Yu-Jin Zhang

3D Computer Vision

Foundations and Advanced Methodologies

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn
PHEI

Springer

3D Computer Vision

Yu-Jin Zhang

3D Computer Vision

Foundations and Advanced Methodologies

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Springer

Yu-Jin Zhang
Department of Electronic Engineering
Tsinghua University
Beijing, China

ISBN 978-981-19-7602-5 ISBN 978-981-19-7603-2 (eBook)
https://doi.org/10.1007/978-981-19-7603-2

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Computer vision is an information discipline that uses computers and electronic devices to realize human visual functions. The original purpose of computer vision research is to identify and locate the objects in it with the help of the images related to the scene, determine the structure of the object and the relationship between the objects, so that the objects and scenes in the objective world can be meaningfully explained and judged.

The research and application of computer vision has a history of more than half a century. In recent years, with the introduction of technologies such as artificial intelligence and deep learning, related theories and methods have developed rapidly, and the application fields have been expanding.

The content of this book basically covers the main aspects of 3D computer vision. In addition to the introduction of basic concepts and related disciplines in the computer vision overview chapter, the main content of the book is divided into 10 chapters, which respectively introduce 10 types of computer vision technologies. They are: camera imaging and calibration technology, depth image acquisition technology, 3D point cloud data acquisition and processing technology, binocular stereo vision technology, multi-ocular stereo vision technology, monocular multi-image scene restoration technology, monocular single-image scene restoration technology, generalized matching technology, simultaneous location and mapping technology, and spatio-temporal behavior understanding technology.

This book focuses on the fundamentals and recent advances in computer vision. In the 10 chapters introducing computer vision technology, the basic concepts and basic principles of the technology are first described, and the typical methods for implementing the technology are analyzed in detail (including algorithm description, specific steps, effect examples, etc.), and some new developments in the field of technology are introduced, summarized, and classified, which can help readers understand the latest development trends.

On the one hand, this book can be used as a professional course material for senior undergraduates and graduate students in related disciplines, helping them master the basic principles, carry out scientific research activities, and complete graduation projects and dissertations; on the other hand, this book is also suitable for the company's research and development personnel to understand the latest progress information and serves as a scientific research reference.

This book has approximately 500,000 words. It has 11 chapters, which consist of a total of 59 sections (second-level headings) and 156 subsections (third-level headings). There are 215 numbered figures, 37 numbered tables, and 660 numbered equations. Finally, the list of more than 300 references cited (there were over 100 in the 2020s), and more than 500 subject terms used for indexing are listed at the end of the book to facilitate further access to related literatures.

Finally, I would like to thank my wife Yun He, daughter Heming Zhang, and other family members for their understanding and support in all aspects.

Beijing, China Yu-Jin Zhang

Contents

About the Author

Yu-Jin Zhang Yu-Jin Zhang received his doctor's degree in Applied Science from the University of Liège in Belgium in 1989. From 1989 to 1993, he successively engaged in postdoctoral research and served as a full-time researcher at Delft University in the Netherlands. Since 1993, he has worked in the Department of Electronic Engineering of Tsinghua University in Beijing, China. He has been a professor since 1997, a Ph.D. supervisor since 1998, and a tenured-professor since 2014. During the sabbatical year of 2003, he was a visiting professor of Nanyang Technological University in Singapore.

At Tsinghua University, he has offered and taught over 10 undergraduate and graduate courses, including "Image Processing," "Image Analysis," "Image Understanding," and "Content Based Visual Information Retrieval." At Nanyang Technological University, he offered and taught a postgraduate course: "Advanced Image Analysis (English)." More than 30 Chinese and English textbooks have been written and published (with a total of over 300,000 printed copies). More than 30 teaching research papers have been published both domestically and internationally.

The main scientific research fields are image engineering (image processing, image analysis, image understanding, and their technical applications) and related disciplines that it actively advocates. He has published over 500 research papers on image engineering both domestically and internationally. He published the monographs *Image Segmentation* and *Content Based Visual Information Retrieval* (Science Press, China), *Subspace Based Face Recognition* (Tsinghua University Press, China); has written *English Chinese Dictionary of Image Engineering* (1st, 2nd, and 3rd editions; Tsinghua University Press, China); has written *Selected Works of Image Engineering Technology* and *Selected Works of Image Engineering Technology (2)* (Tsinghua University Press, China); has translated *Computational Geometry, Topology, and Physics of Digital Images with Applications* (Springer Verlag, Germany) into Chinese; has led the compilation of *Advances in Image and Video Segmentation* and *Semantic Based Visual Information Retrieval* (IRM Press, USA), as well as *Advances in Face Image Analysis: Technologies and Technologies* (IGI Global, USA); has written *Handbook of Image Engineering* (Springer Nature, Singapore); has written *A Selection of Image Processing Techniques*, *A Selection of*

Image Analysis Techniques, and *A Selection of Image Understanding Techniques* (CRC Press, USA).

He is currently a fellow member and honorary supervisor of the Chinese Society of Image Graphics; a fellow member of the International Society of Optical Engineering (SPIE), for achievements in image engineering; and formerly served as the Chairman of the Program Committee of the 24th International Conference on Image Processing (ICIP 2017).

Chapter 1
Introduction

Vision is an important function and means for human beings to observe and recognize the world. Computer vision, as a subject using computer to realize human visual function, has not only received great attention and in-depth research but also been widely used [1].

Visual process can be seen as a complex process from sensing (feeling the image of the objective world) to perception (understanding the objective world from the image). This involves the knowledge of optics, geometry, chemistry, physiology, psychology, and so on. To complete such a process, computer vision should not only have its own corresponding theory and technology but also combine the achievements of various disciplines and the development of various engineering technologies.

The sections of this chapter are arranged as follows.

Section 1.1 gives a general introduction to computer vision, including the key points of human vision, the research methods and objectives of computer vision, and the connections and differences with several major related disciplines.

Section 1.2 introduces the theory and framework of computer vision, mainly including the important visual computational theory and its existing problems and improvements, and also discusses some situations of other theoretical frameworks.

Section 1.3 introduces the overview of various image processing technologies that are the basis of computer vision technology. Under the overall framework of image engineering, three levels of image technology, as well as various recent research directions and application fields, are discussed in detail.

Section 1.4 provides an overview of deep learning methods that have rapidly promoted the development of computer vision technology in recent years. In addition to listing some basic concepts of convolutional neural networks, it also discusses the core technology of deep learning and its connection with computer vision.

Section 1.5 introduces the main content involved in the book and its organizational structure.

Y.-J. Zhang, *3D Computer Vision*, https://doi.org/10.1007/978-981-19-7603-2_1

1.1 Introduction to Computer Vision

The following is a general introduction to the origin, objectives, and related disciplines of computer vision.

1.1.1 Visual Essentials

Computer vision originates from human vision, which is generally called vision. Vision is a kind of human function, which plays an important role in human observation and cognition of the objective world. According to statistics, about 75% of the information obtained by humans from the outside world comes from the visual system, which not only shows that the amount of visual information is huge but also shows that humans have a high utilization rate of visual information. Human visual process can be seen as a complex process from sensing (feeling the image obtained by 2D projection of 3D world) to perception (recognizing the content and meaning of 3D world from 2D image).

Vision is a very familiar function that not only helps people obtain information but also helps people process it. Vision can be further divided into two levels: visual sensation and visual perception. Here, sensation is at a lower level, which mainly receives external stimuli, while perception is at a higher level, which converts external stimuli into meaningful content. In general, sensation receives external stimuli almost indiscriminately and completely, while perception determines which parts of the external stimulus could combine together to form the "object" of interest.

Visual sensation mainly explains the basic properties (such as brightness, color, etc.) of people's response to light (i.e., visible radiation) from the molecular level and point of view, and it mainly involves physics, chemistry, and other disciplines. The main research contents of visual sensation are (1) the physical properties of light, such as light quantum, light wave, and spectrum, and (2) the degree to which light stimulates the visual receptors, such as photometry, eye structure, visual adaptation, visual intensity and sensitivity, vision time, and space characteristics.

Visual perception mainly discusses how people respond to visual stimuli from the objective world and the way they respond. It studies how to make people form an interpretation of the spatial representation of the external world through vision, so it also has psychological factors. As a form of reflection on the current objective things, visual perception only relies on the principle of projection of light onto the retina to form a retinal image and the known mechanism of the eye or nervous system. It is difficult to explain the whole (perception) process clearly. Visual perception is a group of activities carried out in the nerve center, which organizes some scattered stimuli in the visual field to form a whole with a certain shape to understand the world. As early as 2000 years ago, Aristotle defined the task of visual perception as determining "What is where" [2].

In a narrow sense, the ultimate goal of vision is to make meaningful explanations and descriptions of objective scenes for the observer. In a broad sense, it also includes formulating behavior plans based on these explanations and descriptions, as well as according to the surrounding environment and the wishes of the observer, so as to act on the surrounding world. This is also actually the goal of computer vision.

1.1.2 The Goal of Computer Vision

Computer vision is the use of computers to realize human visual functions, that is, the sensation, perception, processing, and interpretation of three-dimensional scenes in the objective world. The original purpose of vision research is to grasp and understand the image of the scene; identify and locate the objects in it; determine their own structure, spatial arrangement, and distribution; and explain the relationship between objects. The research goal of computer vision is to make meaningful judgments about actual objects and scenes in the objective world based on perceived images [3].

There are two main research methods of computer vision at present: one is the method of bionics, which refers to the structural principle of the human visual system to establish corresponding processing modules to complete similar functions and tasks; the other is the method of engineering, which starts from analyzing the function of human visual process, and does not deliberately simulate the internal structure of the human visual system, but only considers the input and output of the system, and adopts any existing and feasible means to realize the function of the system. This book discusses the second approach primarily from an engineering point of view.

The main research goals of computer vision can be summarized into two: they are interrelated and complementary. The first research goal is to build computer vision systems to accomplish various vision tasks. In other words, it enables the computer to obtain images of the scene with the help of various visual sensors (such as CCD and CMOS camera devices), from which to perceive and recover the geometric properties, posture structure, motion, mutual position, etc. of objects in the 3D environment, and to identify, describe, and explain objective scenarios and then make judgments and decisions. The technical mechanism for accomplishing these tasks is mainly studied here. At present, the work in this area focuses on building various specialized systems to complete specialized visual tasks that appear in various practical occasions; in the long run, it is necessary to build a more general system (closer to the human visual system) to complete general vision tasks. The second research goal is to use this research as a means to explore the visual working mechanism of the human brain and to master and understand the visual working mechanism of the human brain (such as computational neuroscience). The main research here is the biological mechanism. For a long time, people have carried out a lot of research on the human brain visual system from the aspects of physiology, psychology, nerve, cognition, etc., but it is far from revealing all the mysteries of the visual process, especially the research and understanding of the visual mechanism is still far away. It lags still behind the research and mastery of visual information processing. It should be pointed out that a full understanding of human brain vision will also promote in-depth research in computer vision [2]. This book mainly considers the first research objective.

It can be seen from the above that computer vision uses computers to realize human visual functions, and its research has obtained many inspirations from human vision. Much important research in computer vision has been accomplished by understanding the human visual system; typical examples include the use of pyramids as an efficient data structure, the use of the concept of local orientation, the use of filtering techniques to detect motion, and the recent artificial neural network. In addition, with the help of understanding and research on the function of the human visual system, it can also help people develop new computer vision algorithms.

The research and application of computer vision has a long history. Overall, early computer vision systems mainly relied on 2D projected images of objective scenes in 3D. The research goal of computer vision was to improve the quality of images, so that users could obtain the information more clearly and conveniently, or focus on automatically obtaining various characteristic data in the image to help users analyze and recognize the scenery. This aspect of work can be attributed to 2D computer vision, which is currently relatively mature with many application products available. With the development of theory and technology, more and more research focuses on fully utilizing the 3D spatial information obtained from objective scenery (often combined with temporal information), automatically analyzing and understanding the objective world, so as to making judgments and decisions. This includes further obtaining depth information on the basis of 2D projection images to comprehensively grasp the 3D world. This area of work is still being explored and requires the introduction of technologies such as artificial intelligence, which is currently the focus of research in computer vision. The recently related work can be categorized under 3D computer vision and will be the concentration of this book.

1.1.3 Related Disciplines

As a discipline, computer vision is inextricably linked with many disciplines, especially with some related and similar disciplines. The following is a brief introduction to several disciplines that are closest to computer vision. The connections and differences between related disciplines and fields are shown in Fig. 1.1.

1. Machine vision or robot vision

   Machine vision or robot vision is inextricably linked with computer vision and is used as a synonym in many cases. Specifically, it is generally believed that computer vision focuses more on the theory and method of scene analysis and image interpretation, while machine vision focuses more on acquiring the image

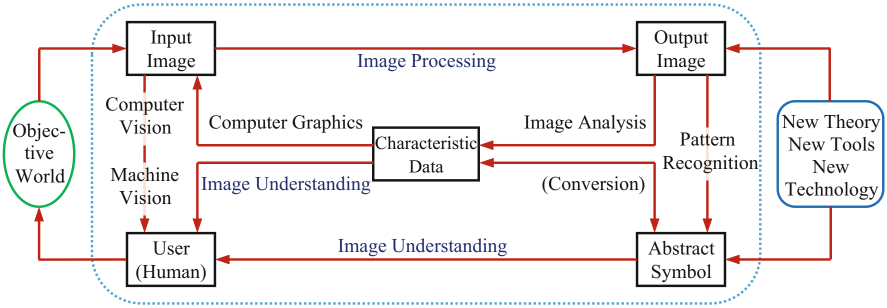

Fig. 1.1 The connections and differences between related disciplines and fields

of the environment through visual sensors, building a system with visual perception function, and realizing the algorithm for detecting and identifying objects. On the other hand, robot vision emphasizes more on the machine vision of robot, so that robot has the function of visual perception.

2. Computer graphics

Graphics refers to the science of expressing data and information in the form of graphics, charts, drawings, etc. Computer graphics studies how to use computer technology to generate these forms, and it is also closely related to computer vision. Computer graphics is generally referred to as the inverse problem of computer vision, because vision extracts 3D information from 2D images, while graphics use 3D models to generate 2D scene images (more generally based on nonimage forms of data description to generate realistic images). It should be noted that, compared with the many uncertainties in computer vision, computer graphics mostly deals with deterministic problems, which can be solved through mathematical methods. In many practical applications, people are more concerned with the speed and accuracy of graphics generation, that is, to achieve some kind of compromise between real time and fidelity.

3. Image engineering

Image engineering is a very rich discipline, including three levels (three subdisciplines) that are both related and different: image processing, image analysis, and image understanding, as well as their engineering applications.

Image processing emphasizes the conversion between images (image in and image out). Although image processing commonly refers to various image technologies, image processing in a narrow sense mainly focuses on the visual observation effect of the output image [4]. This includes making various processing adjustments to the image to improve the visual effect of the image and facilitate the subsequent high-level processing; or compress and encode the image to reduce the required storage space or transmission time on the basis of ensuring the required visual perception, so as to meet the requirements of a given transmission path; or add some additional information to the image without affecting the appearance of the original image; etc.

Image analysis is mainly to detect and measure the objects of interest in the image to obtain their objective information, thereby establishing the description of the objects in the image (image in and data out) [5]. If image processing is a process from image to image, image analysis is a process from image to data. The data here can be the result of the measurement of the object characteristics, or the symbolic representation based on the measurement, or the identification conclusion of the object category. They describe the characteristics and properties of objects in the image.

The focus of **image understanding** is to further study the nature of each object in the image and their mutual relations on the basis of image analysis and obtain an understanding of the meaning of the whole image content and an explanation of the original imaging objective scene, so that people can make judgments (know the world) and guide and plan actions (transform the world) [6]. If image analysis mainly focuses on the observer to study the objective world (mainly on observable things), image understanding, to a certain extent, focuses on the objective world, as well as to grasp and explain the entire objective world (including things not directly observed) with the help of knowledge and experience.

4. Pattern recognition

   Patterns refer to categories of objective things or phenomena that are similar but not identical. Patterns cover a wide range, and images are one of them. (Image) pattern recognition is similar to image analysis in that they have the same input, while the different outputs can be easily converted. Recognition refers to mathematics and technology that automatically establish symbolic descriptions or logical reasoning from objective facts, so people define pattern recognition as the discipline of classifying and describing objects and processes in the objective world. At present, the recognition of image patterns mainly focuses on the classification, analysis, and description of the content of interest (object) in the image. On this basis, the goal of computer vision can be further realized. At the same time, many concepts and methods of pattern recognition are used in computer vision research; however, visual information has its particularity and complexity, so traditional pattern recognition (competitive learning model) cannot include all computer vision.

5. Artificial intelligence

   Artificial intelligence can be counted as a new theory, new tool, and new technology that has been widely studied and applied in the field of computer vision in recent years. **Human intelligence** mainly refers to the ability of human beings to understand the world, to judge the things, to learn the environment, to plan the behavior, to reason the thinking, to solve the problems, etc. Visual function is a manifestation of human intelligence, and similarly, computer vision is closely related to artificial intelligence. Many artificial intelligence technologies are used in the research of computer vision. Conversely, computer vision can also be regarded as an important application field of artificial intelligence, which requires the help of theoretical research results and system implementation experience of artificial intelligence. **Machine learning** is the core of artificial intelligence, which studies how to make computers simulate or implement human learning behaviors, thereby acquiring new knowledge or skills. This is the basis for computer vision to complete complex vision tasks. **Deep learning**, which has recently gained a lot of attention, improves and enhances basic machine learning methods. It tries to imitate the working mechanism of the human brain, to build neural networks that can learn to analyze, recognize, and interpret data such as images.

In addition to the above related disciplines, from a broader perspective, computer vision needs to use various engineering methods to solve some biological problems and complete the inherent functions of biology, so it is also related (mutual learning and mutual dependence) to biology, physiology, psychology, neurology, and other disciplines. In recent years, computer vision researchers have been closely integrated with visual psychology and physiology researchers and have obtained a series of research results. Computer vision belongs to engineering application science and is inseparable from disciplines such as industrial automation, human–computer interaction, office automation, visual navigation and robotics, security monitoring, biomedicine, remote sensing mapping, intelligent transportation, and military public security. On the one hand, the research of computer vision fully combines and utilizes the achievements of these disciplines; on the other hand, the application of computer vision also greatly promotes the in-depth research and development of these disciplines.

1.2 Computer Vision Theory and Framework

As a discipline, computer vision has its own origins, theories, and frameworks. The origin of computer vision should be traced back to the invention and application of computers. In the 1960s, the earliest computer vision technology has been studied and applied.

1.2.1 Visual Computational Theory

The research on computer vision did not have a comprehensive theoretical framework in the early days. In the 1970s, the research on object recognition and scene understanding basically detected linear edges as the primitives of the scene and then combined them to form more complex scene structures. However, in practical applications, comprehensive primitive detection is difficult and unstable, so the computer vision system can only input simple lines and corners to form the so-called building block world.

Marr's book *Vision*, published in 1982, summarizes a series of research results of his and his colleagues on human vision, proposes a **visual computational theory**, and provides a framework for understanding visual information. This framework is both comprehensive and refined and is the key to making the study of visual information understanding rigorous and moving visual research from the descriptive level to the mathematical science level. Marr's theory states that the purpose of vision must be understood before going to the details. This is applicable to a variety of information processing tasks. The gist of the theory is as follows.

1.2.1.1 Vision Is a Complex Information Processing Process

Marr believes that vision is a far more complex information processing task and process than people imagine, and its difficulty is often not recognized by people. A major reason here is that while it is difficult for a computer to understand an image, it is often a breeze for a human.

To understand the complex process of vision, two issues must first be addressed. One is the representation of visual information; the other is the processing of visual information. "Representation" here refers to a formal system (such as Arabic numeral system, binary numeral system, etc.) that can clearly express certain entities or certain types of information as well as some rules that explain how the system works. In the representation, some information is prominent and explicit, while other information is hidden and vague. Representation has a great influence on the difficulty of subsequent information processing. The "processing" of visual information refers to the transformation and gradual abstraction of different forms of representation through continuous processing, analysis, and understanding of the information.

Solving the problem of representation and processing of visual information is actually solving the problem of computability. If a task needs to be done by a computer, it should be computable, which is the problem of computability. In general, for a particular problem, a problem is computable if there is a program that gives an output in finite steps for a given input.

1.2.1.2 Three Essential Factors of Visual Information Processing

To fully understand and interpret visual information, three essential factors need to be grasped at the same time, namely, computational theory, algorithm implementation, and hardware implementation.

First, the highest level of visual information understanding is abstract **computational theory**. The question of whether vision can be calculated by modern computers needs to be answered by computational theory, but there is no clear answer so far. Vision is a process of feeling and perception. From the perspective of microscopic anatomical knowledge and objective visual psychology knowledge, people still lack the grasp of the mechanism of human visual function. Therefore, the discussion on visual computability is still relatively limited, mainly focusing on the number and symbol processing ability of existing computers to complete some specific visual tasks and so on.

Table 1.1 The meaning of the three essential factors of visual information processing

Essential factor	Name	Meaning and problems to be solved
1	Computational theory	What is the computation goal? Why is it computed like this?
2	Representation and algorithm	How to realize computational theory? What is input and output representation? What algorithm is used to realize the conversion between representations?
3	Hardware implementation	How to physically implement representations and algorithms? What are the specific details of computing structures?

Second, the objects of computer operation are discrete numbers or symbols, and the storage capacity of the computer is also limited. Therefore, after having the calculation theory, we must also consider the implementation of the algorithm. Therefore, we need to select an appropriate representation for the entities operated by the machining. On the one hand, the input and output representations of machining should be selected; on the other hand, we should determine the algorithm to complete the representation transformation. **Representation and algorithm** restrict each other, so three points should be paid attention to: (1) in general, there are many optional expressions; (2) the determination of the algorithm often depends on the selected representation; (3) given a representation, there can be many algorithms to complete the task. Generally, the instructions and rules used for machining are called algorithms.

Finally, how the algorithm is physically implemented must also be considered. Especially with the continuous improvement of real-time requirements, the problem of dedicated **hardware implementation** is often raised. It should be noted that the determination of an algorithm usually depends on the characteristics of the hardware that physically implements the algorithm, and the same algorithm can be implemented through different technical approaches.

After summarizing the above discussion, the content shown in Table 1.1 can be obtained.

There is a certain logical causal connection between the above three essential factors, but there is no absolute dependence. In fact, there are many different options for each essential factor. In many cases, the problems involved in explaining each essential factor are basically irrelevant to the other two essential factors (each essential factor is relatively independent), or one or two essential factors can be used to explain certain visual phenomena. The above three essential factors are also called by many people the three levels of visual information processing, and they point out that different problems need to be explained at different levels. The relationship among the three essential factors is often shown in Fig. 1.2 (in fact, it is more appropriate to regard it as two levels), in which the positive arrow indicates that it has a guiding meaning and the reverse arrow has a meaning of as basic. Note that once there is a computational theory, representations and algorithms as well as hardware implementations influence each other.

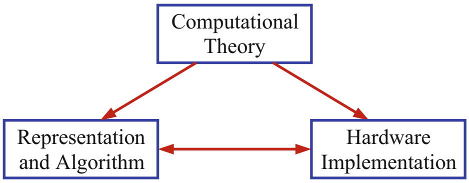

Fig. 1.2 The relationship of the three essential factors of visual information processing

1.2.1.3 Three-Level Internal Representation of Visual Information

According to the definition of visual computability, visual information processing can be decomposed into multiple transformation steps from one representation to another. Representation is the key to visual information processing. A basic theoretical framework for computer vision research and information understanding is mainly composed of the three-level representation structure of the visible world established, maintained, and explained by visual processing. For most philosophers, what are the nature of visual expressions, how they relate to perception, and how they support action are open to different interpretations. However, they agreed that the answers to these questions were all related to the concept **representation**.

1. Primal sketch

   The primal sketch denotes a 2D representation, which is a collection of image features and describes the contour part where the properties of the object surface change. The primal sketch representation provides the information of the contour of each object in the image and is a form of sketch representation of the 3D object. This way of representation can be proven from the human visual process. When people observe a scene, they always notice the drastic changes in it. Therefore, primal sketch should be a stage of the human visual process.

2. 2.5D sketch

   The 2.5D sketch is completely proposed to adapt to the computing functions of the computer. It decomposes the object according to the principle of **orthogonal projection** according to a certain sampling density, so that the visible surface of the object is decomposed into many facets (face element) of a certain size and geometric shape; each facet has its own orientation. Using a normal vector to represent the orientation of the facet in which it is located and composing a set of needles (the vector is shown with an arrow/needle) constitute a 2.5D sketch map (also called a needle diagram). In this type of diagram, the normal of each orientation takes the observer as the center. The specific steps to obtain the 2.5D sketch map (Fig. 1.3 shows an example) are as follows:

   (a) Decompose the orthogonal projection of the visible surface of the object into a collection of facets.
   (b) Use the normal lines to represent the orientation of the facet.
   (c) Draw each normal line and superimpose all normal lines on the visible surface within the outline of the object.

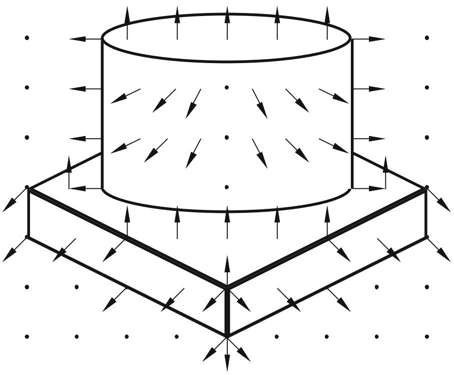

Fig. 1.3 2.5D sketch example

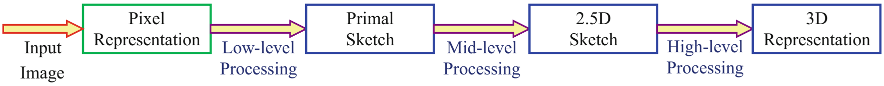

Fig. 1.4 The three-level representation decomposition of the Marr's framework

The 2.5D sketch map is actually an intrinsic image (see Sect. 1.3.2), because it shows the orientation of the surface element of the object, thus giving the information of the surface shape. Surface orientation is an intrinsic characteristic, and depth is also an intrinsic characteristic. The 2.5D sketch map can be converted into a (relative) depth map.

3. 3D representation

   3D representation is a representation form centered on the object (i.e., it also includes the invisible part of the object). It describes the shape and spatial organization of 3D objects in the object-centered coordinate system. Some basic 3D entity representations can be found in Chap. 9.

   Now come back to the problem of visual computability. From the perspective of computer or information processing, the problem of visual computability can be divided into several steps. Between the steps is a certain form of representation, and each step consists of a calculation/processing method that connects the two forms of representation (see Fig. 1.4).

According to the abovementioned three-level representation viewpoint, the problem to be solved by visual computability is how to start from the pixel representation of the original image, through the primal sketch representation and 2.5D sketch representation, and finally obtain the 3D representation. They can be summarized in Table 1.2.

Table 1.2 Representation framework of visual computability problem

Representation	Goal	Primitive
Image	Represent the brightness of the scene or the illuminance of the object	Pixel (values)
Primal sketch	Represent the location of brightness changes in the image, the geometric distribution of the object outline, and the organizational structure	Zero crossing point, end point, corner point, inflection point, edge segment, and boundary
2.5D sketch	Represent the orientation, depth, contour, and other properties of the visible surface of the object in the observer-centered coordinate system	Local surface orientation ("needle" primitives), surface orientation discontinuities, depth, and discontinuous point in depth
3D representation	Represent the object shapes and their spatial organization, by using voxels or surface elements, in a coordinate system centered on an object	3D model, with the axis as the skeleton; attach the volume element or face element to the axis

1.2.1.4 Visual Information Understanding Is Organized in the Form of Functional Modules

The idea of viewing the visual information system as a set of relatively independent functional modules is not only supported by the evolutionary and epistemological arguments in computing, but also some functional modules can be separated by experimental methods.

In addition, psychological research also shows that people obtain various intrinsic visual information by using a variety of clues or a combination of them. This suggests that the visual information system should include many modules. Each module obtains specific visual cues and performs certain processing, so that different weight coefficients can be combined with different modules to complete the visual information understanding task according to the environment. According to this point of view, complex processing can be completed with some simple independent functional modules, which can simplify research methods and reduce the difficulty of specific implementation. This is also very important from an engineering perspective.

1.2.1.5 The Formal Representation of Computational Theory Must Consider Constraints

During the image acquisition process, the information in the original scene will undergo various changes, including the following:

1. When a 3D scene is projected as a 2D image, the depth of the object and the invisible part of the information are lost.

2. Images are always obtained from a specific viewing direction. Different perspective images of the same scene will be different. In addition, information will be lost due to mutual occlusion of objects or mutual occlusion of various parts.
3. Imaging projection makes the illumination, object geometry and surface reflection characteristics, camera characteristics, and the spatial relationship between the light source, the object, and the camera all integrated into a single image gray value, which are difficult to be distinguished.
4. Noise and distortion will inevitably be introduced in the imaging process.

For a problem, if its solution is existing, unique, and continuously dependent on the initial data, then it is well-posed. If one or more of the above is not satisfied, it is ill-posed (under-determined). Due to the information changes in the various original scenes mentioned above, the method of solving the vision problem as the inverse problem of the optical imaging process becomes an ill-posed problem (becoming an ill-conditioned problem), so it is very difficult to solve. In order to solve this problem, it is necessary to find out the constraints of the relevant problems according to the general characteristics of the external objective world and turn them into precise assumptions, so as to draw conclusive and testable conclusions. Constraints are generally obtained with the aid of prior knowledge. The use of constraints can change ill-conditioned problems. This is because adding constraints to the calculation can make its meaning clear, thus enabling the problem to be solved.

1.2.2 Framework Issues and Improvements

Marr's visual computational theory is the first theory that has a greater impact on visual research. This theory has actively promoted research in this field and has played an important role in the research and development of image understanding and computer vision.

Marr's theory also has its shortcomings, including four problems about the overall framework (see Fig. 1.6):

1. The input in the framework is passive: what image is input, the system will process what image.
2. The processing goal in the framework remains unchanged, and the position and shape of the objects in the scene are always restored.
3. The framework lacks or does not pay enough attention to the guiding role of high-level knowledge.
4. The information processing process in the entire framework is basically bottom-up, one-way flow, and no feedback.

In response to the above problems, people have proposed a series of improvement ideas in recent years. Corresponding to the framework of Fig. 1.4, these improvements can be incorporated into new modules to obtain the framework of Fig. 1.5.

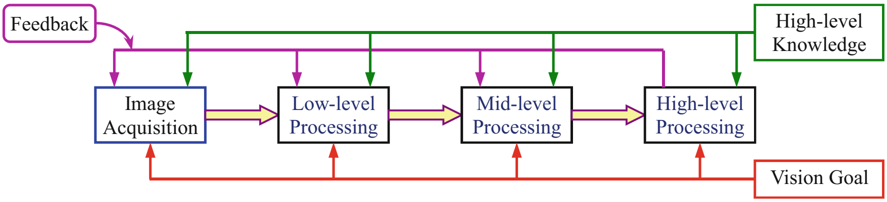

Fig. 1.5 Improved visual computational framework

In the following, with conjunction to Fig. 1.5, the four aspects of the original framework of Fig. 1.4 will be discussed in detail.

1. Human vision has initiative

   People will change the line of sight or perspective as needed to help observation and cognition. **Active vision** means that the vision system can determine the movement of the camera according to the existing analysis results and the current requirements of the vision task to obtain the corresponding image from the appropriate position and perspective. Human vision is also selective, one can stare (observing the region of interest at a higher resolution), or one can turn a blind eye to certain parts of the scene. **Selective vision** means that the vision system can determine the focus of the camera to obtain the corresponding image based on the existing analysis results and the current requirements of the vision task. Taking these factors into account, an image acquisition module is added to the improved framework, which is also considered together with other modules in the framework. This module should choose the image acquisition modes according to the visual purpose.

   The aforementioned active vision and selective vision can also be regarded as two forms of active vision: one is to move the camera to focus on a specific object of interest in the current environment; the other is to focus on a specific region in the image and dynamically interact with it to get an interpretation. Although the two forms of active vision look very similar, in the first form, the initiative is mainly reflected in the observation of the camera, while in the second form, the initiative is mainly reflected in the processing level and strategy. Although there is interaction in both forms, that is, vision has initiative, mobile cameras need to record and store all the complete scenes, which is a very expensive process. In addition, the overall interpretations obtained in this way are not necessarily used. Collecting only the most useful part of the scene, narrowing its scope, and enhancing its quality to obtain useful interpretations mimic the process of human interpretation of the scene.
2. Human vision can be adjusted for different purposes

   Purposive vision means that the vision system makes decisions based on the purpose of vision, such as whether to fully recover information like the position and shape of objects in the scene or just detect whether there is an object in the scene. It may give a simpler solution to vision problems. The key issue here is to determine the purpose of the task. Therefore, a visual purpose box (vision goal) is added to the improvement framework. Qualitative analysis or quantitative analysis can be determined according to different purposes of understanding (in practice, there are quite a lot of occasions where only qualitative results are sufficient; no complex quantitative result is needed). However, the current qualitative analysis still lacks complete mathematical tools. The motivation of purposive vision is to clarify only part of the information that is needed. For example, the collision avoidance of autonomous vehicles does not require precise shape descriptions, and some qualitative results are sufficient. This kind of thinking does not have a solid theoretical basis, but the study of biological vision systems provides many examples.

Qualitative vision, which is closely related to purposive vision, seeks a qualitative description of an object or scene. Its motivation is not to express geometric information that is not needed for qualitative (nongeometric) tasks or decisions. The advantage of qualitative information is that it is less sensitive to various unwanted transformations (such as slightly changing perspectives) or noise than quantitative information. Qualitative or invariant can allow easy interpretation of observed events at different levels of complexity.

3. Humans have the ability to completely solve visual problems with only partial information obtained from images

   Humans have this ability due to the implicit use of various knowledge. For example, after obtaining object shape information with the aid of CAD design data (using object model library), it can help solve the difficulty of restoring the object shape from a single image. The use of high-level (domain) knowledge can solve the problem of insufficient low-level information, so a high-level knowledge frame (module) is added to the improved framework.

4. There is an interaction between the sequential processing processes in human vision

   The human visual process has a certain sequence in time and different levels in meaning, and there is a certain interaction between the various steps. Although the mechanism of this interaction is not yet fully understood, the important role of high-level knowledge and feedback from the later results to low-level processing has been widely recognized. From this perspective, the feedback control flow is added to the improvement framework, and the existing results and high-level knowledge are used to improve visual efficiency.

1.2.3 A Discussion on Marr's Reconstruction Theory

Marr's theory emphasizes the reconstruction of the scene and uses the reconstruction as the basis for understanding the scene.