of pulsed Time-of-Flight cameras. Point-wise Time-of-Flight sensors use a pan-tilt mechanism to obtain a time sequence of point measurements. This technique is also known as *Light Detection And Ranging* (LiDAR). Matrix-based Time-of-Flight cameras estimate a complete depth image for every time step based on a CMOS or CCD image sensor. They employ light pulses generated by a laser that are a few nanoseconds apart. Current commodity sensors belong to the second category, while Light Detection And Ranging is more employed for long-range outdoor sensing, e.g., in the context of self-driving cars. Due to the immensely high speed of light of approximately 300,000 km per second, the used clock for measuring the travel time has to be highly accurate, otherwise the depth measurements are imprecise.

The second type of Time-of-Flight camera uses a time-modulated light pulse and measures the phase shift between the emitted and returning pulse. For *Modulated Time-of-Flight* cameras, the light pulse is normally modulated by a continuous wave. A phase detector is used to estimate the phase of the returning light pulse. Afterward, the scene depth is obtained by the correlation between phase shift and scene depth. Multi-frequency techniques can be employed to further improve the accuracy of the obtained depth measurements and the effective sensing range of the cameras. Examples of current commodity Time-of-Flight cameras that are based on modulated time of flight include the Microsoft Kinect One and the Creative Senz3D.

1.4.2 Structured Light Sensors

Structured light sensing, similar to stereo reconstruction, is based on triangulation. The key idea is to replace one of the two cameras in a stereo system by a projector. The projector can be interpreted as an inverse camera. By projecting a known unique structured pattern [14] into the scene, additional artificial features are introduced into the scene. This drastically simplifies correspondence matching, thus the quality of the reconstruction does not depend on the amount of natural color features in the scene. Some sensors, such as the Microsoft Kinect, project a unique dot pattern [4], others project a temporal sequence of black and white stripes. Structured Light cameras are widespread and often used in research. The commodity sensors of this category normally work in the infrared domain to not interfere with human vision and enable the simultaneous capture of an additional color image. Examples of commodity sensors based on this technology are the Microsoft Kinect, Primesense Carmine, Asus Xtion Pro, and Intel Realsense. Actually, the Intel Realsense is a hybrid of a passive and active sensing approach. One problem of structured light cameras is that the sun's infrared radiation can saturate the sensor, making the pattern indiscernible. This results in missing depth information. The Intel Realsense alleviates this problem by combining active and passive vision. To this end, it combines two infrared cameras with one infrared projector that is used to add additional features to the scene. If the projector is overpowered by the ambient scene illumination the Intel Realsense defaults to standard stereo matching between two captured infrared images. Normal working ranges for such commodity sensors are between 0.5 and 12 m. Similar to stereo systems, the accuracy of such sensors directly depends on the distance to the scene, i.e., the accuracy degrades with increasing distance. The captured depth and color images of RGB-D sensors are not aligned, since the infrared and the color sensor are at different spatial locations, but the depth map can be mapped to the color image if the position and orientation of the two sensors is known.

1.5 Comparison of the Sensing Technologies

So far, we have discussed the most prevalent technologies for obtaining depth measurements. More specifically, we had a look at passive stereo reconstruction and active structured light as well as time-of-flight sensing. These three types of approaches are based on different physical and computational principles and thus have different advantages and disadvantages. For example, they have differing working ranges and noise characteristics. It is important to understand the advantages and disadvantages of the different technologies to be able to pick the right sensor for the application one wants to build. In the following, we compare the discussed three technologies in detail.

1.5.1 Passive Stereo Sensing

Stereo reconstruction is based on finding correspondences between points observed in both camera views and triangulation to obtain the depth measurements. Thus, the quality and density of the depth map directly depends on the amount of color and texture features in the scene. More specifically, the quality and density of the depth measurements degrades with a decreasing amount of available features. One extreme case, that is often found in indoor scenes, are walls of uniform color, which can not be reconstructed, since no reliable matches between the left and right camera can be found. Similar to uniformly colored objects, also low light, e.g., scanning in a dark room, can heavily impact the ability to compute reliable matches. Repeated structures and symmetries in the scene can lead to wrong feature associations. In this case, multiple equally good matches exist and sophisticated pruning strategies and local smoothness assumptions are required to select the correct match. Passive stereo is a triangulation-based technique. Therefore, it requires a baseline between the two cameras, which leads to a larger form factor of the device. Similar to all approaches based on triangulation, the quality of the depth measurements degrades with increasing distance to the scene and improves for larger baselines. The noise characteristics of stereo vision systems have been extensively studied [2]. One significant advantage of passive stereo systems is that multiple devices do not interfere with each other. This is in contrast to most active sensing technologies. In addition, stereo sensing can have a large working range if a sufficiently large baseline between the two cameras is used. Since stereo systems are built from off-the-shelf monochrome or color cameras, they are cheap to build and are quite energy efficient. One great use case for passive stereo sensing is outdoor 3D scene reconstruction.

1.5.2 Structured Light Sensing

Active range sensing techniques, such as structured light sensing, remove one of the fundamental problems of passive approaches, i.e., the assumption that the scene naturally contains a large amount of color or texture features. This is made possible, since the projected pattern introduces additional features into the scene which can be used for feature matching. For example, this allows to reconstruct even completely uniformly colored objects, but comes at the price of a higher energy consumption of the sensor, since the scene has to be actively illuminated. In addition, structured light sensors do not work under strong sunlight, since the sensor will be oversaturated by the sun's strong IR radiation and thus the projected pattern is not visible. Due to the projection of a structured pattern, a few problems might occur: If the projected pattern is partially occluded from the sensor's viewpoint, which is especially a problem at depth discontinuities in the scene, the depth cannot be reliably computed. Normally, this leads to missing depth estimates around the object silhouette, which leads to a slightly "shrunken" reconstruction. This also complicates the reconstruction of thin objects. The projected pattern might also be absorbed by dark objects, reflected by specular objects, or refracted by transparent objects, all of these situations might lead to wrong or missing depth estimates. Active structured light depth sensing technology has a limited working range, normally up to 15 m, since otherwise too much energy would be required to consistently illuminate the scene. The noise characteristics of structured light sensors have been extensively studied [11, 12]. Using multiple sensors at the same time might result in a loss of depth accuracy due to interference of multiple overlapping patterns, since the correspondences can not be reliably computed. Geometric structures that are smaller than the distance between the projected points are lost. One great use case for structured light sensing is the face identification system of the current Apple iPhone X.

1.5.3 Time-of-Flight Sensing

In contrast to stereo vision and structured light, Time-of-Flight cameras are based on a different physical measurement principle, i.e., measuring time of flight/phase shift of a light pulse instead of triangulation. This leads to a different set of failure modes and drastically different noise characteristics. One of the biggest artifacts in time-of-flight depth images are the so-called “flying pixels” at depth discontinuities. Flying pixels have depth values between the fore- and background values that exist in reality. They appear if the light pulse is reflected back by multiple parts of the scene and then measured at the same sensor location. This is related to the much wider class of multi-path interference effects ToF cameras suffer from, i.e, multiple indirect light paths being captured by the sensor. Examples of this are multi-path effects caused by materials that exhibit reflections or refractions, e.g., mirrors or glass. Even in relatively diffuse scenes, indirect bounces of the light pulse might influence the reconstruction quality. Dark materials do not reflect light. Therefore, no returning light pulse can be measured which leads to holes in the depth map. Similar to other active sensing modalities, Time of Flight suffers from interference between multiple sensors if they use the same phase shift. This can be alleviated by using different modulation frequencies for each sensor. Similar to active Structured Light, Time-of-Flight depth sensing struggles under strong sunlight. Since Time-of-Flight cameras require a certain integration time to obtain a good signal-to-noise ratio, fast motions lead to motion-blurred depth estimates. The noise characteristics of Time-of-Flight cameras have been extensively studied [1]. One great use case for time-of-flight sensors is body tracking in the living room to enable immersive gaming experiences.

1.6 Conclusion and Outlook

We had a detailed look into the different types of existing range sensors. All depth sensing techniques have their own advantages and disadvantages and it is important to pick the right sensor for the application one wants to build. In the future, higher resolution sensors and projectors will further help to increase the achievable quality of depth measurements. On the software side, deep learning techniques have the potential to further improve the captured depth data by learning depth denoising, upsampling, and super-resolution. This will lead to an even wider democratization of range sensing technology and many more compelling new use cases.

References

1. Belhedi A, Bartoli A, Bourgeois S, Gay-Bellile V, Hamrouni K, Sayd P (2015) Noise modelling in time-of-flight sensors with application to depth noise removal and uncertainty estimation in three-dimensional measurement. IET Comput Vis 9(6):967–977
2. Bier A, Luchowski L (2009) Error analysis of stereo calibration and reconstruction. In: Gagalowicz A, Philips W (eds) Computer vision/computer graphics collaboration techniques. Springer, Berlin, pp 230–241
3. Bradski G, Kaehler A (2013) Learning OpenCV: computer vision in C++ with the OpenCV Library, 2nd edn. O'Reilly Media, Inc, Sebastopol (2013)
4. Cruz L, Lucio D, Velho L (2012) Kinect and RGBD images: challenges and applications. In: Proceedings of the 2012 25th SIBGRAPI conference on graphics, patterns and images tutorials, SIBGRAPI-T '12, pp. 36–49. IEEE Computer Society, Washington, DC (2012). https://doi.org/10.1109/SIBGRAPI-T.2012.13
5. Foix S, Alenya G, Torras C (2011) Lock-in time-of-flight (ToF) cameras: a survey. IEEE Sens J 11(9):1917–1926. https://doi.org/10.1109/JSEN.2010.2101060
6. Forsyth DA, Ponce J (2002) Computer vision: a modern approach. Prentice Hall Professional Technical Reference (2002)
7. Hartley R, Zisserman A (2000) Multiple view geometry in computer vision. Cambridge University Press, New York
8. Huhle B, Jenke P, Straßer W (2008) On-the-fly scene acquisition with a handy multi-sensor system. IJISTA 5(3/4):255–263
9. Lindner M, Kolb A, Hartmann K (2007) Data-fusion of PMD-based distance-information and high-resolution RGB-images. In: International symposium on signals, circuits and systems, vol 1, pp 1–4. https://doi.org/10.1109/ISSCS.2007.4292666
10. Magnor M, Grau O, Sorkine-Hornung O, Theobalt C (eds.) (2015) Digital representations of the real world: how to capture, model, and render visual reality. A K Peters/CRC Press, Massachusetts
11. Mallick T, Das PP, Majumdar AK (2014) Characterizations of noise in kinect depth images: a review. IEEE Sens J 14(6):1731–1740. https://doi.org/10.1109/JSEN.2014.2309987
12. Nguyen CV, Izadi S, Lovell D (2012) Modeling kinect sensor noise for improved 3D reconstruction and tracking. In: Proceedings of the 2012 second international conference on 3D imaging, modeling, processing, visualization and transmission, 3DIMPVT '12, pp. 524–530. IEEE Computer Society, Washington, DC (2012). https://doi.org/10.1109/3DIMPVT.2012.84
13. Sarbolandi H, Lefloch D, Kolb A (2015) Kinect range sensing: structured-light versus time-of-flight kinect. Comput Vis Image Underst 139:1–20. https://doi.org/10.1016/j.cviu.2015.05.006

14. Saty TP, Gupta RK (2007) Model and algorithms for point cloud construction using digital projection patterns
15. Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of the 2006 IEEE computer society conference on computer vision and pattern recognition, CVPR '06, vol 1, pp 519–528. IEEE Computer Society, Washington, DC (2006). https://doi.org/10.1109/CVPR.2006.19
16. Sturm P, Ramalingam S, Tardif JP, Gasparini S, Barreto JA (2011) Camera models and fundamental concepts used in geometric computer vision. Found Trends Comput Graph Vis **6**(1-2), 1–183. https://doi.org/10.1561/0600000023
17. Tippetts B, Lee DJ, Lillywhite K, Archibald J (2016) Review of stereo vision algorithms and their suitability for resource-limited systems. J Real-Time Image Process 11(1):5–25. https://doi.org/10.1007/s11554-012-0313-2
18. Zhang Z (2000) A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 22(11):1330–1334. https://doi.org/10.1109/34.888718
19. Zollhöfer M, Nießner M, Izadi S, Rehmann C, Zach C, Fisher M, Wu C, Fitzgibbon A, Loop C, Theobalt C, Stamminger M (2014) Real-time non-rigid reconstruction using an RGB-D camera. ACM Trans Graph **33**(4), 156:1–156:12. https://doi.org/10.1145/2601097.2601165

Chapter 2
Dealing with Missing Depth: Recent Advances in Depth Image Completion and Estimation

Amir Atapour-Abarghouei and Toby P. Breckon

Abstract Even though obtaining 3D information has received significant attention in scene capture systems in recent years, there are currently numerous challenges within scene depth estimation which is one of the fundamental parts of any 3D vision system focusing on RGB-D images. This has lead to the creation of an area of research where the goal is to complete the missing 3D information post capture. In many downstream applications, incomplete scene depth is of limited value, and thus, techniques are required to *fill the holes* that exist in terms of both missing depth and colour scene information. An analogous problem exists within the scope of scene filling post object removal in the same context. Although considerable research has resulted in notable progress in the synthetic expansion or reconstruction of missing colour scene information in both statistical and structural forms, work on the plausible completion of missing scene depth is contrastingly limited. Furthermore, recent advances in machine learning using deep neural networks have enabled complete depth estimation in a monocular or stereo framework circumnavigating the need for any completion post-processing, hence increasing both efficiency and functionality. In this chapter, a brief overview of the advances in the state-of-the-art approaches within RGB-D completion is presented while noting related solutions in the space of traditional texture synthesis and colour image completion for hole filling. Recent advances in employing learning-based techniques for this and related depth estimation tasks are also explored and presented.

A. Atapour-Abarghouei (✉)
Department of Computer Science, Durham University, Durham, UK
e-mail: amir.atapour-abarghouei@durham.ac.uk

T. P. Breckon
Departments of Engineering & Computer Science, Durham University, Durham, UK
e-mail: toby.breckon@durham.ac.uk

P. L. Rosin et al. (eds.), *RGB-D Image Analysis and Processing*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-28603-3_2

2.1 Introduction

Three-dimensional scene understanding has received increasing attention within the research community in recent years due to its ever-growing applicability and widespread use in real-world scenarios such as security systems, manufacturing and future vehicle autonomy. As mentioned in Chap. 1, a number of limitations pertaining to environmental conditions, inter-object occlusion and sensor capabilities still remain despite the extensive recent work and many promising accomplishments of 3D sensing technologies [33, 134, 149, 158]. It is due to these challenges that a novel area of research has emerged mostly focusing on refining and completing missing scene depth to increase the quality of the depth information for better downstream applicability.

Although traditional RGB image inpainting and texture synthesis approaches have been previously utilized to address scene depth completion [7, 39, 64], challenges regarding efficiency, depth continuity, surface relief and local feature preservation have hindered flawless operation against high expectations of plausibility and accuracy in 3D images [4]. In this vein, this chapter provides a brief overview of the recent advances in scene depth completion, covering commonly used approaches designed to refine depth images acquired through imperfect means.

Moreover, recent progress in the area of monocular depth estimation [6, 44, 55, 152] has lead to a cheap and innovative alternative to completely replace other more expensive and performance-limited depth-sensing approaches such as stereo correspondence [129], structure from motion [27, 41] and depth from shading and light diffusion [1, 132] among others. Apart from computationally intensive demands and careful calibration requirements, these conventional depth-sensing techniques suffer from a variety of quality issues including depth inhomogeneity, missing or invalid values and alike, which is why the need for depth completion and refinement in post-processing arises in the first place.

As a result, generating complete scene depth from a single image using a learning-based approach can be of significant value. Consequently, a small portion of this chapter is dedicated to covering the state-of-the-art monocular depth estimation techniques capable of producing complete depth which would eliminate any need for depth completion or refinement.

2.2 Missing Depth

As explained in the previous chapter, different depth-sensing approaches can lead to various issues within the acquired scene depth, which in turn make depth completion and refinement an important post-processing step.

Passive scene-sensing approaches such as stereo correspondence [129] have long been established as a reliable method of dense depth acquisition. Although stereo imaging is well equipped to estimate depth where highly granular texture is present, even the smallest of issues in calibration and synchronization can lead to noisy, invalid

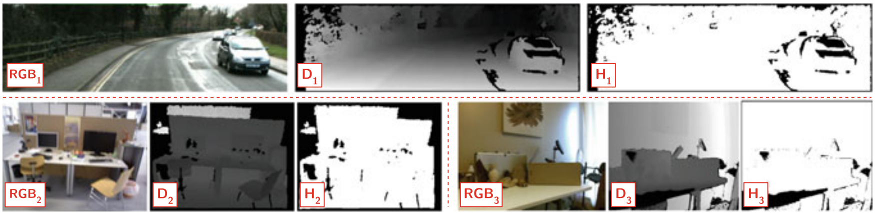


Fig. 2.1 Examples of depth acquired via stereo correspondence (top), structured light device (bottom left) and time-of-flight camera (bottom right). **RGB:** colour image; **D:** depth image; **H:** hole mask indication missing depth values

or missing depth values. Additionally, missing values are prevalent in sections of the scene that contain occluded regions (i.e. groups of pixels that are seen in one image but not the other), featureless surfaces, sparse information for a scene object such as shrubbery, unclear object boundaries, very distant objects and alike. Such issues can be seen in Fig. 2.1 (top), wherein the binary mask marks where the missing depth values are in a disparity image calculated via a stereo correspondence algorithm [65].

On the other hand, consumer devices such as structured light and time-of-flight cameras are active range sensors that are more widely utilized for a variety of purposes due to their low cost and wide availability in the commercial market with factory calibration settings [14, 23, 46].

However, due to a number of shortcomings such as external illumination interference [23], ambient light saturation [46], inaccurate light pattern detection in the presence of motion [125] and active light path error caused by reflective surfaces or occlusion [126], consumer structured light devices can result in missing depth or noisy values that are best handled by removal and subsequent filling. An example of such a depth image and its missing values can be seen in Fig. 2.1 (bottom left). Time-of-flight cameras can also suffer from complications detrimental to output deployment due to issues such as external illumination interference [123], light scattering caused by semi-transparent surfaces [59, 72] and depth offset for non-reflective objects [96]. Such issues are exemplified in Fig. 2.1 (bottom right).

Completing depth images, captured through these active or passive depth-sensing technologies, can lead to significant performance boost in any 3D vision application even though many current systems simply cope with challenges created by noisy and incomplete depth images without any post-processing. In the next section, we will focus on various approaches to the problem of image completion in the context of RGB-D imagery.

2.3 RGB-D Completion

While object removal, inpainting and surface completion [2, 15, 17–20, 36, 43, 133] has been a long-standing problem addressed within the literature in the past few decades, depth completion is a relatively new area of research with its own challenges and limitations. However, scene depth is still represented and processed in the form of images, and some researchers still directly apply classical RGB image inpainting methods to depth images or use depth completion approaches heavily inspired by RGB completion techniques. Consequently, an overview of image inpainting within the context of scene colour image (RGB) can be beneficial for a better understanding of the multi-facet subject of depth filling. In the following section, relevant image inpainting methods are briefly discussed before moving on to a more detailed description of the depth completion literature.

2.3.1 RGB Image Inpainting

Inpainting deals with the issue of a plausibly completing a target region within the image often created as a result of removing a certain portion of the scene. Early image inpainting approaches attempted to smoothly propagate the isophotes (lines within the image with similar intensity values) into this target area. However, most of these approaches [15, 133] tend to ignore an important aspect significant to an observer's sense of plausibility, which is the high-frequency spatial component of the image or texture. Consequently, later inpainting techniques began to incorporate ideas from the field of texture synthesis (in which the objective is to generate a large texture region given a smaller sample of texture without visible artefacts of repetition within the larger region [42, 43, 118]) into the inpainting process to compensate for the lack of texture commonly found in the target region post completion [2, 36, 79] (exemplar-based inpainting).

In one of the most seminal works on image inpainting [15], the problem is addressed using higher order partial differential equations and anisotropic diffusion to propagate pixel values along isophote directions (Fig. 2.2). The approach demonstrated remarkable progress in the area at the time but more importantly, it contained a set of guidelines for image inpainting created after extensive consultations with scene composition experts, which have now standardized the functionalities of an inpainting algorithm. These remain highly relevant even in depth completion:

- **1:** Upon completion of the inpainting process, the target region must be consistent with the known region of the image to preserve global continuity.
- **2:** The structures present within the known region must be propagated and linked into the target region.

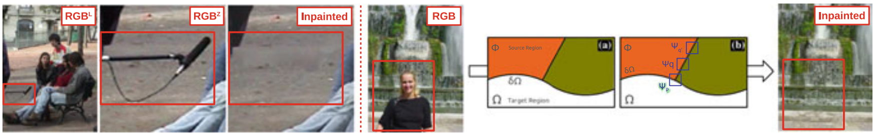


Fig. 2.2 **Left:** results of [15]. The foreground microphone has been removed and inpainted, but the texture is not accurate, leading to a perception of blurring. **Right:** an example of the results and process of exemplar-based inpainting [36]

- **3:** The structures formed within the target region must be filled with colours consistent with the known region.
- **4:** Texture must be added into the target region after or during the inpainting process.

Improved inpainting approaches were subsequently proposed employing a variety of solutions including the fast marching method [133], total variational (TV) models [28, 121], and exemplar-based techniques [16, 36]. In one such approach, the authors of [36] follow traditional exemplar-based texture synthesis methods [43] by prioritizing the order of filling based on the strength of the gradient along the target region boundary. Although the authors of [36] are not the first to carry out inpainting via exemplar-based synthesis [16], previous approaches are all lacking in either structure propagation or defining a suitable filling order that could prevent the introduction of blurring or distortion in shapes and structures. This exemplar-based method [36] is not only capable of handling two-dimensional texture but can plausibly propagate linear structures within the image. An example of the results of this method can be seen in Fig. 2.2 (right), in which water texture has been plausibly synthesized after the person is removed from the image. However, this approach cannot cope with curved structures and is heavily dependent on the existence of similar pixel neighbourhoods in the known region for plausible completion. Even though the approach relies on fine reflectance texture within the image to prioritize patches and can fail when dealing with large objects in more smooth depth images (Fig. 2.3—left), it has been a great step towards focusing on granular texture within the image completion literature.

Other image completion techniques have also been proposed that would address different challenges in the inpainting process. For instance, certain methods use schemes such as reformulating the problem as metric labelling [85], energy minimization [12, 140], Markov random field models with labels assigned to patches [83], models represented as an optimal graph labelling problem, where the shift-map (the relative shift of every pixel in the output from its source in the input) represents the selected label and is solved by graph cuts [119], and the use of *Laplacian pyramids* [91] instead of the gradient operator in a patch correspondence search framework due to the advantageous qualities of Laplacian pyramids, such as isotropy, rotation invariance and lighter computation. There have also been attempts to complete images in an exemplar-based framework using external databases of semantically similar images [60, 141] (Fig. 2.3—right).

Deep neural networks have recently revolutionized the state of the art in many computer vision tasks such as image stylization [52, 54, 76, 80], super-resolution [111, 138] and colourization [156]. Image completion has also seen its fair share of progress using such techniques. In [113], an approach is proposed that is capable of predicting missing regions in an RGB image via adversarial training of a generative model [56]. In a related work, the authors of [150] utilize an analogous framework with similar loss functions to map the input image with missing or corrupted regions to a latent vector, which in turn is passed through their generator network that recovers the target content. The approach in [146] proposes a joint optimization framework

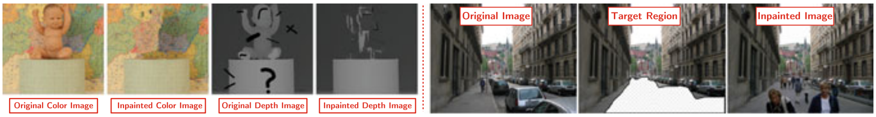


Fig. 2.3 **Left:** results of exemplar-based inpainting [36] applied to RGB and depth images. Note that the objective is to remove the object (baby) from both the RGB and depth images and to fill the already existing holes (pre-removal) in the depth image. The approach is significantly more effective when applied to colour images. **Right:** result of exemplar-based inpainting using an external database [60]

composed of two separate networks, a content encoder, based on [113], which is tasked to preserve contextual structures within the image, and a texture network, which enforces similarity of the fine texture within and without the target region using neural patches [95]. The model is capable of completing higher resolution images than [113, 150] but at the cost of greater inference time since the final output is not achievable via a single forward pass through the network.

More recently, significantly better results have been achieved using [73], which improves on the model in [113] by introducing global and local discriminators as adversarial loss components. The global discriminator assesses whether the completed image is coherent as a whole, while the local discriminator concentrates on small areas within the target region to enforce local consistency. Similarly, the authors of [151] train a fully convolutional neural network capable of not only synthesizing geometric image structures but also explicitly using image features surrounding the target region as reference during training to make better predictions.

While these learning approaches are highly capable of generating perceptually plausible outputs despite the significant corruption applied to the input, when it comes to depth, they are incapable of producing high-quality outputs due in part to the significantly higher number of target regions (holes) both large and small over the smoother surfaces in depth images. Examples of these novel approaches applied to depth images can be seen in Fig. 2.4, which indicates how ineffective learning-based RGB image inpainting approaches can be within the depth modality [4].

While RGB completion techniques in various forms have previously been used with or without modifications [100, 144, 154] to complete depth images, significant differences between RGB and depth images prevent a successful deployment of RGB inpainting techniques to perform depth completion. For instance, the lack of

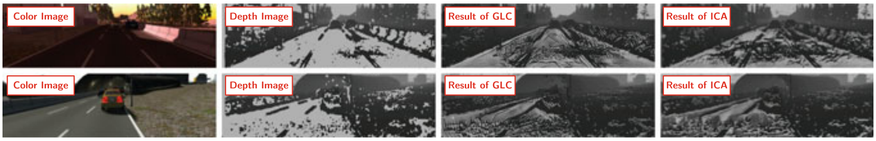


Fig. 2.4 Results of global and local completion (GLC) [73] compared to inpainting with contextual attention (ICA) ([151]) applied to depth images