SUNG-EUI YOON, KAIST

RENDERING

FREELY AVAILABLE ON THE INTERNET

http://sglab.kaist.ac.kr/~sungeui/render

First printing, August 2018

Contents

Dedicated to TaeYoung and JaeHa.

Preface

Rendering is a way of visualizing various 3D models in 2D images or videos. It is one of fundamental tools in the field of computer graphics. Thanks to its ubiquitous demand, it is not only used for applications in computer graphics, but also widely used for many other fields.

There have been tremendous progress on rendering techniques. One of epitomes for rendering techniques is games, where we can see real-time, yet high-quality rendering images. These real-time techniques are commonly based on the concept of rasterization, which is the main theme of Part I of this book. Another successful application of computer graphics is movie. Unlike games, the movie production accommodates much longer computational time for higher rendering quality. Ray tracing based rendering techniques, therefore, are utilized more frequently for movies. These ray tracing techniques are mainly discussed in Part II.

These techniques have been developed for many decades. For example, the concept of ray tracing was introduced to the field of computer graphics at 1980. Since rendering techniques have been studied for a long period of time, it is very hard to catch up all the major concepts, unless properly guided. Also, new concepts and techniques have been constantly proposed.

Many graphics books are available, but only a few rendering books are available. Furthermore, most of them focuses either one of two main rendering techniques, rasterization and ray tracing, in an advanced manner. Given this situation, I decided to treat both of them, while covering most fundamental concepts of those techniques. I will also cover advanced topics as I have more time, built on top of those basic concepts.

In order to save time of writing this book and better explain concepts, I re-used many existing materials (e.g., images) of lecture slides and papers. For each of them, I mentioned its source, but here I'd like to point out that I borrowed many images from lecture slides used in a Computer Graphics course given at University of North Carolina at Chapel Hill for Part I and lecture slides of Prof. Kavita

Bala for Part II. Also, the latex template of this book is based on the Tufte's design style and is based on the Apache license.

Finally, many students of CS380, CS480, CS580 offered at KAIST posed many interesting questions that are the basis of many Q&A parts of this book. Also, many of them gave useful comments on different parts of this book.

Sung-eui, KAIST

July, 2018

1
Introduction

Rendering is one of the fundamental techniques in computer graphics, visualization, and many other related fields. Since the rendering technique has been widely used in many different applications, its perceived meaning can vary a lot depending on people using it.

For users and developers for games, rendering techniques should be interactive and can support many interesting visual effects (e.g., magic fire). In terms of the performance, the rendering part used in games should take less than 10 ms, since the whole frame takes 33 ms assuming 30 frames per second, and other parts (e.g., game logics and network) can take 10 ms to 20 ms. As a result, rendering methods adopted in such games should be extremely fast[1].

For viewers, artist, and developers for movies, rendering techniques should be photo-realistic and provide even artistic controls on effects that they want to express. In many movies (e.g., Jurassic Park), we see scenes captured from real cameras and mixed together with computer generated effects and virtual objects. Rendering methods for these movies should be indistinguishable between real and virtual scenes. As a result, these techniques are usually based on physics and simulations of light and material interactions. Furthermore, artists and directors making such movies are not satisfied with such realistic looking results[2]. They want to convey particular emotion and mood on computer generated effects. We thus need techniques accommodating such user inputs.

As you can see, there are such a wide variety of rendering applications with different characteristics. Therefore, a single rendering method satisfying all those characteristics and requirements is hard to be developed. As a result, many different rendering and visualization methods have been developed. Instead of covering them in detail in this book, we would like to cover main techniques and their variations.

[1] Games requires real-time rendering techniques spending only 10 ms for each frame

[2] Rendering used for movies needs to provide realistic results, while supporting various artistic directions

1.1 Rendering Techniques

At a high level, there are two main, but different rendering techniques: rasterization and ray tracing.

There are two main rendering techniques: rasterization and ray tracing

Rasterization is to traverse triangles of a model and project triangles to the frame buffer. Rasterzation is classified as an object driven rendering method and has been widely accelerated by various hardwares (e.g., GPUs) because of its simplicity. Thanks to the simplicity and the hardware acceleration, rasterization based rendering methods can show an interactive rendering performance even for massive models consisting of more than hundreds of millions of triangles [3]. Thanks to these features, rasterization techniques are available in OpenGL and DirectX, graphics APIs, and adopted in many games through game engines (e.g., Unity).

[3] Sung-Eui Yoon, Brian Salomon, Russell Gayle, and Dinesh Manocha. Quick-VDR: Interactive View-dependent Rendering of Massive Models. In *IEEE Visualization*, pages 131–138, 2004

Ray tracing, however, generates rays per each pixel and finds triangles that intersect with these rays by traversing an acceleration hierarchy. Ray tracing is classified as a view-driven rendering method and requires random access on meshes and hierarchies. Therefore, it requires much complex control logics and caches and in turn has been showing much (e.g., two orders of magnitude) slower performance than that of rasterization based methods.

Although ray tracing shows much slower performance than rasterization, it can naturally support physically-correct rendering because its algorithm follows the physical intersections between lights and materials. Therefore, it has been widely used in offline applications (e.g., movies) that require high-quality rendering results. On the other hand, rasterization has been widely used for interactive applications such as games.

While there are such stereotypical usages of rasterization and ray tracing, these techniques are still under active, yet steady development, and are thus improved in many different directions. For example, many games want to interactively support realistic rendering effects that ray tracing has been able to support in the domain of rasterization. Furthermore, some of recent applications such as Poketmon Go, an AR (augmented reality) application, needs to seamlessly integrate camera-captured scenes and computer generated effects. To realize this, ray tracing and rasterization techniques are used together to achieve both the performance and quality[4].

[4] Ray tracing and rasterization are used together for achieving the fast performance and high quality.

Relationship with other fields. Computer graphics commonly assumes that virtual scenes are represented by various types of models such as triangles for the scene geometry and BRDF for material appearance (Fig. 1.1). The main output of various computer graphics methods is an image or a series of images known as video. Com-

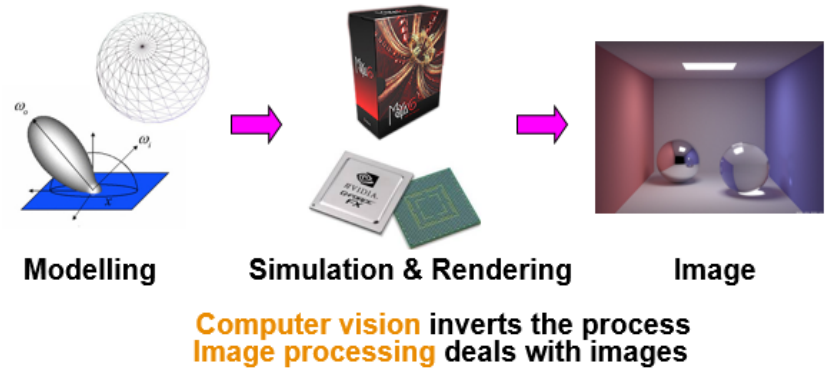


Figure 1.1: An overall structure of computer graphics. Images are adopted from Google image search.

mon methods for computer graphics include rendering, a type of simulation for light and material interactions, and many other types of simulations such as cloth, fire, character simulations. Computer vision commonly starts with images and attempts to extract models (e.g., geometry and BRDF), and image processing deals with images for denoising or many other image improvement. One of well known image processing tools is Photoshop from Adobe.

These different approaches have been developed and matured in its own fields (e.g., computer graphics and vision). Recently, these techniques developed from different fields are mixed together to create novel applications and approaches. As a result, their boundaries become rather blurred in these days.

Applications of computer graphics. Numerous applications of computer graphics exist. A lot of them are in the entertainment business for making games and movies (Fig. 1.2). Some movies are generated totally based on computer graphics, or some scenes of movies have various special effects. They also get renewed attentions with other related technology advances such as introduction of 3D TV to consumer markets and head mounted display (HMD) for virtual reality (VR) and augmented reality (AR).

In addition to various entertainment applications, various product designs and analysis such as computer-aided design (CAD) uses computer graphics. Also, medical and scientific visualization is a big part of computer graphics. Finally, information visualization that associates various geometric meaning to complex data (e.g., graphs) are getting bigger and bigger.

Organization of the book. Rendering has been studied in various aspects covering optics and novel applications. As a result, we focus on the following two parts in this book.

Figure 1.2: Applications of computer graphics. From the top left, image cuts of startcraft, toy story, a CT image of mouse skull, a weather visualization from LLNL, and a double eagle oil tanker for CAD.

1. **Rasterization.** Rasterization is an efficient rendering technique that mainly works in an image space that can be easily accelerated in GPUs. This approach is discussed in Part I.

2. **Ray tracing.** Ray tracing is a common approach of simulating the physical interaction between the light and materials. It is therefore widely used for providing physically-based rendering. This is discussed in Part II.

1.2 *Related Materials*

Rendering techniques have been studied for several decades, and excellent books are available. We list some of them here:

- Fundamentals of Computer Graphics, by Peter Shirley et al. [5]. This book covers various fundamental topics of computer graphics.

- Physically based rendering by Pharr et al. [6].. This book also covers a wide variety of topics of physical-based rendering. It also provides source codes for all the concepts discussed in the book. If you want to have hands-on experience on physics-based rendering, this book provides both theoretical concepts and practical programming tools.

- Advanced Global illumination, by Dutre et al. [7]. This book covers physics-based rendering techniques.

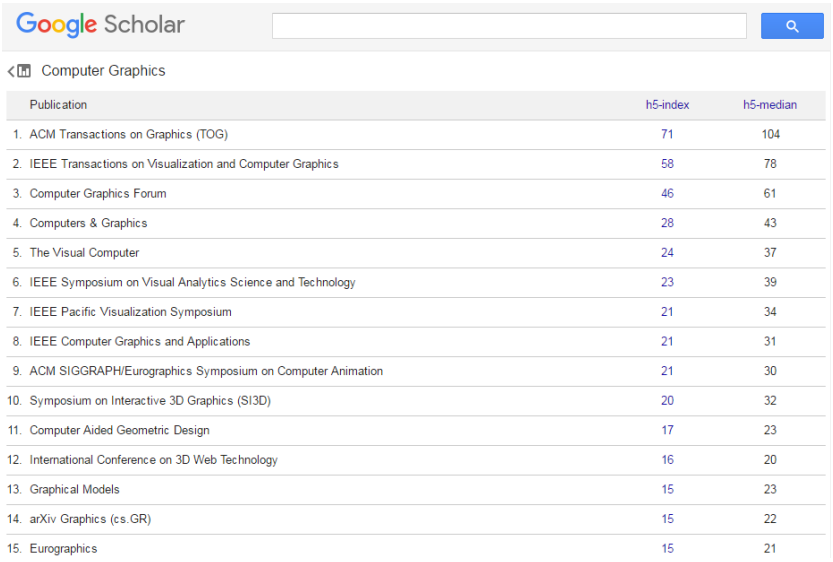
- Realistic ray tracing, by Shirley et al. [8]. While this book is rather old, it covers various concepts and detailed information of ray tracing, which is one of main ingredients of building physics-based rendering.

[5] Peter Shirley and Steve Marschner. *Fundamentals of Computer Graphics*. A. K. Peters, Ltd., 3rd edition, 2009

[6] Matt Pharr and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation 2nd*. Morgan Kaufmann Publishers Inc., 2010a

[7] Philip Dutre, Kavita Bala, and Philippe Bekaert. *Advanced Global Illumination*. AK Peters, 2006

[8] Peter Shirley and R. Keith Morley. *Realistic Ray Tracing*. AK Peters, second edition, 2003



Google Scholar

Computer Graphics

	Publication	h5-index	h5-median
1.	ACM Transactions on Graphics (TOG)	71	104
2.	IEEE Transactions on Visualization and Computer Graphics	58	78
3.	Computer Graphics Forum	46	61
4.	Computers & Graphics	28	43
5.	The Visual Computer	24	37
6.	IEEE Symposium on Visual Analytics Science and Technology	23	39
7.	IEEE Pacific Visualization Symposium	21	34
8.	IEEE Computer Graphics and Applications	21	31
9.	ACM SIGGRAPH/Eurographics Symposium on Computer Animation	21	30
10.	Symposium on Interactive 3D Graphics (SI3D)	20	32
11.	Computer Aided Geometric Design	17	23
12.	International Conference on 3D Web Technology	16	20
13.	Graphical Models	15	23
14.	arXiv Graphics (cs.GR)	15	22
15.	Eurographics	15	21

Figure 1.3: This shows a list of graphics related conferences and journals according to Google Scholar at 2016. Note that many conferences papers in computer graphics are published at journals, and thus journals (e.g., ACM Trans. on Graphics) are ranked higher than well-known conferences (e.g., SIGGRAPH).

These books cover fundamental concepts of rendering, but lacks recent developments. If you want to follow those recent techniques, you can find recent papers through the following:

- Google scholar. You can find recent technical papers from various search engines. Especially, Google scholar is useful, since it also identifies papers that refer to a particular paper. By looking this information, you can find prior and future works given a particular paper.

- Graphics conferences and journals. Novel ideas are generated in every where. One can easily learn those novel ideas by looking at recent papers published at graphics conferences and journals. One of well-known of them is ACM SIGGRAPH, whose papers are published at a journal called ACM Trans. on Graphics (ToG). Google Scholar also provides a list of influential conferences and journals with their ranking (Fig. 1.3).

1.3 *Common Q & A*

Do we need an excellent artistic sense to study computer graphics or to become an technical expert in this field? Not really. Of course, it is always better to have a good artistic sense to work on visual data processing. However, if some jobs require such a high standard of artistic senses, those jobs may be for artistic designers, not for engineers. In my opinion, it is more important to have better

engineering backgrounds (e.g., mathematical backgrounds and algorithm developments) and problem-solving skills. For example, I don't have any sense of art, but I work on computer graphics!

I have found that something like tea pot and bunny models are widely used in many papers and technical videos. Why? You made a good observation. Some of models including the Utah teapot and Stanford bunny have been created earlier as research results or research benchmarks. Then, these models are distributed to other researchers for their follow-on research. That's why these models are widely used in many papers.

Part I

Rasterization

Rasterization is one of most popular rendering techniques developed for computer graphics. It simply projects triangles in a scene into a viewing space and color pixels overlapped with those triangles. This approach is very simple and thus can be implemented efficiently in specialized hardwares. Especially, many graphics hardware and GPUs support this rasterization scheme.

It, however, does not simulate the natural interaction between light and materials. Simply speaking, in reality, objects are not projected into our eyes! Due to this issue, rasterization schemes have fundamental drawbacks of simulating various rendering effects such as shadows, transparency, and so on. Nonetheless, thanks to its fast performance, many techniques and fixes have been proposed to improve its rendering quality.

In this part, we discuss the fundamental engine of rasterization, which is developed in many graphics library such as OpenGL and DiretX accelerated by GPUs. In other parts , we study global illumination that physically simulates interactions between lights and materials.

1.4 Related Materials

Many useful resources for rasterization techniques are available. Some of them are listed here:

- OpenGL Programming Guide. OpenGL is one of very popular computer graphics library that can be used in a wide variety of computing platform including Windows, Linux, and mobile OS. OpenGL provides various useful low-level graphics APIs, and they are well explained in this book and in its reference book. Early version of these books are available on free at internet. We also explain some of OpenGL APIs and their concepts, when we explain concepts of rasterization for delivering concrete examples.
- Real-time rendering [9] and its resource. This book covers a vast amount of topics that are related to rasterization and real-time rendering techniques. Its resource cite [10] has many useful web pages and links.
- OpenGL tutorials. Many OpenGL tutorials exist at Web. Some of them are based on the legacy OpenGL, but `http://www.opengl-tutorial.org/` discusses useful tutorials based on a recent OpenGL (ver. 3.3 and later).

[9] Tomas Akenine-Möller, Eric Haines, and Naty Hoffman. *Real-Time Rendering 3rd Edition*. A. K. Peters, Ltd., 2008

[10] `http://www.realtimerendering.com/`

2
Rendering Pipeline

Rendering triangles for scenes requires an excessive amount of computation time, since there could be many triangles representing scenes, and each triangle can map to hundreds of pixels in the screen space. As a result, carefully designed steps, known as rendering pipeline, has been proposed.

2.1 Classic Rendering Pipeline

Let us first discuss the classic rendering pipeline, before studying a modern, but complex one.

Fig. 2.1 shows an example of a classic rendering pipeline running on a GPU. An graphics application runs on a CPU in general and sends geometry of the scene and a camera setting that its user wants to see to a GPU by using a graphics library such as OpenGL. The rendering pipeline implemented in a GPU processes such requests and computes an output image displayed in a screen.

In general, the rendering pipeline consists of many steps for drawing an image from the user's camera position and orientation in an efficient manner. At a high level, they usually breaks into vertex processing and pixel processing units. The vertex processing step transforms input geometry into ones mapped in the screen space. Those ones are converted into pixels with appropriate colors by the pixel processing step, and this step is commonly known as the rasterization step.

Historically, these steps take a high computation time and thus are implemented in a chip in a hard-wired manner. These steps, therefore, are rather fixed functions and invoked through graphics APIs. As we have more processing power and developers request more flexibility on programming, the GPU implementing these steps become more general like CPU and can run various graphics programs such as OpenGL shaders.

While more accurate rendering techniques (e.g., global illumi-

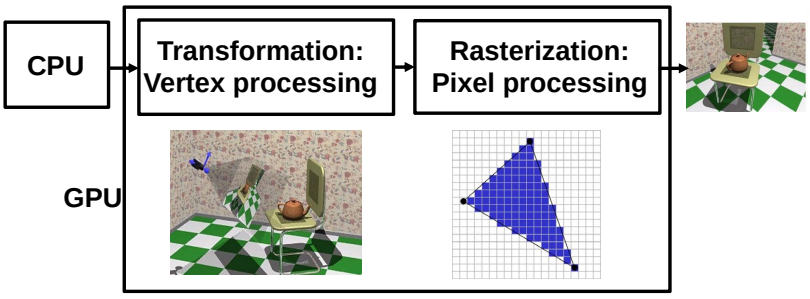


Figure 2.1: This shows a schematic diagram of classic rendering pipeline consisting only two steps: vertex and pixel processing steps.

nation) have been proposed with high performance, rasterization scheme is one of the most efficient rendering algorithms specializing on local illumination, which considers the light energy transfer between a surface and a light source. We therefore study this scheme in a detailed manner in Chapter 3 and 7.

2.2 *Modern Rendering Pipeline*

Fig. 2.2 shows a schematic view on a modern rendering pipeline adopted in OpenGL 3.0. While this differs a lot from the classical one, it shares both vertex and pixel (e.g., fragment) processing steps.

- **Vertex specification.** Vertices and triangles are defined and passed to the following step.
- **Vertex processing.** Each vertex is processed by a vertex shader, a program working on each vertex. It performs various modeling transformation, viewing, and projection transformations.
- **Vertex post-processing.** It performs various basic operations after the vertex processing step and serves as a setup stage for the following steps such as rasterization. It includes clipping (Sec. 6.4), homogeneous divide (Sec. 4.2.1), and viewport transformation (Sec. 3.1).
- **Primitive assembly.** Face culling is performed in this step.
- **Rasterization.** This step converts a triangle represented by vertices into a number of fragments.
- **Fragment shader.** It also processes each fragment generated by the prior rasterization step.

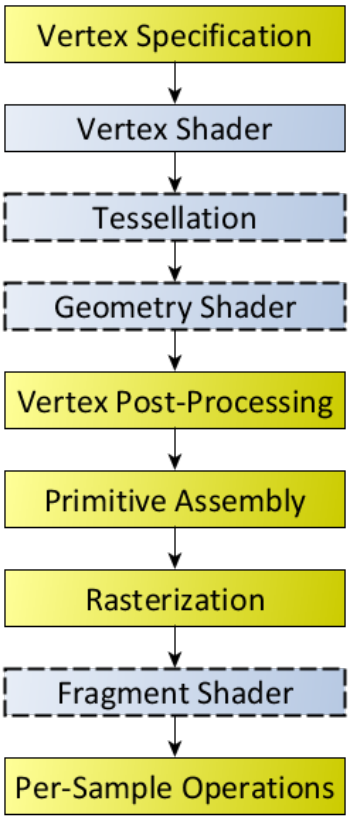


Figure 2.2: This shows a rendering pipeline adopted in OpenGL 3.0. This image is excerpted from the OpenGL homepage.

2.3 OpenGL and Other Tools

The rendering pipeline has been implemented and accelerated in GPUs. To enjoy such hardware acceleration, we use OpenGL and DirectX. OpenGL is more widely available in different operation systems and devices, since DirectX depends on Windows OS. Most concepts and techniques that are covered in this part are available at such APIs. Nonetheless, it is useful to know what other tools related to graphics are available and their goals. Fig. 2.3 shows other tools and languages that can utilize various features of GPU other than the rasterization.

Recently, Vulkan was introduced for achieving even higher performance on mobile phones [1] that have lower performance than PCs. For achieving its goal, Vulkan allows users to various low-level APIs with low overheads and multi-tasking. Nonetheless, it comes with certain costs such as higher programming burdens to users.

[1] G. Sellers and J.M. Kessenich. *Vulkan Programming Guide: The Official Guide to Learning Vulkan*. Addison Wesley, 2016

While these APIs provide the full features of the rendering pipeline, they are rather low-level APIs. When we want to develop high-level applications such as a game, we need to utilize a more powerful set of tools and SWs. This is a gap that modern game and rendering engines such as Unity try to fill in. Additionally, in graphics applications (e.g., games and movies), content creation is one of main tasks, and many modeling and animation tools are available.

Initially, GPU is designed as a specialized hardware to accelerate the rendering process, which is captured in the rendering pipeline. However, as the performance of GPU is getting higher and various demands on programmability on the rendering pipeline arise. As a result, parts of vertex and fragment stages can be programmable through a dedicated language, i.e., GLSL and HLSL.

While these shading languages are designed to effectively utilize functions of GPUs for graphics applications, non-traditional needs on using GPUs for non-graphics applications keep increasing, thanks to its higher performance on streaming tasks than CPUs. To accommodate such demands, a general purpose language for utilizing GPUs has been proposed, and CUDA and OpenCL are two examples.

2.3.1 Common Questions

What if we have new input devices (e.g., joystick, or multiple input devices used in PlayStation or XBox)? How can we handle those devices in OpenGL programs? OpenGL does not have any functionality to support those various input devices. GLUT library supports some of basic input devices such as keyboard and mouses. For other devices, you need to use other external libraries that support those

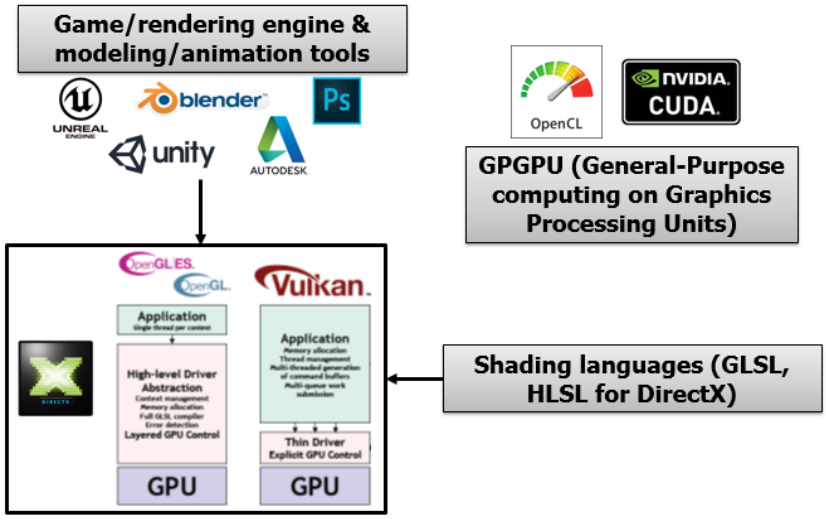


Figure 2.3: This figure shows other APIs, SWs, and languages that are related to OpenGL and computer graphics. In this book, we mainly discuss the core rendering pipeline that rasterizes input models. Nonetheless, many game and rendering engines (e.g., Unity) are commonly used as convenient, high-level tools. Also, shading languages are used in recent OpenGL verions, to add various details on rendering results. Additionally, general purpose computing languages for GPU (e.g., CUDA) are also used for implementing arbitrary programs on GPUs. Images are excerpted from the Vulkan overview and Google images.

devices.

In what cases, is OpenGL used rather than DirectX? OpenGL is cross-platform graphics API, while DirectX is proprietary library for Windows. Because of the openness of OpenGL, it, more specifically, OpenGL ES, is widely used for many embedded systems including mobile phones.

In what portions of my OpenGL program are executed in CPU and GPU? In a typical OpenGL program, rendering parts (e.g., portions started with glBegin and ended with glEnd) are performed in GPU, graphics hardware, if your computer is equipped with such GPU. All the control parts, e.g., calling OpenGL functions and handling events, are performed in CPU. In other words, various functionality inside OpenGL APIs are commonly performed in GPU, while all the other parts are performed in CPU.

3
Transformation

Many components of rasterization techniques rely upon different types of transformation. In this chapter, we discuss those transformation techniques.

3.1 Viewport Transformation

In this section, we explain the viewport transformation based on an example. Fig. 3.1 show different spaces that we are going to explain.

Suppose that you have an arbitrary function, $f(x, y)$, as a function of 2 D point (x, y); e.g., $f((x, y)) = x^2 + y^2$. Now suppose that you want to visualize the function in your computer screen with a particular color encoding method, e.g., heat map that assigns hot and cold colors depending on values of $f(x, y)$.

This function is defined in a continuous space, say x and y can be any real values. In computer graphics, we use a term of world to denote a model or scene that we would like to visualize or render. In this case, the function $f(x, y)$ is our world. Our goal is to visualize this function so that we can understand this function better. In many cases, the world is too large and thus we cannot visualize the whole world in a single image. As a result, we commonly introduce a camera to see a particular region of the world.

Unfortunately, our screen is not in the continuous space and has only a limited number of pixels, which is represented by a screen resolution. Our graphics application can use the whole screen space or some part of it. Let us call that area as a screen space. Fig. 3.2 show common conventions of the screen space. Finally, we visualize a part of the world seen through the camera into a part of our screen space, which is commonly known as a viewport; note that we can have multiple viewports in our screen.

Suppose a position, x_w, in the world that we are now seeing in the camera. In the end, we need to compute its corresponding position, x_s, in our screen space of the viewport. If we know x_s, we can draw

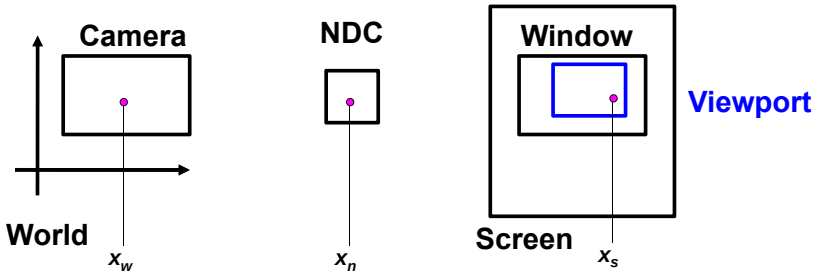


Figure 3.1: This shows a mapping from a viewable region in the world through a camera to the viewport in our screen space pass through the intermediate space, normalized device coordinate (NDC).

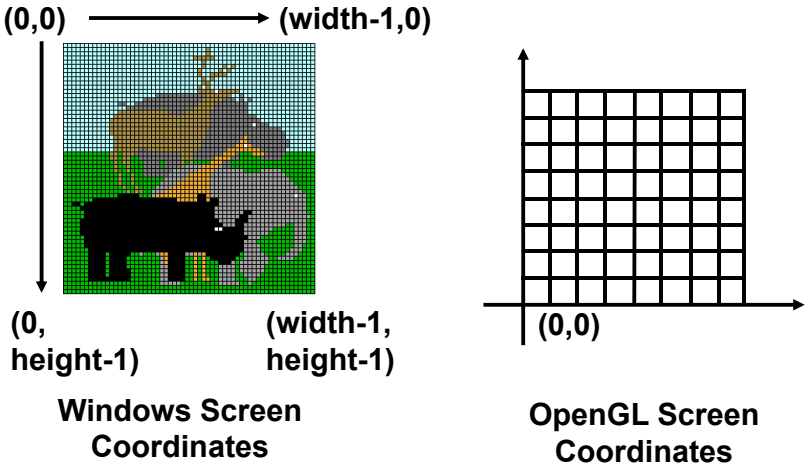


Figure 3.2: This shows two different conventions of screen coordinate spaces.

the color of the world position x_w at x_s. The question is how to compute x_s from x_w, i.e., the mapping from the world space to the viewport or screen space.

Normalized device coordinate (NDC). While world and screen spaces are two fundamental spaces, we also utilize NDC. NDC is a canonical space, whose both X and Y values are in a range of $[-1, 1]$. NDC serves as an intermediate space that is transformed to the screen space, which is hardware-dependent space. As a result, given the world potion x_w, we first need to compute a position in the NDC space, x_n, followed by mapping to x_s. We will also see various benefits of using NDC later, which include simplicity and thus efficiency of various rasterization operations.

Mapping from the world space to NDC. Suppose that the part of the world that we can see through a camera is represented by $[w.l, w.r] \times [w.b, w.t]$, where $w.l$ and $w.r$ are the visible range along X-axis and $w.b$ and $w.t$ define the visible range in Y-axis, while the

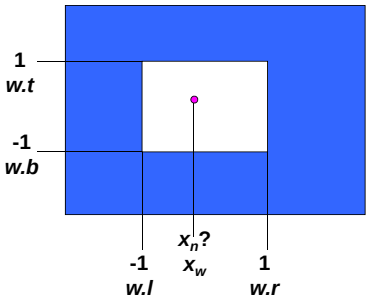


Figure 3.3: Mapping between the world space and NDC.

NDC space is represented by $[-1,1] \times [1,1]$.

Since the relative ratio of x_w and x_n is same in each space, we have the following relationship:

$$\frac{x_n - (-1)}{1 - (-1)} = \frac{x_w - (w.l)}{w.r - w.l}.$$
$$x_n = 2\frac{x_w - w.l}{w.r - w.l} - 1.$$
$$x_n = Ax_w + B,$$

where $A = \frac{2}{w.r - w.l}, B = -\frac{w.r + w.l}{w.r - w.l}$. This equation indicates that given the information, we can compute the NDC coordinate with one multiplication and one summation. Similarly, we can derive the mapping equation from x_n to x_s.

An issue of this approach is that there are too many pixels and thus evaluating such simple equations requires computational time. Since most graphics applications require interactive or real-time performance, we need to think about efficient way of handling these operations early in the history of computer graphics. Furthermore, it turns out that such mapping and similar transformations are very common operations in many graphics applications. The most common way of handling them in an efficient and elegant way is to adopt linear algebra and use matrix operations.

3.1.1 Common Questions

Can glBegin () with GL_POLYGON support concave polygons? According to its API description, GL_POLYGON works only with convex polygons. But, what may happen with concave polygons? Since it is not part of the specification of OpenGL, each vendor can have their own handling method for that kind of unspecified cases. If you are interested, you can try it out and let us know.

In the case of rendering circles, shown as an example in the lecture note, we render them by using lines. Is there a direct primitive that supports the circle? OpenGL has a limited functionality that supports continuous mathematical representations including circles, since a few model representations (e.g., triangles) have been widely used and it is hard to support all the possible representations. However, OpenGL keeps changing and it may support many continuous functions in a near future. At this point of time, we need to discretize continuous functions with triangles or other simple primitives and render them.

We use the NDC between the world space and the screen space. Isn't it inefficient? Also, don't we lose some precision during this

process? There is certainly some overhead by introducing the NDC. However, it is very minor compared to its benefits in terms of simplifying various algorithms employed throughout the rendering process. Yes. We can lose more precision during the conversion process due to float operations. However, it may be very small and may not cause significant problems for rendering purposes. Nonetheless, the transformation is based on analytic equations, not pixels, and thus can be easily recovered to the original information.

OpenGL is designed for cross-platform. But, I think that it means that we cannot use assembly programming for higher optimizations. Yes. You're right. We cannot use assembly languages for such optimizations. However, programmers for graphics drivers for each graphics vendor definitely use an assembly language and attempt to achieve the best performance. High-level programmers like us rely on such drivers and optimize programs with OpenGL API available to us.

Multi-threading with OpenGL: Since OpenGL has been designed very long time ago and has many different threads, it requires some cares to use multiple threads for OpenGL. There are many articles in internet about how to use multiple threads with OpenGL. I recommend you to go over them, if you are interested in this topic.

Why do we use a viewport? The viewport space doesn't need to be the whole window space. Given a window space, we can decompose it into multiple sub-spaces and use sub-spaces for different purposes. An example of using multiple viewports is shown in Fig. 3.4.

3.2 *2D Transformation*

In this section, we discuss how to represent various two dimensional transformation in the matrix form. We first discuss translation and rotation.

2D translation has the following forms:

$$x' = x + t_x, \tag{3.1}$$

$$y' = y + t_y, \tag{3.2}$$

where (x, y) is translated as an amount of (t_x, t_y) into (x', y'). They are also represented by a matrix form:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}. \tag{3.3}$$

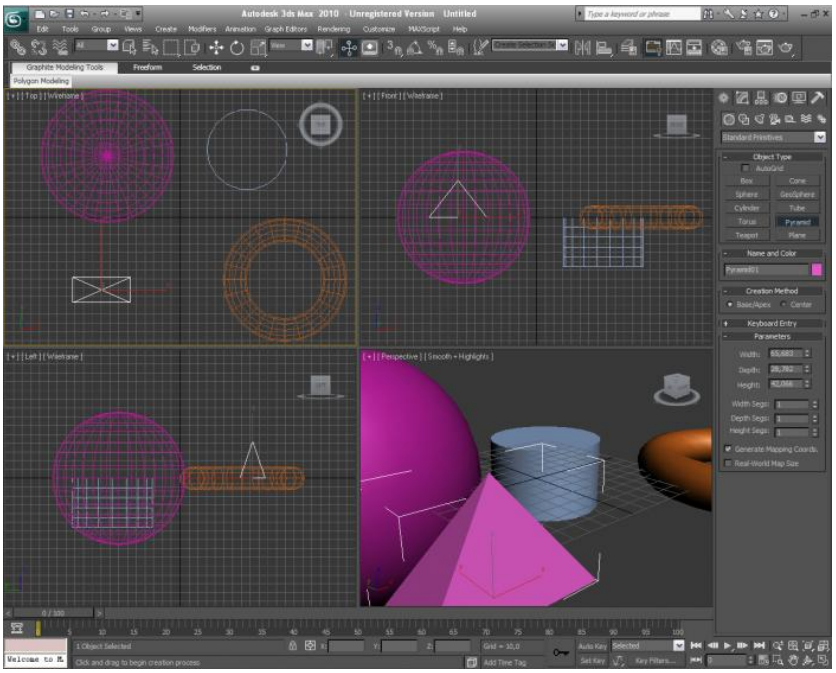


Figure 3.4: This figure shows multiple viewports, each of which shows an arbitrary 3D view in addition to top, front, and side views. The image is excerpt from screenshots.en.sftcdn.ne.

Given the 2D translation, its inverse function that undoes the translation is:

$$x = x' - t_x, \tag{3.4}$$
$$y = y' - t_y. \tag{3.5}$$

Also, its identity that does not change anything is:

$$x' = x + 0, \tag{3.6}$$
$$y' = y + 0. \tag{3.7}$$

Let us now consider 2D rotations. Rotating a point (x, y) as an amount of θ in the counter-clock wise is:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = R_\theta \begin{bmatrix} x \\ y \end{bmatrix}, \tag{3.8}$$

where R_θ is the rotation matrix. Its inverse and identity are defined as the following:

$$R^{-1} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \tag{3.9}$$
$$R_{\theta=0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.10}$$

Suppose that you want to rotate an object by 30 degrees, followed by rotating it again with 60 degrees. We intuitively know that rotating 90 degrees in a single time gives the same effect of rotating 30 degrees and 60 degrees again.V Formally, one can prove the following equation:

$$R_{\theta_2} R_{\theta_1} = R_{\theta_1 + \theta_2}. \tag{3.11}$$

3.2.1 *Euclidean Transformation*

In this subsection, we would like to discuss a particular class of transformation, Euclidean transformation. The Euclidean transformation preserves all the distances between any pairs of points. Its example includes translation, rotation, and reflection. Since the shape of objects under this transformation is preserved, the Euclidean transformation is also known as rigid transformation.

This rigid transformation is one of most common transformation that we use for various game and movie applications. For example, camera rotation and panning are implemented by the rigid transformation.

Mathematically, the Euclidean transformation is represented by:

$$T(x) = Rx + t, \tag{3.12}$$

where R and t are rotation matrix and 2D translation vector.

While this is a commonly used mathematical representation, this representation has a few drawback for graphics applications. Typically, we have to perform a series of rotation and translation transformation for performing the viewport transformation, camera operations, and other transformation applied to objects. As a result, it can take high memory and time overheads to apply them at runtime. Furthermore, there is cases that we need to compute a invert operation from a coordinate from the screen space to the corresponding one in the world space. Given the series of rotation and translation operations, the inverting operation can require multiple steps.

As an elegant and efficient approach to these issues, the homogeneous coordinate has been introduced and explained in the next section.

3.2.2 *Homogeneous Coordinate*

Homogeneous coordinates are originally introduced for projective geometry, but are widely adopted for computer graphics, to represent the Euclidean transformation in a single matrix.

Homogeneous coordinates provides various benefits for transformation and are thus commonly used in graphics.

Suppose a 2D point, (x, y) in the 2D Euclidean space. For the homogeneous coordinates, we introduce an additional coordinate,

and (x, y) in the 2D Euclidean space corresponds to $(x, y, 1)$ in the 3D homogeneous coordinates. In fact, $(zx, zy, z), z \neq 0$ also corresponds to (x, y) by dividing the third coordinate z to the first and second coordinates, to compute the corresponding 2D Euclidean coordinate.

Intuitively speaking, (zx, zy, z) represents a line in the 3D homogeneous coordinate space. Nonetheless, any points in the line maps to a single point (x, y) in the 2D Euclidean space. As a result, it can describe a projection of a ray passing through a pin hole to a point.

Let us now describe its practical benefits for our problem. Before we describe the Euclidean transformation (Eq. 3.12) and its problems. By using the 3D homogeneous coordinate, the Euclidean transformation is represented by:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & t_x \\ \sin\theta & \cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{3.13}$$

Note that the translation amount t_x and t_y are multiplied with the homogeneous coordinate, which is one. As a result, the translation is incorporated within the transformation matrix that also encodes the rotation part simultaneously.

One of benefits of using the homogeneous coordinates is to support the translation and rotation in a single matrix. This property addresses problems of the Euclidean transformation (Sec. 3.2.1). Specifically, even though there are many transformations, we can represent each transformation in a single matrix and thus their multiplication is also represented in a single matrix. Furthermore, its inversion can be efficiently performed. Thanks to these properties resulting in a higher performance, the homogeneous coordinates have been widely adopted.

Revisit to mapping from the world space to NDC. We discussed viewport mapping, one of which operation transforms world space coordinates to those in NDC space (Sec. 3.1). Since this transformation uses multiplication, followed by the additions, it can be represented by homogeneous coordinates and thus in a single matrix:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{w.r-w.l} & 0 & -\frac{w.r+w.l}{w.r-w.l} \\ 0 & \frac{2}{w.t-w.b} & -\frac{w.t+w.b}{w.t-w.b} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{3.14}$$

Nonetheless, the matrix is not exactly in the Euclidean transformation, since it involves scaling. This is covered in the affine transformation in the next section.

3.2.3 *Affine Transformation*

We discussed the Euclidean transformation that is a combination of rotation and translation in Sec. 3.2.1. We now study on an affine transformation, which covers wider transformation than the Euclidean transformation.

In the 2D case, the affine transformation has the following matrix representation:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{3.15}$$

The affine transformation preserves parallel lines under the transformation, but does not necessarily preserve angles of lines. The affine transformation covers a combination of rotation, translation, shearing, reflection, scaling, etc. The transformation is also called projective transformation, since it also supports projection, which is discussed in Sec. 4.2.

OpenGL functions. Various transformation functions (e.g., $glTranslate(\cdot)$) available at early versions of OpenGL (e.g., version 2) are deprecated in recent versions. Nonetheless, it is informative to see its usage with corresponding matrix transformations, which are adopted in the recent OpenGL.

The following code snippet shows a display function of rendering a rectangle with a rotation matrix.

```
void display(void)
{
  // we assume the current transformation matrix to be the identify matrix.
   glClear(GL_COLOR_BUFFER_BIT); // initialize the color buffer.

   glPushMatrix(); // store the current matrix, the identify matrix, in the matrix stack
   glRotatef(spin, 0.0, 0.0, 1.0); // create a rotation matrix, M_r.
   glColor3f(1.0, 1.0, 1.0);
   glRectf(-25.0, -25.0, 25.0, 25.0); // create geometry, say, v.
   glPopMatrix(); // go back to the initial identify matrix.

   glFinish (); // send all OpenGL commands to GPU and finish once they are done.

   glutSwapBuffers();
}
```

The actual rasterization done in GPU occurs once $glFinish()$ is called. Before rasterizing the rectangle, we perform the specified transformation, which is to compute v', where $v' = M_r v$. We then rasterize the rectangles with transformed geometry, v'.

3.2.4 *Common Questions*

Is there any benefit of using column-major ordering for the matrix over row-major ordering? Not much. Some people prefers to use column-major, while others like to use row-major. Somehow, people who designed OpenGL may prefer column-major ordering.

3.3 *Affine Frame*

In this chapter, we started with viewport transformation, followed by 2D transformation. Overall, an underlying question along these discussions is this: suppose that we have two different frames and we know coordinates of a point in a frame. What is the coordinates of the point in the different frame? For example, the viewport transformation is an answer to this question with the world and viewport frames.

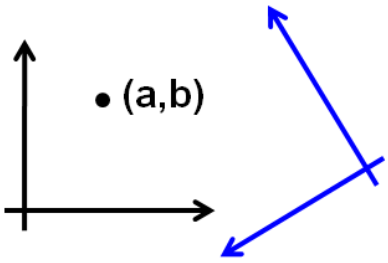


Figure 3.5: What is the coordinate of the point against the blue frame?

We use a set of linearly independent basis vectors to uniquely define a vector. Suppose that $\vec{V}_1, \vec{V}_2, \vec{V}_3$ are to be three such basis vectors represented in a column-wise vector. We can then define a vector, $\vec{X}$, with three different coordinates, c_1, c_2, c_3, as the following:

$$\vec{X} = \sum_{i=1}^{3} c_i \vec{V}_i = \begin{bmatrix} \vec{V}_1 & \vec{V}_2 & \vec{V}_3 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \mathbf{V}c, \tag{3.16}$$

where $\mathbf{V}$ is a 3 by 3 matrix, whose columns corresponds to the basis vectors.

Now let's consider how we can represent a point, $\dot{p}$, in the 3D space. Unfortunately, the point cannot be represented in the same manner as we used for defining a vector in above. To define a point in the space, we need an anchor, i.e., origin, of the coordinate system. This makes the main difference between points and vectors. Specifically, points are absolute locations, while vectors are relative quantity.

A point, $\dot{p}$, is defined with respect to the absolute origin, $\dot{o}$, as the following:

$$\dot{p} = \dot{o} + \sum_{i=1}^{3} c_i \vec{V}_i = \begin{bmatrix} \vec{V}_1 & \vec{V}_2 & \vec{V}_3 & \dot{o} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ 1 \end{bmatrix}. \tag{3.17}$$

Simply speaking, we can define a point by using 4 by 4 matrix $\begin{bmatrix} \vec{V}_1 & \vec{V}_2 & \vec{V}_3 & \dot{o} \end{bmatrix}$, whose each column includes three basis vectors and the origin. As a result, this matrix is also called a affine frame; in this chapter, we will just use a frame for simplicity.

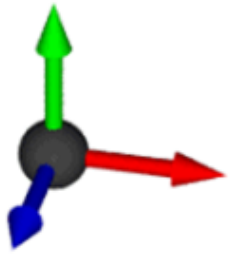


Figure 3.6: The affine frame consisting of three basis vectors and the origin.

We can also define a vector with the frame as the following:

$$\vec{x} = \sum_{i=1}^{3} c_i \vec{V}_i = \begin{bmatrix} \vec{V}_1 & \vec{V}_2 & \vec{V}_3 & o \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ 0 \end{bmatrix}. \tag{3.18}$$

Interestingly, the fourth coordinate for any vector with the frame has 0, since the vector is not based on the origin.

Defining points and vectors with the frame has various benefits. Here are some of them:

1. **Consistent model.** Various operations between points and vectors reflects our intuition. For example, subtracting two points yields a vector and adding a vector to a point produces a point. These operations are consistent with respect to our representations with the frame:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ 1 \end{bmatrix} - \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ 1 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ 0 \end{bmatrix}. \tag{3.19}$$

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ 1 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ 0 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ 1 \end{bmatrix}. \tag{3.20}$$

2. **Homogeneous coordinate.** We introduced the homogeneous coordinate to represent the rotation and translation in a single matrix (Sec. 3.2.2). Such homogeneous coordinates are actually defined in the affine frame, and the fourth coordinate indicates whether it represents points or vectors depending on its values.

3. **Affine combinations.** Adding one point to another point does not make sense. Nonetheless, there is a special case that makes sense. Suppose that we add two points with weights of α_1 and α_2, where the sum of those weights to be one, i.e., $\alpha_1 + \alpha_2$. We then have the following equation:

$$\alpha_1 \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ 1 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ 1 \end{bmatrix}. \tag{3.21}$$

Intuitively speaking, this affine combination results in a linear interpolation between those two points. This idea can be also extended to any number of points. One example with three points includes the barycentric coordinate (Sec. 10.2).

3.4 Local and Global Frames

We would like to conclude this section by discussing local and global frames, followed by revisiting the viewport transformation in these frames.

Suppose that you have a point, $\dot{p}$, defined in the affine frame, $\mathbf{W}$, with a coordinate of c; i.e., $\dot{p} = \mathbf{W}c$. We now want to translate the point with $\mathbf{T}$ and then rotate it with $\mathbf{R}$. The transformed point, $\dot{p}'$, is defined as the following and can be interpreted in two different directions:

$$\begin{aligned}\dot{p}' =& \mathbf{WRT}c \\ =& \mathbf{W}(RTc) = \mathbf{W}c' \text{ // use the global frame} \quad (3.22)\\ =& (\mathbf{WRC})c = \mathbf{W}'c \text{ // use a local frame.} \quad (3.23)\end{aligned}$$

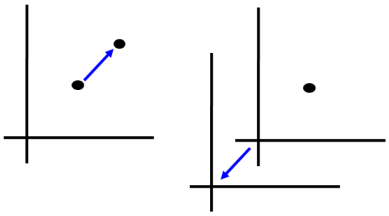


Figure 3.7: Global (left) and local (right) frames of the same transformation.

The second equation is derived by changing the coordinate given the global frame $\mathbf{W}$. The third equation is derived by modifying the frame itself into a new local frame, say $\mathbf{W}'$, while maintaining the coordinate. These two different interpretation can be useful for understanding different transformations.

Let us remind you that we started with this chapter by discussing the viewport transformation. Let's apply local and global frames to the viewport transformation. During the viewport transformation, the point does not move. Instead, we want to compute a coordinate in the viewport space, $\mathbf{V}$, from that in the world space, $\mathbf{W}$. In other words, we can represent them as the following:

$$\dot{p} = \mathbf{W}c = \mathbf{V}c', \quad (3.24)$$

where the relationship between the world and viewport spaces is represented by $\mathbf{V} = \mathbf{WS}$.

In this case, the coordinate c' in the viewport space is computed as the following:

$$\dot{p} = \mathbf{W}c = \mathbf{VS}^{-1}c = \mathbf{V}(\mathbf{S}^{-1}c) = \mathbf{V}c'. \quad (3.25)$$

This approach, considering coordinates with different frames, can be very useful for considering complex transformation. We will use this approach for explaining 3D rotation transformations in the next section.

3.5 3D Modeling Transformation

To create a scene consisting of multiple objects, i.e., models, we need to place those models in a particular place in the world. This

operation is modeling transformation that commonly consists of translation and rotation.

3D translation is extended straightforwardly from the 2D translation:

$$c' = \begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_y \\ 0 & 0 & 0 & 1 \end{bmatrix} c. \tag{3.26}$$

The rotation in the 3D space along the canonical axis is easily extended from the 2D case. For example, the rotation along the X axis is computed as the following:

$$\mathbf{R}_{\grave{\mathbf{X}}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & \sin\theta & 0 \\ 0 & -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3.27}$$

The 3D rotation against an arbitrary vector requires additional treatments. Nonetheless, the affine frame we studied in Sec. 3.2.3 simplifies this process and we thus discuss this approach here in this section.

Suppose that we would like to rotate a vector, $\vec{x}$, given a rotation axis vector, $\vec{a}$. When the rotation axis aligns with one of canonical X, Y, or Z axis, we can easily extend the 2D rotation matrix to 3D rotation matrix. Unfortunately, the rotation axis may not be aligned with those canonical axes, complicating the derivation of the rotation matrix. We now approach this problem in the perspective of the affine frame. The vector $\vec{x}$ can be considered to be defined in the frame of three basis vectors consisting of $\vec{a}$, the red one, and two other orthogonal vectors, the black and green vectors in the figure.

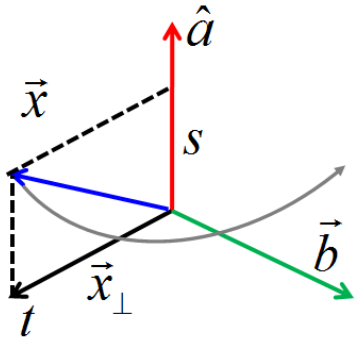


Figure 3.8: Geomtry for 3D rotation.

Let's first compute the black vector $\vec{x}_\perp$, which is orthogonal to $\vec{a}$, and the plane spanned by these two vectors contains the rotation vector $\vec{x}$. We can decompose two coordinates, s and t, of $\vec{x}$ in the plane defined by $\vec{a}$ and $\vec{x}_\perp$, respectively. To compute such coordinates, we can apply the dot product. s and t, and $\vec{x}_\perp$ are then computed by canceling the coordinate of $\vec{a}$, as follows:

$$\begin{aligned} s &= \vec{x} \cdot \vec{a}, \\ \vec{x}_\perp &= \vec{x} - s\vec{a}, \\ t &= \vec{x} \cdot \vec{x}_\perp. \end{aligned}$$

The green vector $\vec{b}$ that is orthogonal to both $\vec{a}$ and $\vec{x}_\perp$ is computed by the cross product between $\vec{a}$ and $\vec{x}_\perp$; i.e., $\vec{b} = \vec{a} \times \vec{x}_\perp$.

So far, we have computed three basis vectors of a local affine frame that can define the vector $\vec{x}$. Specifically, the vector $\vec{x}$ is defined as the

following:

$$\begin{bmatrix} \vec{a} & \vec{x}_\perp & \vec{b} & \dot{o} \end{bmatrix} \begin{bmatrix} s \\ t \\ 0 \\ 0 \end{bmatrix}, \tag{3.28}$$

where $\dot{o}$ is a virtual origin of our local affine frame. The rotation in the amount of θ along the rotation axis $\vec{a}$ is transformed to the rotation along the X axis in the local affine frame. As a result, coordinates of the rotated vector are computed as the following:

$$\begin{bmatrix} \vec{a} & \vec{x}_\perp & \vec{b} & \dot{o} \end{bmatrix} \mathbf{R}_{\grave{\mathbf{X}}} \begin{bmatrix} s \\ t \\ 0 \\ 0 \end{bmatrix}. \tag{3.29}$$

Quaternion is an popular alternative for the 3D rotation, and many tutorials are available for the topic.

4
Camera Setting

In this chapter, we discuss two important aspects of a camera setting: 1) how to setup camera parameters, and 2) how to project objects into a 2D viewing space.

For the simplicity, we discuss these issues with a pinhole camera, one of simple camera setting. Modern cameras employ many different types of lenses and thus are much more complex than the pinhole camera. We also discuss how to extend such realistic cameras in other chapters .

4.1 Viewing Transformation

To see a particular portion of the world scene, it is natural to specify the camera. The camera is specified with its origin, and X, Y, and Z axis in the world space (Fig. 4.1), which define the affine frame of the camera space. The viewable image is then mapped to the X-Y space in the camera space. As a result, the goal of the viewing transformation is to convert the coordinates defined in the world space into those in the camera or viewing space.

Unfortunately, defining those parameters, e.g., X-axis of the camera, in the world space is neither an intuitive nor easy task. Instead, we would like to design an intuitive and easy way of defining those parameters. Following quantities are commonly adopted ones for defining the viewing space:

1. **Eye point,** e**.** This is simply the position of the camera.

2. **Look-at point,** p**.** We typically have a specific target that we want to look at. As a result, requiring such a look-at point is not a big burden to users.

3. **Up-vector,** $\vec{u}_a$**.** While we have the look-at point, the orientation of the camera is not specified. For example, we can look at the target point, while we maintain our head upward or downward.

Figure 4.1: To generate an image, we specify a camera in the world space, which consists of the origin and X, Y, and Z axis of the camera.

As a result, we require to specify an up-vector, $\vec{u}_a$, that define the orientation of the camera.

While we prepared an intuitive way of defining the camera, we still need to define the affine frame of the viewing space. The next goal is to define the affine frame from these parameters, as the following:

1. **Look-at vector, $\vec{l}$.** The Z-direction of the camera can be computed by computing the look-at vector, $\vec{l}$, which is computed by $p - e$ with a proper normalization, $\hat{l} = \frac{l}{|l|}$. Note that we use the hat notation, $\hat{}$, to denote a normalized vector, whose magnitude is one.

2. **Right vector, $\vec{r}$.** The X-axis of the camera is computed by the cross product between the look-at vector $\hat{l}$ and the given up-vector $vecu_a$:

r

$$\begin{aligned} \vec{r} &= \vec{l} \times \vec{u}_a, \\ \hat{r} &= \frac{\vec{r}}{|r|}. \end{aligned} \tag{4.1}$$

3. **Adjusted up-vector, $\hat{u}$.**

   The given up-vector may not be perpendicular to the look-at and right vectors. As a result, we recompute a new up-vector, $\hat{u}$, that is perpendicular to both of them: $\hat{u} = \hat{r} \times \hat{l}$. Since it is difficult and cumbersome for users to specify the initial up-vector in this way, we adjust the up-vector in this way. Usually, this process is performed within a graphics library such as OpenGL.

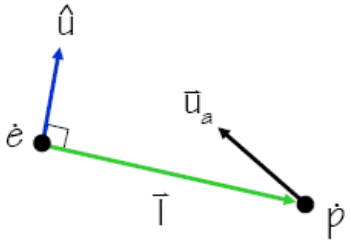


Figure 4.2: Adjusting the initial up-vector.

Let's consider how to transform coordinates in the world space to the viewing space defined in the camera space. This problem is exactly same one that we discussed for local and global frames of Sec. 3.4. As a result, we apply the concept of changing frames to this problem.

Suppose that the coordinate in the world space is c. What we want is to translate the camera origin such that the camera origin becomes the origin in the viewing space. This is represented by $\mathbf{T}_{-\mathbf{e}}$. We then rotate the coordinate with a rotation matrix, $\mathbf{R_v}$, into the camera space. As a result, we have the following equation:

$$\mathbf{W}c = \mathbf{E}\mathbf{R_v}\mathbf{T}_{-\mathbf{e}}c, \tag{4.2}$$

where $\mathbf{W}$ and $\mathbf{E}$ are frames of the world space and camera space. Therefore, the viewing matrix is defined as $\mathbf{R_v}\mathbf{T}_{-\mathbf{e}}$ that convert the world space coordinate c into one in the camera space.

For the world space, we use canonical basis vectors and thus $\mathbf{W} = \mathbf{I}$. Also, the viewing space $\mathbf{E}$ is represented by the three basis vectors. As a result, we have the following relationship:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \hat{r} & \hat{u} & -\hat{l} \end{bmatrix} \mathbf{R_v} \tag{4.3}$$

$$\begin{bmatrix} \hat{r} & \hat{u} & -\hat{l} \end{bmatrix}^{-1} = \mathbf{R_v} \tag{4.4}$$

The matrix of $\begin{bmatrix} \hat{r} & \hat{u} & -\hat{l} \end{bmatrix} = M$ is an orthonormal matrix, whose columns are orthogonal to each other and unit normal vectors. In this case, $M^T M = I$ is satisfied and thus M^{-1} can be easily computed by M^T. As a result, the rotation matrix $\mathbf{R_v}$ is computed as the following:

$$\mathbf{R_v} = \begin{bmatrix} \hat{r}^T \\ \hat{u}^T \\ -\hat{l}^T \end{bmatrix} \tag{4.5}$$

Given the rotation matrix and translation matrix, the viewing matrix $\mathbf{V}$ is computed as the following:

$$\mathbf{V} = \mathbf{R_v}\mathbf{T}_{-\mathbf{e}} = \begin{bmatrix} \hat{r}_x & \hat{r}_y & \hat{r}_z & 0 \\ \hat{u}_x & \hat{u}_y & \hat{u}_z & 0 \\ -\hat{l}_x & -\hat{l}_y & -\hat{l}_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & -e_x \\ 0 & 1 & 0 & -e_y \\ 0 & 0 & 1 & -e_z \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4.6}$$

Connections to OpenGL. In an old version of OpenGL, the viewing transformation is setup by calling "gluLookAt $(\cdot)$". This function simply constructs the viewing matrix (Eq. 4.6) and composes it with the current matrix that OpenGL maintains. In a recent OpenGL version, e.g., 3.0, gluLookAt is no longer available, and one needs to maintain their own viewing transformation in a vertex shader . Fortunately, there are many available codes to implement equivalent functions in recent versions of OpenGL.

4.2 *Projection*

Figure 4.3: Orthographic projection.

Projection occurs right after viewing transformation. Projection maps 3D points defined in the camera or eye space into 2D points in the image space. There are two common projection methods: orthographic and perspective projection.

The orthographic projection simply flattens 3D objects into the 2D image space. It preserves parallel lines before and after the projection. It is used for top and side views in various modeling tools (e.g., 3ds Max). It can, however, appear unnatural due to the lack of perspective foreshortening.

In a simplest form, the orthographic projection is defined as the following:

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{4.7}$$

As an additional details to the viewing and projection transformation, we also define a view volume for each camera. Fig. 4.4 shows an example of the view volume for the orthographic projection with related parameters defining the view volume. After the orthographic projection, we map those 3D coordinates into ones in the NDC space (Sec. 3.1).

In this context, the orthographic projection mapping to the NDC space is computed as the following:

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{r-l} & 0 & 0 & \frac{-(r+l)}{r-l} \\ 0 & \frac{2}{t-b} & 0 & \frac{-(t+b)}{t-b} \\ 0 & 0 & \frac{2}{f-n} & \frac{-(f+n)}{f-n} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{4.8}$$

where r, l, t, b, f, n indicates right, left, top, bottom, far, and near, respectively. As a sanity check, when we have a coordinate of $(l, 0, 0, 1)$, it should give us -1 after the orthographic projection. This is verified as the following:

$$x'(l) = \frac{2l}{r-l} - \frac{r+l}{r-l} = -\frac{r-l}{r-l} = -1. \tag{4.9}$$

Note that we do not cancel the Z-coordinate even after the orthographic projection. We actually use the Z-coordinate for an important rendering task, visibility check using the depth buffer (Ch. 7.4).

4.2.1 *Perspective Projection*

Perspective projection is very common in modern computer animation. It, however, takes a long history to be fully understood and

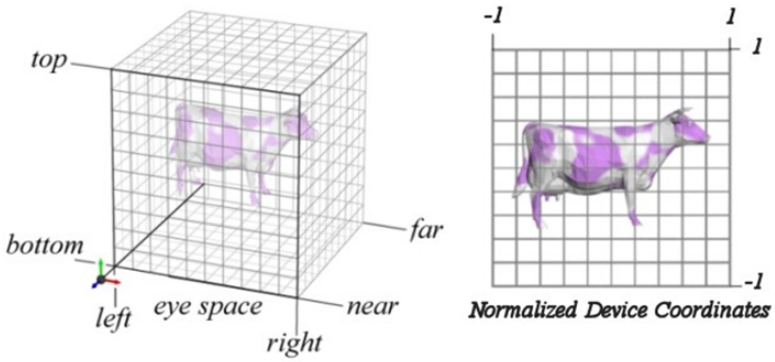


Figure 4.4: The left figure shows a view volume for the orthographic projection. After the orthographic projection, we map 3D coordinates into ones in the NDC space.

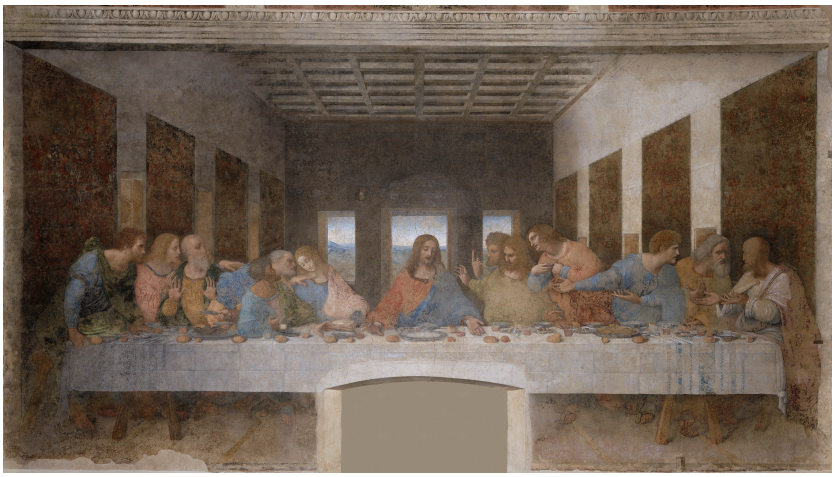


Figure 4.5: This shows the last supper drawn by Leonardo da Vince. This painting shows that objects are drawn under the perspective projection. Furthermore, the vanishing point is located at the Jesus to emphasize the theme of the paining. In other words, perspective projection may be intentionally used for artistic expression.

used in arts. Fig. 4.5 shows an early example of a painting adopting the perspective projection and its intentional use for artistic expression.

A key characteristic of the perspective projection is foreshortening of far-away objects compared to close objects. Another characteristic of perspective projection is that parallel lines in perspective projection always intersect at a point, i.e., vanishing point.

In this section, we discuss such a perspective projection under a simplistic camera model, pinhole camera. Fig. 4.7 shows a 2D schematic illustration of a point into a view plane under a pinhole camera. The point, p, has (y, z) coordinates in the Y-Z world space. Under the pinhole, we can see the point by observing on the ray that is reflected from the point and passes through the pinhole. We draw the ray in the blue color.

In a camera we commonly have some kind of sensors (e.g., camera sensors or film) to capture the light energy that the ray carries at the

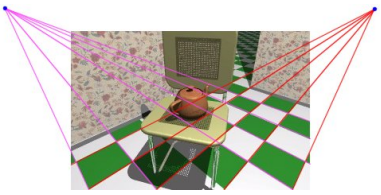


Figure 4.6: Vanishing points.

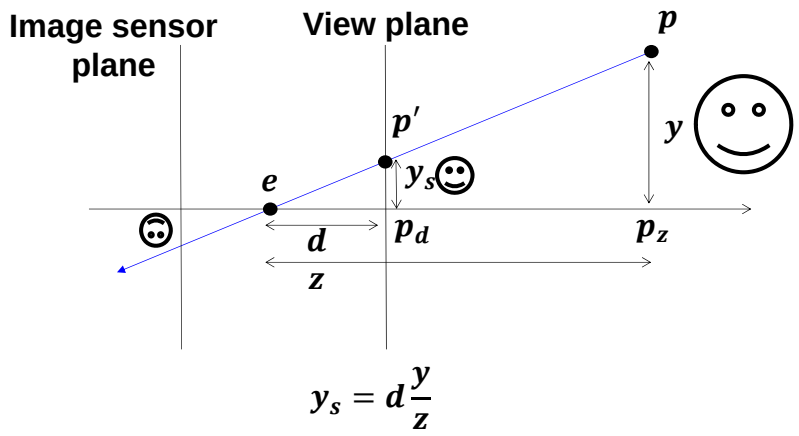


Figure 4.7: This figure illustrates how a point maps in the world space maps to one in the view plane space.

end of the optical systems behind of the eye point, i.e., focal point. In computer graphics, we, however, have such image recording plane in front of the eye position, i.e., the camera center.

Given this configuration of the view plane, our goal is to compute coordinates of the point, p', in the view plane that is projected from the 3D point p. Since the projected point p' is in the view plane, its Z-coordinate is d, which is the distance from the camera origin e to the view plane. The unknown of p' is its Y-coordinate.

To derive this, we apply properties of similar triangles between $\triangle p'ep_d$ and $\triangle pep_z$, and we then have the following relationship based on the same proportion of same sides:

$$\frac{y_s}{d} = \frac{y}{z} \Rightarrow y_s = d\frac{y}{z}, \tag{4.10}$$

where d and z are Z-coordinates of points p_d and p_z, respectively.

The next question is how to represent this equation in a matrix form. The bottom line is how to represent $\frac{1}{z}$ in a matrix form. We address this problem by utilizing the homogeneous coordinate with the following simple matrix form:

$$\begin{bmatrix} wx' \\ wy' \\ wz' \\ w \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{4.11}$$

The trick is on the homogeneous coordinate. In this case, the homogeneous coordinate after applying the perspective matrix is set to the depth of the point, i.e., $w = z$. We then have the following the homogeneous divide and accomplish the perspective projection:

$$w = z, x' = \frac{x}{w} = \frac{x}{z}, y' = \frac{y}{w} = \frac{y}{z}, z' = 0. \tag{4.12}$$

The final homogeneous coordinate after the homogeneous divide is $1 = \frac{w}{w}$.

We also define a view volume for the perspective projection and convert it to the unit view volume, followed by mapping to the NDC space. Based on this, we can also setup a perspective projection matrix for the NDC space. In an old OpenGL version, this function is supported by call $glFrustum(\cdot)$ or $gluPerspective(\cdot)$.

4.2.2 Common Questions

Can we support other projections than orthographic and perspective projections? For example, a projection simulating the image observed from bug's eyes? What if this projection is not represented as a simple matrix? Yes, we can support many other projections that are represented as some mathematical equations. Also, current GPU can support arbitrary projections although the projection is not represented as a simple matrix.

I felt that there are something missed in the image generated by using perspective projection. Then, I realized that those images do not have effects like out-of-focusing and in-focusing. How can we support these effects? To correctly simulate these kinds of effects, we need to simulate a lens that we are using in camera. This can be supported by using ray tracing, but may take long computation time. Instead, we can mimic similar effects by considering depth values of rasterized objects. For example, the depth values of the rasterized objects are far away from the user-defined focal depth, we blur the image of the object. This is not a correct solution, but a hacky solution that can run very fast in the rasterzation rendering mode.

5
Interaction

In this chapter, we discuss basic ways of interacting with 3D objects. We first discuss a file format of 3D objects (Sec. 5.1), and how to select and manipulate those objects (Sec. 5.2). We then discuss a simple way of supporting 3D rotation based on a concept of the virtual trackball (Sec. 5.3), followed by handling hierarchically defined models (Sec. 5.4).

5.1 Loading Objects

One can create a 3D object using various modeling tools such as Blender, a free and open-source software, and Autodesk 3ds Max, a commercial tool. Also, many 3D models have been created and available commercially and freely at various websites. As a result, it is also common to load those models and compose a 3D scene with them.

As a step to compose and render such a scene, it is necessary to read and load 3D objects. Many file formats are proposed to enable such operations easily. In this section, we discuss an obj format, one of simplest and widely available formats. A simple example of an obj file format is shown in Frame 5.1.

```
# A simple cube in an obj file format          // strings starting
with # are comments
   v 1 1 1                                     // vertex specification
   v 1 1 -1
   v 1 -1 1
   v 1 -1 -1
   v -1 1 1
   v -1 1 -1
   v -1 -1 1
   v -1 -1 -1
```

```
f 1 3 4                                  // face specification
f 5 6 8
f 1 2 6
f 3 7 8
f 1 5 7
f 2 4 8
```

Basic obj file tokens are explained in below:

- **# comments.** The rest of the line starting with # is comment.
- **v float float float.** It specifies X, Y, and Z coordinates of a vertex.
- **vn float float float.** It defines a normal.
- **vt float float.** It specifies U and V texture coordinates.
- **f int int int ..** It defines a triangle (or other polygon) with vertices with specified indices. These arguments are 1-based indices. When we do not have normal information associated with the triangle, we compute the normal out of the plane passing the triangle. The direction, i.e., inward or outward, of the normal vector depends on the ordering of those vertices (Ch. 6.3). As a result, an extra attention is required on the ordering of vertices.

We can also read and store these files in an ASCII mode or binary mode. It is usually more intuitive for human to use the ASCII mode, since we can effective understand what the file describes. On the other hand, the binary mode has benefits in terms of compact storage and thus fast I/O operations.

Layouts. One can have an arbitrary ordering, i.e., layout, of vertices and triangles. Nonetheless, the layout has been identified to play an important role in terms of performance. Since modern computer architectures adopt a block-based cache, the cache fetches a block containing consecutively located data, when one of those data is accessed. As a result, data that are likely to be accessed together are recommended to be stored closely. This idea leads to cache-coherent and cache-oblivious layouts [1]. .

[1] Sung-Eui Yoon, Peter Lindstrom, Valerio Pascucci, and Dinesh Manocha. Cache-Oblivious Mesh Layouts. *ACM Transactions on Graphics (SIGGRAPH)*, 24 (3):886–893, 2005

5.2 *Selection*

To interact with objects in graphics applications, we first need a way of selecting a particular object in the 3D scene. Suppose that we would like to select an object that the mouse pointer is locating at.

Many possible approaches are possible, and two of them are listed here:

1. **Object-space approach.** Given the point of the mouse cursor, we can imagine a virtual ray passing from the camera origin to the point. We can then identify objects that are intersecting objects and choose the object that has the closest intersection point to the viewer. Overall, this approach is ray casting, which is the basis for ray tracing, a critical component of physically-based rendering (Ch. 10).

2. **Image-space approach.** Since we have a rendered image in the color buffer, we can directly access the pixel in the buffer, where the mouse cursor is located at. Unfortunately, the pixel has only the color of a rendered triangle, not the ID of the triangle. We explain a concept of an item buffer that encodes the ID of each triangle based on a color. This approach unlike the object space method based on ray tracing works on the image buffer and thus is an image-space approach.

It is worthwhile to mention that many graphics problems can be approached in either the object-space, image-space, or even a hybrid approach combining both of them, as described for the selection problem.

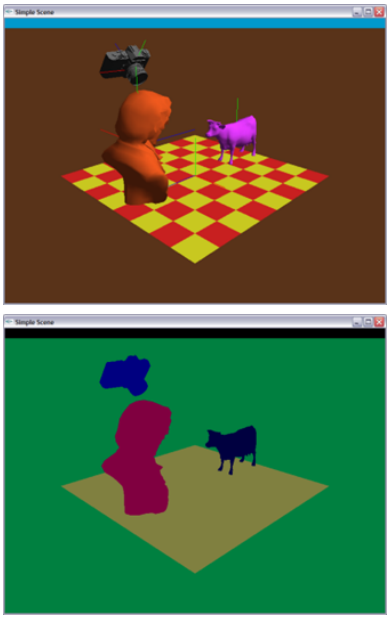


Figure 5.1: The top image shows the color buffer of a scene, while the bottom image shows its item buffer rendered in the back buffer.

5.2.1 *Selection with Item Buffer*

For the selection problem mentioned in the prior section, we want to encode a triangle ID on each pixel on a buffer.

A simple way given the rendering pipeline is to use the concept of the item buffer. The item buffer is simply a different name to the color buffer, with the difference of encoding IDs of triangles, not the original colors of them. To encode an ID for each triangle, we use a unique RGB color value, ID color, for each triangle or each object that serves a smallest selection granularity.

We render all the objects with those ID colors, but we should not show this result to a viewer, since this is not the final result. Therefore, we render it to a back buffer, but do not swap it to the front buffer that is accessed by a display device and thus visible to the viewer. We then read the back buffer by calling an appropriate access function, e.g., $glReadPixels(\cdot)$. Once we fetch the color ID given the chosen pixel, we can identify its associate triangle or object. We then provide a feedback based on the selection operation and render the scene with its original colors.

Note that this selection method works by reading the buffer and thus is categorized by an image space approach. As a result, this

method shows common features of many image space approaches, and some of them are them are:

Accuracy of image-space methods are commonly controlled by the chosen resolution of images.

1. Accuracy depending on the image resolution. A main characteristic of the image-space method is that its accuracy depends on the chosen image resolution, since we identify the triangle ID by accessing an pixel in the item buffer. When we have multiple triangles in a pixel, the pixel can encode only a single triangle. As a result, as we have higher resolutions, we have a higher accuracy in terms of identifying a chosen triangle. Note that when we choose a triangle based on a ray in the object-space method, we do not have such a characteristic.
2. Different performance characteristics to the object space approach. While the image-space method has its accuracy issue, it is commonly used in many different problems including the selection problem, since it is relatively easy to implement and to show high performance, mainly thanks to the support from GPUs. For example, we render triangles and read the buffer through GPU, and thus they can be done quite quickly. Nonetheless, it is less obvious whether this approach has a better time complexity. Specifically, the image-space method using the item buffer explained in this section has a linear time complexity, while the ray tracing based approach using an acceleration structure such as bounding volume hierarchy has sub-linear complexity (Ch. 10.3). As a result, when we have many objects and triangles, the object-space approach can be faster.

In this section, we studied about an image-space selection method using the item buffer. More importantly, we discussed its different characteristics with those of an object-space method using ray tracing.

Figure 5.2: A trackball. The image is excerpted from the homepage of its vendor, Kensington.

5.3 *Virtual Trackball*

In the prior section, we discussed how to pick an object. Once we select an object, we would like to re-position or re-orient the object. For such operations, we can do that through many input devices such as keyboard, mouse, touch screen, etc. For example, many modeling tools (e.g., Autodesk 3ds Max) provide various object and camera manipulations through mouse, which is an inexpensive and widely used input device.

Discussing various interaction operations with available input devices is beyond the scope of this section. Instead, we focus on how to rotate an object in a 3D space. Fig. 5.2 shows a trackball, where a rolling ball is attached. We can use the trackball to intuitively rotate

an object, which is mapped to the ball on the track ball. Unfortunately, the trackball is not widely available compared to keyboard and mouse. We now see how we can support such convenient rotation mechanisms with a mouse.

Suppose that we enclose a sphere on an object that we would like to rotate. Fig. 5.3-top image shows such a configuration. The 2D grid represents our viewing plane. The interaction scenario for rotating the object with the mouse works as the following: 1) the user locates the mouse cursor and clicks a button at a point, $\vec{a}$ [2], and then move and specify the cursor into a different position, $\vec{b}$. Basically, based on this interaction scheme, we want to roll the ball from $\vec{a}$ to $\vec{b}$. The next question is how to compute rotation information, the rotation axis, $\vec{r}$, and its rotation amount, θ, realizing the rolling operation?

Suppose that you grasp the ball from $\vec{a}$ to $\vec{b}$ in your right (or left) hand. The thumb in this case indicates the rotation axis $\vec{r}$. The rotation axis is a vector orthogonal to both $\vec{a}$ and $\vec{b}$, and this can be computed by the cross product between them:

$$\hat{r} = \hat{a} \times \hat{b}, \tag{5.1}$$

where $\hat{r}$ represents a normalized vector whose magnitude is one, i.e., $\hat{r} = \frac{\vec{r}}{|\vec{r}|}$. The rotation angle θ is computed by the inverse of the dot product:

$$\theta = \cos^{-1}(\hat{a} \cdot \hat{b}). \tag{5.2}$$

If necessary, we can also compute a rotation matrix based on computed axis and angle (Sec. 3.5).

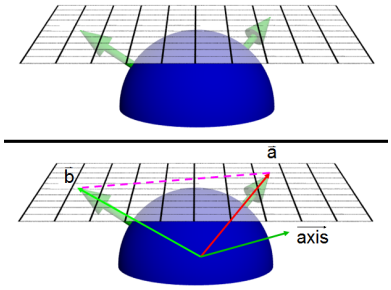


Figure 5.3: Top: we place a ball on an object under the rotation, while the viewing space is touching the ball. Bottom: suppose that we push a button of a mouse at the location of $\vec{a}$ and release it at the another location, $\vec{b}$. In this user input, we want to rotate the ball and its enclosed object from $\vec{a}$ to $\vec{b}$.

[2] We represent this as a vector starting from the ball origin to the point.

5.4 *Transformation Hierarchy*

Some objects have many joints (Fig. 5.4), and we can move each joint independently. In this section, we would like to compute transformations for those parts of an object.

As an example for the study, let's consider an object consisting of two parts with a joint (the rightmost object in Fig. 5.4). Each part is usually defined in its own modeling coordinate; its center is commonly located at the origin, say $(0, 0, 0)$, in its modeling coordinate. We then apply appropriate transformations to those parts to locate it at the world space. Since these parts are defined hierarchically, these transformations are also defined in the same, hierarchical way.

Suppose that $\mathbf{M_b}$ and $\mathbf{M_p}$ are two transformation matrices that converts from the base to the world, and from the part to the base, respectively. In this context, to compute the base, say, its coordinate b in its modeling space, in the world, we compute such transformed locations based on $\mathbf{M_b}b$. For the part, p, we need to apply $\mathbf{M_p}$ to

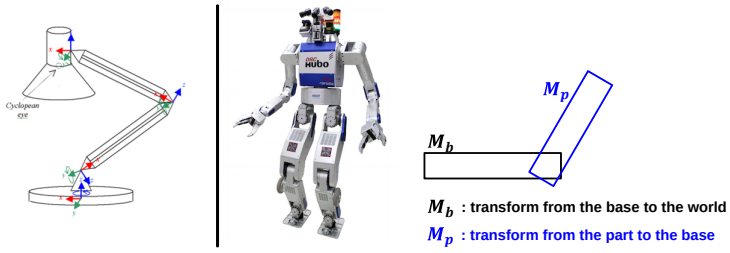


Figure 5.4: The left and middle show two examples of objects with many joints. The left is a lamp with many joints, and the middle shows a humanoid robot, DRC hubo from KAIST, who won DRC (DARPA Robotics Challenge) held at 2015. The right shows an example object consisting of two parts.

locate the part in the base space, followed by $\mathbf{M_b}$ to the world space. As a result, we apply $\mathbf{M_b M_p} p$.

6
Clipping and Culling

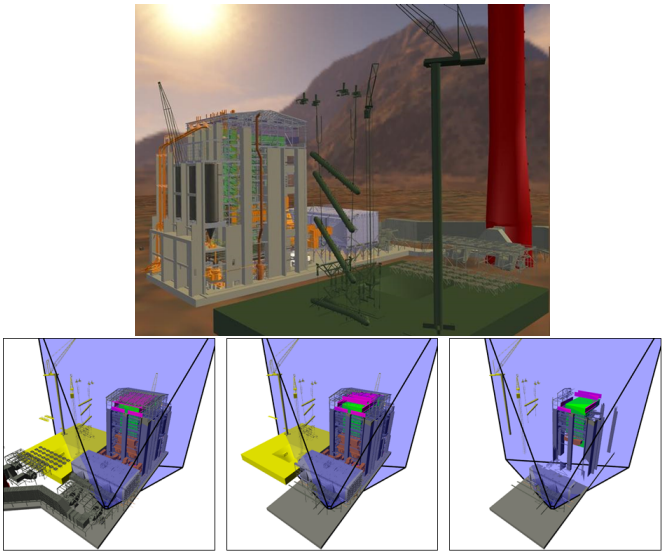


Figure 6.1: The top model shows a coal-fired power plant model consisting of 12 millions of triangles. The model has many pipes within the green, furnace room. It has drastically irregular distributions of triangles across the model ranging from a small bolt to large walls in the furnace. This model is courtesy of an anonymous donor. Bottom images show effects of performing various culling operations. The middle image is the result after performing view-frustum culling to the original power plant model shown in the left. We show these models in a 3rd person view, while the light blue shown in black lines represents the 1st person's view where we perform various culling. The right image shows the result after performing occlusion culling. Since the model has a depth complexity, occlusion culling shows a high culling ratio in this case.

The performance of rasterization linearly increases as we have more triangles. While GPU accelerates the performance of rasterization, it improves only a constant factor, not the time complexity, i.e., growth rate, of the rasterization method. Especially, when we have so many triangles in a scene, it may be prohibitively slow for such scenes. An example includes a power plant scene consisting of 12 millions of triangles (Fig. 6.1).

In this chapter, we discuss two acceleration techniques, clipping and culling, to improve the performance of rasterization. At a high level, their main concepts are:

1. **Culling.** Culling throws away entire primitives (e.g., triangles) and objects that cannot possibly be visible to the user. This is one of important rendering acceleration methods.

2. **Clipping.** Clipping clips off the visible portion of a triangle and throws away the invisible part. This simplifies various parts of the rasterization process.

6.1 *Culling*

Culling conservatively identifies a set of triangles and objects that are invisible to the viewer, and does not pass them to the rendering pipeline. Since the culling process itself can have its own overhead, it is important to design an efficient culling method, while identifying a large portion of invisible triangles among their maximum set.

Fig. 6.1 shows two culling methods, view-frustum culling and occlusion culling, applied to the power plant model. Since this model has a high depth complexity, i.e., many triangles map to a pixel in the screen image, and widely distributed triangles across its scene, such culling methods can be very effective, while they have their own computational overheads. Some of culling methods work as the following:

1. **Back-face culling.** We cannot see triangles heading away from us, unless such triangles are transparent. In opaque models, back-face triangles are blocked by front-face triangles. Back-face culling can be done quite easily and integrated in the rendering pipeline (Sec. 6.5).

2. **View-frustum culling.** The view-frustum (Fig. 6.1) shows an example of the view-frustum and its culling result. Typically, the view-frustum is defined as a canonical view volume within the rendering pipeline and performed by checking whether a triangle or an object is inside the volume or not.

3. **Occlusion culling.** In the case of opaque models, we cannot see triangles located behind the closest triangle to the viewer. As we have more complex models, such models tend to have more numbers of triangles and thus more numbers of triangles map to a single pixel, resulting in a higher depth complexity. In this case, occlusion culling identifies such occluded triangles or objects. Typically, occlusion culling has been more difficult to be adopted, since knowing whether a triangle is occluded or not may require rasterizing the triangle, which we wanted to avoid initially through occlusion culling.

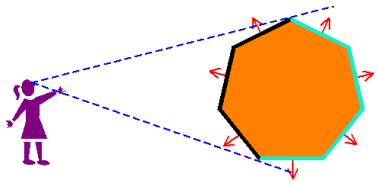


Figure 6.2: Back-face triangles of closed objects are invisible, and back-face culling aims to cull such triangles.

In the next section, we discuss inside/outside tests that are basis for many culling and clipping methods.

6.2 *Inside/Outside Tests*

Many culling and clipping methods check whether a point (or other primitives) is inside or outside against a line in 2D or a plane in 3D. We thus start with a definition of a line for the sake of simplicity; the discussion with the line naturally extends to 3D or other dimensions.

Among many alternatives on definitions on lines, we use the following implicit line representation:

$$\begin{aligned}(n_x, n_y)\cdot(x,y) - d = 0 \rightarrow \\ n_x x + n_y y - d = 0 \rightarrow \\ \begin{bmatrix} n_x & n_y & -d \end{bmatrix}\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = 0 \rightarrow \\ \bar{l}\dot{p} = 0,\end{aligned} \tag{6.1}$$

where $(n_x, n_y) \equiv \vec{n}$ is a unit normal vector of the line equation and $\dot{p}$ is a point in the homogeneous coordinate. We use $\bar{l}$ to denote coefficients of the line.

Given the line equation, we also define the positive half space, $\dot{p}^+$, where $\bar{l}(\dot{p}^+) \equiv \bar{l}\dot{p}^+ > 0$; we also define the negative half space in a similar way. We use the following lemma for explaining culling techniques.

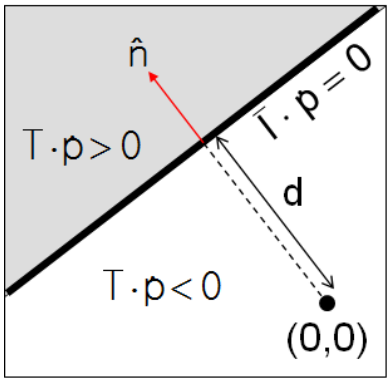


Figure 6.3: Notations of the implicit line equation.

Lemma 6.2.1. *When the normal of the line equation, Eq. 6.1, is a unit normal vector, d gives the L2 distance from the origin of the coordinate system to the line.*

Proof. Let us define (x, y) to be the point in the line realizing the minimum $L2$ distance from the origin to the line, and we then have the following equation:

$$\begin{aligned}(n_x, n_y) = s(x,y), \\ n_x^2 + n_y^2 = 1, \\ s^2(x^2+y^2) = 1.\end{aligned} \tag{6.2}$$

where s is a non-zero constant. Since the point (x, y) is in the line, we have the following equation:

$$\begin{aligned}n_x x + n_y y = d, \\ d = \frac{1}{s}\left(n_x^2 + n_y^2\right) = \frac{1}{s}, \\ = \sqrt{x^2+y^2} \because \text{Eq. 6.2.}\end{aligned} \tag{6.3}$$

□

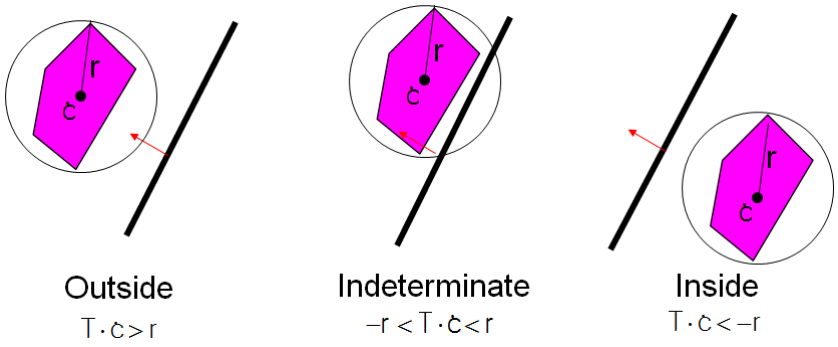


Figure 6.4: This shows three different cases of culling the polygon given its enclosing spherical bounding volume.

In a similar way of proving the lemma, we can see that given a point (x, y) that is or is not on a line whose normal is (n_x, n_y), $n_x x + n_y y$ gives a distance from the point to the line. We also utilize this property for designing culling techniques.

6.3 *View-Frustum and Back-Face Culling*

Let us discuss a simple culling scenario against a line before moving to view-frustum and back-face culling. Suppose that we have a polygon, and we can cull it when the polygon is located totally outside a culling line, as shown in Fig. 6.4. Since it takes a high culling overhead against each vertex of the polygon with many vertices against the line, we use a bounding volume that tightly encloses the polygon.

There are many different types of bounding volumes (BVs) including spheres, boxes, oriented boxes, etc. Commonly, spheres and axis-aligned bounding boxes (AABBs) are frequently chosen bounding volumes, since they are easy to compute with a low computational overhead and a reasonably high culling ratio. Detailed discussions are available in the chapter of bounding volumes and bounding volume hierarchy for ray tracing (Sec. 10.3). In this section, we simply use the sphere for the sake of clear explanation.

Suppose that we use a sphere enclosing the polygon. As a simple culling method in this case, we use its center, c, and radius, r, irrespective of how many vertices the polygon has. Specifically, we test the center against a culling line, $l(\dot{p})$, by plugging its center position to its implicit line equation. There are three different cases (Fig. 6.4), depending on the value of $l(c)$. Since we assume to cull the polygon when it is located outside the line, we focus on this case only in this chapter.

The value of $l(c)$ indicates the $L2$ distance from the line to the center c. When $l(c) > r$ indicates that the sphere is conservatively

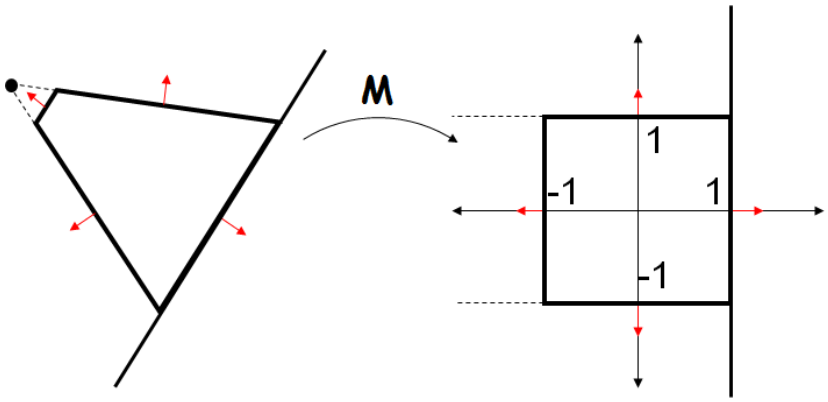


Figure 6.5: The left image shows a view-frustum in 2D, while the right image shows its canonical view volumes. These lines in 2D and planes in 3D of the canonical view are compactly represented and thus can result in fast runtime performance.

outside the line, we can cull it from the further rendering process. With the provided information, it is unclear whether this simple culling operations results in a higher rendering performance than naively rendering all those objects. Nonetheless, we have discussed a basic concept of culling against a line. The performance of this basic approach can be significantly improved by using hierarchically computed bounding volumes, known as bounding volume hierarchy (Sec. 10.3).

Let's see how we perform the view-frustum culling. In rasterization, we assume that we see objects only located within the view-frustum, while this is not the case in reality [1]. Based on this assumption, we can safely cull triangles located outside the view-frustum.

The view-frustum is defined as the left image of Fig. 6.5. We can define such planes with the implicit plane equations, but the view-frustum defined by the given camera setting is transformed to the canonical view volume, which are defined as $x = \pm 1, y = \pm 1, z = \pm 1$, as mentioned in Sec 4.2. The right image of Fig. 6.5 shows the canonical view-volume in 2D.

When a triangle is located outside either one of these six planes, we cull the triangle. This operation applies to each triangle, and is adopted in the rendering pipeline. For large-scale scenes where the view-frustum contains only a portion of them, we can apply the culling method in a hierarchical manner by using a hierarchical acceleration data structure such as bounding volume hierarchy. This approach is more involved and thus a rendering engine supports it.

Back-face culling can be done in a different way. In this section, we discuss a method utilizing the inside/outside tests. One can observe that we cannot see a triangle, when it faces backward (Fig. 6.2). More specifically, suppose that we compute a plane passing the triangle. Then, the triangle is classified as the back-face, when the eye is

[1] We can see other objects that are reflected by objects (e.g., mirror) located within the view-frustum.

located in the negative half side of the plane.

To compute such a plane, we need a normal, the orthogonal vector heading outward to the plane. Given a vertex ordering from v_0, v_1, v_2 in the counter-clock wise, the normal of the triangle, $\vec{n}$, and the distance, d, of the plane is computed as the following:

$$\begin{aligned} \vec{n} &= (v_1 - v_0) \times (v_2 - v_0), \\ d &= \vec{n} \cdot v_0, \end{aligned} \tag{6.4}$$

where the dot product computes the projected distance of the vertex v_0 to the normal direction.

Later, in Ch. 7.3, we discuss a faster back-face culling method, which is more appropriate to be adopted in the rendering pipeline.

Back-face culling in OpenGL. To cull back facing triangles in OpenGL, we use $glCullFace(\cdot)$ after enabling the feature (e.g., GL_CULL_FACE). OpenGL identifies back-face or front-face based on its normal computed from its vertex ordering (Ch. 7.3). OpenGL also provides a way of defining back-face and front-face based on a winding order of vertices between clockwise or counter-clockwise. The counter-clockwise ordering indicates that when we wrap those vertices starting from v_0, passing v_1 to v_2 with the hand, the thumb direction is the front-face. By culling away such back facing triangles, we can avoid to generate fragments from those triangles, resulting in a higher performance.

6.4 *Clipping*

In this section, we discuss clipping that identifies only a visible portion of a primitive, i.e., triangle, and pass it to the following stage (e.g., rasterization stage) in the rendering pipeline.

Let's first discuss a simple case, clipping a line segment consisting of two points, $\dot{p}_0, \dot{p}_1$, against another line, whose coefficient is represented by $\vec{l}$. Our goal here is to identify the clipping point, $\dot{p}$, that intersects with another line $\vec{l}$. To compute the point, we present $\dot{p}$ with a line parameter, t, as the following:

$$\dot{p} = \dot{p}_0 + t(\dot{p}_1 - \dot{p}_0). \tag{6.5}$$

The point should be in another line and thus $\bar{l}\dot{p} = 0$. We then have the following equation:

$$\begin{aligned} &\bar{l} \cdot (\dot{p}_0 + t(\dot{p}_1 - \dot{p}_0)), \\ &t = \frac{-(\bar{l} \cdot \dot{p}_0)}{\bar{l} \cdot (\dot{p}_1 - \dot{p}_0)}. \end{aligned} \tag{6.6}$$

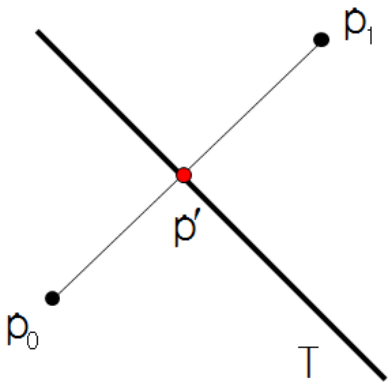


Figure 6.6: A configuration of culling an edge against a line.

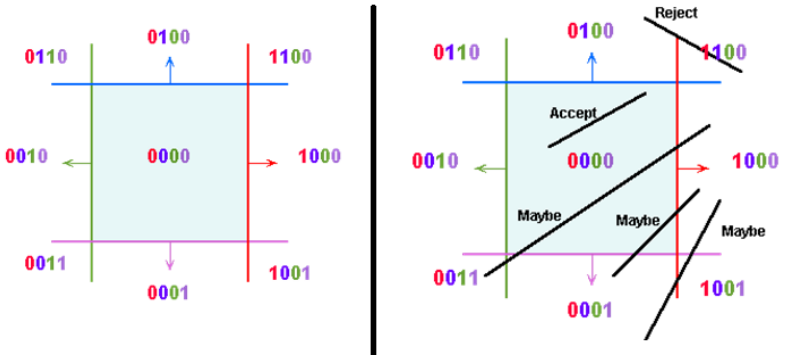


Figure 6.8: The left shows outcodes for each region defined by four lines of the view region. The right shows results of culling edges based on the Cohen-Sutherland method.

Each vertex is also associated with other attributes like colors and texture coordinates. We can also compute those attributes for the clipping point based on the same interpolation method.

Based on this simple line-by-line clipping method, we explain a clipping method, Sutherland-Hodgman algorithm for a polygon including a triangle against a line (e.g., a line of the viewport rectangle) of a convex viewport.

In this method, we traverse each edge of the polygon and check whether the edge is totally inside against the line or not. When it is totally inside or outside, we keep it or throw away it, respectively. Otherwise, we compute two clipping points as shown in Fig. 6.7 and connect them with a new edge. We also apply this process repeatedly against each line of the viewport region.

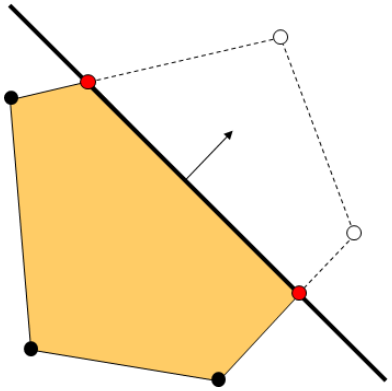


Figure 6.7: The Sutherland-Hodgman method computes a clipped polygon against a line.

6.4.1 *Cohen-Sutherland Clipping Method*

The Cohen-Sutherland method is used to quickly check whether an edge is totally inside or outside given the view region, by using the concept of outcodes. An outcode is assigned to each vertex of primitives, whose each bit encodes whether the vertex is inside or outside against its corresponding line (Fig. 6.8). For example, the first bit in the figure corresponds inside (1) or outside (0) regions against the red line.

When we consider two binary codes, c_1 and c_2, assigned to two vertices of an edge, we have the following conditions and actions:

- If $(c_1 \vee c_2) = 0$, the edge is inside.
- If $(c_1 \wedge c_2) \neq 0$, the edge is totally outside.
- If $(c_1 \wedge c_2) = 0$, the edge potentially crosses the clip region at planes indicated by true bits in $(c_1 \oplus c_2)$. Nonetheless, this could be false positive, meaning that they are identified to be potentially crossing the clip region, but are not actually.

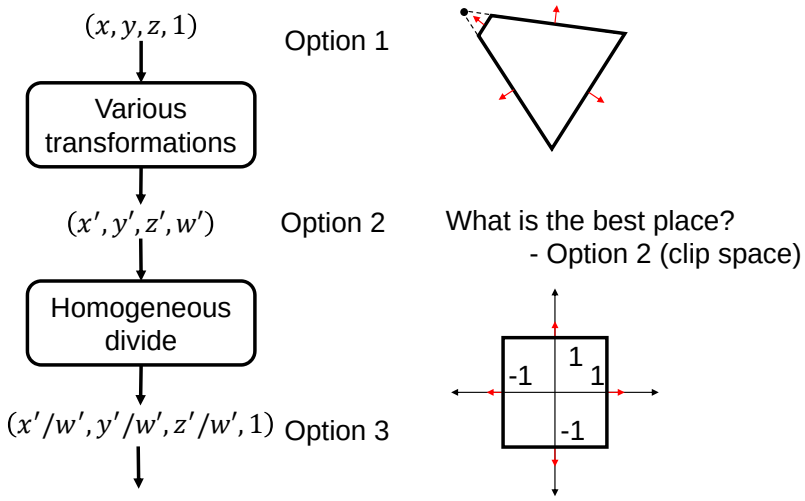


Figure 6.9: This shows different stages of the rendering pipeline with vertex coordinates and view-frustum in each space.

This also applies to a triangle case by utilizing three outcodes computed from three vertices of the triangle.

6.5 *Clipping in the Pipeline*

We discussed how to clip an edge against a plane of the view-frustum before. We would like to now discuss in which stage of the rendering pipeline we perform the clipping operation.

Fig. 6.9 shows how vertex coordinates change as we perform different steps in the rendering pipeline. Overall, there are three different places where we can perform the clipping operation. The first option is the world space where the view-frustum is defined. The second and third options are before and after performing the homogeneous divide.

Each option has its pros and cons. The most intuitive option would be the first one. Also, the third option seems to be good, since the plane equations of the view-frustum in that space are canonical like $x = 1$, and the clipping operation can be done quite quickly. Nonetheless, if we do not clip an edge that spans outside the view-frustum before this stage, the edge flips around due to the projection carried by the homogeneous divide, and generates an unexpected behavior. As a result, the third option is not possible.

Interestingly, the second option has been identified empirically to show the best place to perform the operation, since it does not have the problem of the option three and their plane equations are also defined quite easily. The space of the option two is known as the clip space. Let us discuss how the view-frustum is defined in this

clip space. Specifically, $x'/w' = 1$ in the third space corresponds to $x'/w'w' = w' \rightarrow x' = w'$ in the clip space, which does not depend on the camera setting, and thus can be done efficiently.

[2] As you may realize through this discussion, the rendering pipeline has been heavily tested and optimized to deliver the highest rendering performance. Nonetheless, these choices can change depending on different workloads (e.g., some games use geometry or texture heavily) and hardware performance (e.g., faster memory read or computation).

[2] Structures of the rendering pipeline are not fixed and can be changed for better performance and usability.

6.6 Common Questions

Even though some objects are outside the view frustum, they can be seen though transparent objects or reflected from mirrors. Exactly. The rasterization algorithm is a drastically simplified rendering algorithm over the real interactions between lights and materials. The direct illumination, seen thought primary rays, are well captured by rasterization, while other indirect illuminations are not captured well in the rasterization. To address this problem, many techniques have been proposed in the field of rasterization. However, the most natural way of handling them is to use ray tracing based rendering algorithms.

7
Rasterization

The main idea of rasterization is to project a triangle into the view space and rasterize it into fragments in the color and depth buffers. In this chapter, we assume that vertices of the triangle are projected into the view space, after they undergo various transformations, followed by clipping and NDC transformation.

7.1 Primitive Rasterization

Rasterization is optimized for processing triangles thanks to their simplicity.

For the rasterization process, we commonly use triangles as input primitives, mainly because it is the simplest polygon and simplifies the rasterization process. Nonetheless, these other representations are also decomposed into a set of triangles and fed into the rasterization process.

Rasterization process has two main goals: 1) pixel coverage determination (Fig. 7.1) and 2) parameter interpolation (Fig. 7.5). Given a pixel of the color buffer (or other buffers), we determine whether the pixel belongs to a given triangle or not. Once the pixel is covered by the triangle, we also need to compute its color or other parameters such as its depth value for the depth buffer.

For the coverage problems, many directions are possible. One is to check whether the center of a pixel is inside of a triangle. Another is to measure an area coverage ratio of a pixel against the triangle. The first one is based on a point sample, while the latter one is based on area computation. While the area based computation is more correct, the sample based approach is more efficient, and thus is commonly adopted for rasterization process. They share common pros and cons between point sample based and area based approaches, as we discussed for image-space and object-space methods (Ch. 5.2).

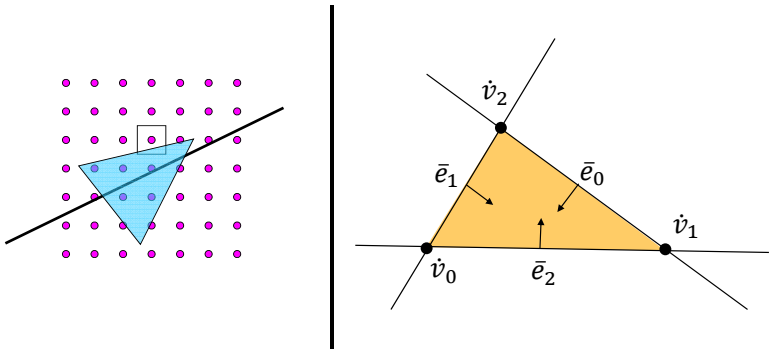


Figure 7.1: The left shows one, pixel coverage determination, of two main goals of the rasterization process. The right shows configurations of vertices and edges of a triangle used for our discussion.

Scanline based triangle rasterization. Some of early techniques for rasterization are based on a concept of scanline, a row of pixels that span a triangle (Fig. 7.2). At those days, the memory was very expensive, and thus having the full resolution of color and depth buffers is not preferred. Instead, these scaneline based approaches maintain a scanline and incrementally update the scanline to raster the whole triangle. Specially, we rasterize an input triangle from top to bottom. Once we meet a vertex of the triangle, we setup the scanline information (e.g., starting and end coordinates shown as red pixels in the figure). For the next scanline, we incrementally update those starting and end coordinates by utilizing slope information of two edges starting from the vertex.

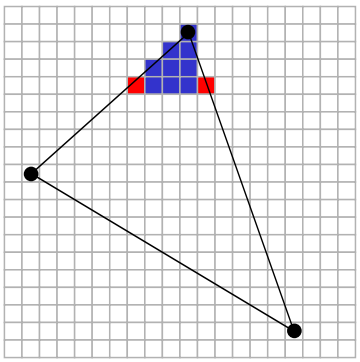


Figure 7.2: This shows a scanline based rasterization. The scanline can be incrementally computed between two neighboring rows.

While this technique was adopted early on, it was identified to show poor scalability to handle scenes with many triangles, since this technique relies on expensive sorting operations and is not friendly for parallelization. Instead, ray tracing and Z-buffer techniques as visible surface determination, i.e., visibility techniques, are prevail techniques in these days (Sec. 10.4).

In the next section, we discuss another rasterization technique combined with the Z-buffer technique.

7.2 *Rasterization with Edge Equations*

In this section, we discuss a rasterization technique for triangles based on edge equations, as shown in the right side of Fig. 7.1. We will see that this approach is simply and friendly for parallelization, to achieve a high performance and thus handle a scene with many triangles.

Let us first compute an edge equation given two vertices, v_0 and v_1, of a triangle (Fig. 7.3). Our goal is to construct an edge equation,

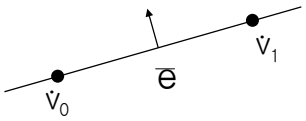


Figure 7.3: An edge representation from two vertices of a triangle.

$\bar{e}$, whose normal vector heads towards the inside of the triangle. Overall, coefficients of the edge equation is given by the cross product between those two vertices:

$$\begin{aligned}\bar{e} &= \dot{v}_0 \times \dot{v}_1 \\ &= \begin{bmatrix} x_0 & y_0 & 1 \end{bmatrix}^t \times \begin{bmatrix} x_1 & y_1 & 1 \end{bmatrix}^t \\ &= \begin{bmatrix} (y_0 - y_1) & (x_1 - x_0) & (x_0 y_1 - x_1 y_0) \end{bmatrix} \\ &= \begin{bmatrix} A & B & C \end{bmatrix}. \end{aligned} \tag{7.1}$$

It is not intuitive to compute the edge equation in this way. Here is the rationale. Think of a line passing $\dot{v}_0$ in the homogeneous space, i.e, $(x_0 w, y_0 w, w)$ with an arbitrary value w. We also think another line passing $\dot{v}_1$. The edge in the 2D space maps to a plane in the 3D homogeneous space. Since these two lines and the plane passes the origin, $(0, 0, 0)$, of the 3D homogeneous space, the normal of the plane, i.e., the edge equation, is computed by the cross product between $\dot{v}_0 - (0, 0, 0)$ and $\dot{v}_1 - (0, 0, 0)$.

Once we set the edge equation $\bar{e}$ in this way, points, $\dot{p}$, inside the triangle have $\bar{e}\dot{p} > 0$. We then see that pixels of the triangle reside in the positive half-spaces against three edge equations from the triangle (Fig. 7.1).

While the aforementioned approach is simple enough to identify which pixels are inside a triangle, there are a few special cases requiring certain treatments. They are two cases of sharing edges and vertices.

Sharing an edge. The left image of Fig. 7.4 shows that a shared edge of two triangles passes the center of a pixel. This case arises rarely, but can happen, since there are many pixels, say 1 M pixels when we use a 1 K by 1 K image resolution. When we assign the pixel to both of those two triangles, the pixel color varies depending on an order of rendering those two triangles, which is not a desirable effect. We thus need a tie-breaker assigning only a single triangle to the pixel.

A simple method is to consider the normal of each edge of a triangle and to assign the pixel to either one of them. For example, we can use the following simple tie-breaker:

$$\text{bool} \quad t = \begin{cases} A > 0 & \text{if } A \neq 0, \\ B > 0 & \text{otherwise,} \end{cases}$$

where (A, B) are the normal vector of an edge computed by Eq 7.1. We then assign a triangle to the pixel, when $(\bar{e}(\dot{p}) > 0) \vee (\bar{e}(\dot{p}) = 0 \wedge t)$.

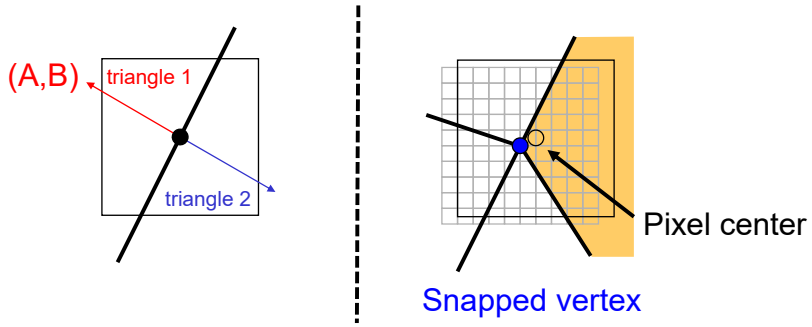


Figure 7.4: This figure shows two cases requiring special treatments for the pixel coverage.

Sharing a vertex. The right image of Fig. 7.4 shows another degenerated case, where a shared vertex of triangles is located at the center of the pixel. For handling this case, one can use a similar tie-breaker that we designed for the shared edge case. Another approach is to snap or quantize vertices of triangles in a way that those snapped or quantized vertices are not aligned with center coordinates of pixels.

7.3 *Interpolation Parameters*

In the last section, we discussed which pixels are covered by a triangle based its edge equations. In this section, we study how to compute colors and other parameters for the pixel, given associated information of the triangle.

Suppose that each vertex has associated information such as color, normal, etc. For the sake of simplicity, we explain various concepts based on the color, especially, red channel information, $r(x, y)$, given a pixel (x, y). Given three red values associated with three vertices of a triangle, we need a way of interpolating these values for a pixel within the triangle. The simplest method is to pick a red value among those three values. While this is simple, it does not produce reasonably high-quality rendering results.

Among many options, we use the linear interpolation from those available values associated with three vertices (Fig. 7.5). The linear red plane is then defined as the following:

$$r(x, y) = A_r x + B_r + C_r, \tag{7.2}$$

where A_r, B_r, C_r are three coefficients of the 2D plane. There are three unknowns and we thus need three equations to compute the plane. Fortunately, these three equations are defined by available

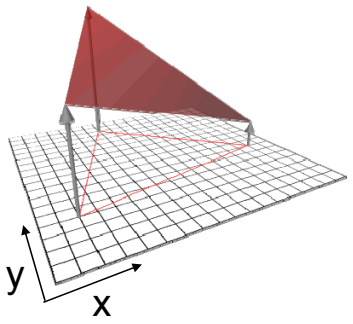


Figure 7.5: This shows the linear interpolation of color values associated with three vertices.

information of three vertices as the following equation:

$$\begin{bmatrix} r_0 & r_1 & r_2 \end{bmatrix} = \begin{bmatrix} A_r & B_r & C_r \end{bmatrix} \begin{bmatrix} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ 1 & 1 & 1 \end{bmatrix} \rightarrow$$

$$\begin{bmatrix} A_r & B_r & C_r \end{bmatrix} = \begin{bmatrix} r_0 & r_1 & r_2 \end{bmatrix} \frac{\begin{bmatrix} (y_1 - y_2) & (x_2 - x_1) & (x_1 y_2 - x_2 y_1) \\ (y_2 - y_0) & (x_0 - x_2) & (x_2 y_0 - x_0 y_2) \\ (y_0 - y_1) & (x_1 - x_0) & (x_0 y_1 - x_1 y_0) \end{bmatrix}}{\det \begin{bmatrix} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ 1 & 1 & 1 \end{bmatrix}}, \tag{7.3}$$

where r_0, r_1, r_2 are three red values associated to corresponding three vertices.

An interesting fact is that the area of the triangle, $A_{\dot{v}_0\dot{v}_1\dot{v}_2}$, is computed as the following by utilizing the determinant of a matrix:

$$\begin{aligned} A_{\dot{v}_0\dot{v}_1\dot{v}_2} =& \frac{1}{2} \det \begin{bmatrix} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ 1 & 1 & 1 \end{bmatrix} && (7.4) \\ =& \frac{1}{2} \left((x_1 y_2 - x_2 y_1) + (x_2 y_0 - x_0 y_2) + (x_0 y_1 - x_1 y_0) \right) \\ =& \frac{1}{2} (C_{20} + C_{12} + C_{01}), && (7.5) \end{aligned}$$

where C_{20}, C_{12}, C_{01} are coefficients of three edge equations, $\bar{e}_{20}, \bar{e}_{12}, \bar{e}_{01}$, respectively; $\bar{e}_{20}$ indicates the edge equation constructed from $\dot{v}_2$ to $\dot{v}_0$.

Note that when the area is zero, the triangle is invisible. Furthermore, when the area is negative, the triangle is back-facing. If the back-face culling is enabled (Ch. 6.3), we cull the triangle for later rasterization. Otherwise, we flip normals of edge equations and perform later rasterization.

Let's consider the interpolation equation (Eq. 7.3). Actually, other components of the top matrix are coefficients of three edge equations! We then have the following interpolation equation:

$$\begin{bmatrix} A_r & B_r & C_r \end{bmatrix} = \frac{1}{2A_{\dot{v}_0\dot{v}_1\dot{v}_2}} \begin{bmatrix} r_0 & r_1 & r_2 \end{bmatrix} \begin{bmatrix} \bar{e}_{20} \\ \bar{e}_{12} \\ \bar{e}_{01} \end{bmatrix}, \tag{7.6}$$

where $\bar{e}_{20}$ represents an 1 by 3 vector containing its three coefficients, A_{20}, B_{20}, C_{20}.

Once we compute coefficients of the red plane (Eq. 7.2), we can compute a color on any pixel within the triangle.

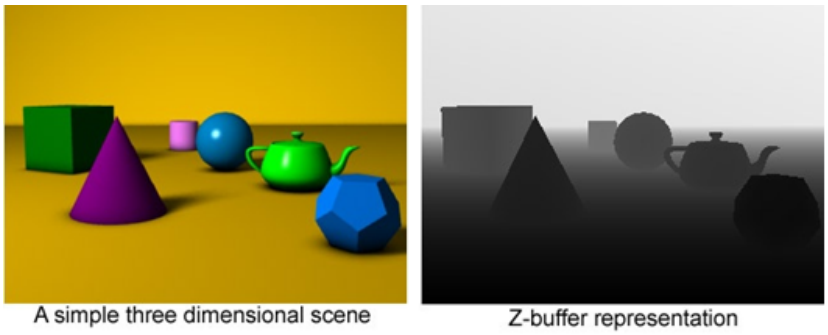


Figure 7.6: The left shows an input scene, while the right image shows its Z-buffer. The white color represents the farthest depth value, 1, while the black one indicates the closest value, 0.

7.4 *Z-Buffering*

The Z-buffer technique is a visibility determination technique, which encodes the depth value of a visible triangle per each pixel. Overall, it is an image-space technique to determine the visible triangle by using a 2D buffer, i.e., depth-buffer.

Z-buffer is one of the most important concepts for rasterization. Simply speaking, we address a complex problem of visibility determination using a 2D map.

Fig. 7.6 visualizes the depth buffer, i.e., Z-buffer, given a scene. The depth buffer simply contains depth values of visible triangles. Note that each vertex of a triangle has its position information (x, y, z). Once we project it to the image space, we also have its depth value in the canonical view volume (Ch. 4.2). The depth value in the canonical view volume spans in the range of $[0, 1]$, where 0 indicates the closest one, while 1 indicates the farthest one.

Given the depth value of each pixel, more correctly, fragment, rasterized from a triangle, we can easily know that whether the fragment has a depth value smaller than the one stored in the depth buffer and thus visible. Once the fragment has a smaller depth value, we update the depth buffer with that depth value at the pixel. We continue this process until we process all the fragments generated from the rasterization process.

As you can see, this Z-buffer is very simple, and thus can be well adopted to a hardware implementation. While there have been many advanced techniques, this Z-buffer technique is the most common technique adopted in rasterization. Nonetheless, recent ray tracing techniques are getting wider attentions thanks to its conceptual simplicity and better functionality supporting realistic rendering effects (Ch. 10).

Processing order. Note that the rasterization method based on the edge equation can be parallelized among different pixels. For example, a rasterization result of a pixel does not depends on anything of another pixel. This opens up various approaches to parallelize the

process for achieving higher performance.

Fig. 7.7 shows two examples of the processing ordering of pixels for rasterizing the triangle. In practice, we identify a bounding box covering the triangle and process the region based on tiles. A tile is a sub-region, say 4 by 4 pixels, of the image space. A GPU core is assigned to process each tile. Different GPU cores process those tiles in a parallel manner, to achieve a high performance. A GPU core assigned to a tile needs to setup three edge equations for a pixel, (x, y), in the tile. For the neighboring pixel, say $(x, y+1)$, we incrementally compute those edge equations, as the following:

$$\begin{aligned} E(x,y) &= Ax + By + C, \\ E(x+1,y) &= A(x+1) + By + c \\ &= E(x,y) + A. \end{aligned} \tag{7.7}$$

So far, we have discussed the rasterization process converting a triangle into a set of fragments. This is one of main concepts of rasterization, setting apart it from ray tracing.

While the rasterization process adopted back-face culling, it can be very slow, especially, when the given scene has so many triangles. There have been many scalable techniques (e.g., mesh simplification) to handle such cases.

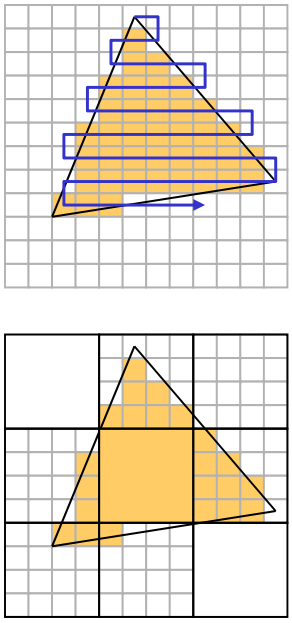


Figure 7.7: Rasterization process can be parallelized, and any ordering of processing pixels or tiles can be possible.

8
Illumination and Shading

In this chapter, we look into basic concepts of illumination and shading. These topics can have different meanings depending on the context. In this chapter, we focus on computing effects of lights for illumination per vertex, followed by applying those illumination results to fill triangles. Before we talk about these concepts, let us first think about how we see things.

8.1 How can we see objects?

Each one of us might have thought about how we can see objects at one point in the past, since seeing things is a part of our daily activity. To fully explain the whole process is beyond the scope of this book. Instead, we would like to point out main components of this process.

At a high level, seeing objects means that we receive the light energy in our eye, which is in the end transferred to our brain. Let us first talk about the light. The light is electromagnetic waves, and our human eyes see only a portion of the spectrum of those electromagnetic waves, commonly called visible light (Fig. 8.1). Visible light refers to wavelengths in a range of 400 and 700 nm.

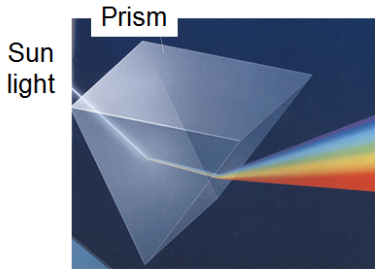


Figure 8.1: This illustrates that the sun light is composed of different wavelengths, which are perceived in different colors. The image is excerpted from the Newton magazine.

Our eye has multiple layers and one of them is retina, which contains photoreceptor cells sensing the visible light. There are mainly two types of such cells: rod and cones. The rod cell is extremely sensitive to photons and can be even triggered by even a single photon. The rod cells give information mainly about intensity of the light, while cone cells are about the color information. There are also three types of cones, each of which responds to different wavelengths, which we call red, green, and blue colors (Fig. 8.2). In reality, color does not exist, but based on response levels from these three cone types, our brain reconstructs the color.

Color does not exist in reality. Instead, our brain reconstructs based on response levels of red, green, blue cones.

Now let's consider how the light interacts with materials. This process can be explained in different levels including quantum physics, but in this chapter, we give only a high level idea on the process.

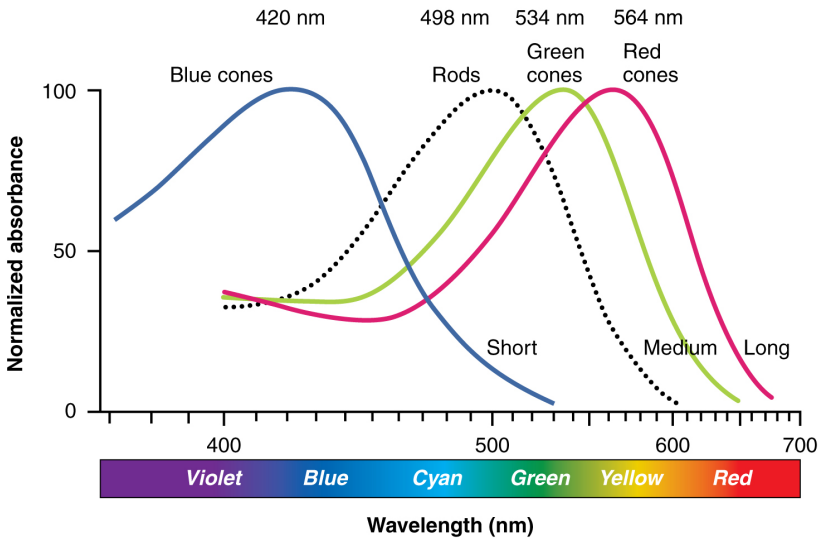


Figure 8.2: This figure shows the response level of rod and cone cells as a function of wavelengths. This figure is available from Anatomy and Physiology under the Creative Commons Attribution 3.0 license.

Once a material or an atom receives a photon, the atom enters into its excited sate. It then returns back to its normal state, while emitting its energy into the space. The energy can be interpreted into another photon or wave, thanks to the duality of the light. The key factor that we need to know is directions of the emitted photon and their wavelengths that determine the perceived color.

As an concrete example, please consider a leaf shown in Fig. 8.3. The incoming sun light is a mix of various electromagnetic waves with a set of different wavelengths, and thus can be perceived as a white light. Once the sun light hits with the leaf, the leaf receives its energy, which is used for its photosynthesis and dissipated as heat. Nonetheless, some of its received energy is emitted into waves (and particles) with different wavelengths. In this leaf case, the emitted energy has the wavelengths corresponding to the green color. As a result, we see the green color to the leaf. Furthermore, the emitted energy is distributed into all the possible directions, and thus we can see the leaf in any directions towards it.

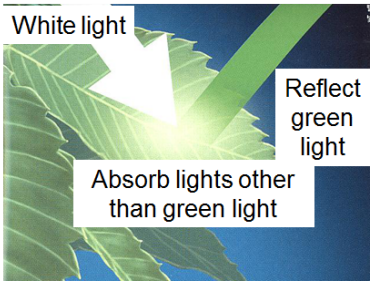


Figure 8.3: This illustrates how the leaf interacts with the coming light. This image is excerpted from the Newton magazine.

8.2 *Bi-Directional Reflectance Distribution Function*

Depending on materials, they have different reflectance behaviors. For example, the chalk is diffuse, meaning that it reflects the light in all possible directions. We thus see the chalk in any view directions. On the other hand, when we look at the surface of an apple, there can be a highlight, a bright spot, when we have a particular view direction. We call such materials to be glossy.

BRDF is used to explain the reflection behavior of a material.

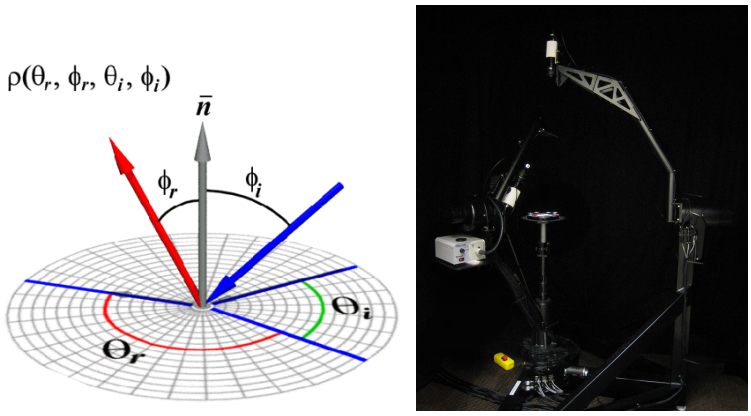


Figure 8.4: The left image shows incoming and outgoing directions, which are four parameters of a BRDF. The right image shows a gonioreflectormeter measuring the BRDF; it is from Univ. of Virginia.

Since materials have different reflectance distributions, we need a function to encode such behaviors. Bi-directional reflectance distribution function (BRDF) is introduced to meet the requirement. BRDF, $f(\cdot)$, is defined over incoming light direction and outgoing light directions. Each direction in the 3D space is encoded with two parameters, θ and ϕ. As a result, BRDF is a four dimensional function (Fig. 8.4). Detailed explanation on BRDF is available at the chapter on radiometric quantities (Ch. 12).

Gonioreflectormeter is used to measure BRDF, by rotating a light source and sensor location (Fig. 8.4). This approach takes very long time, and thus it is one of active research areas to efficiently measure the BRDF.

Measured BRDF itself can be very large in terms of memory footprint. It is thus common to encode and use them in a compact representation. In the following section, we discuss one of most simple illumination models.

8.3 *Phong Illumination Model*

The Phong model is an empirical model, but is used a lot for its simplicity.

The Phong illumination model is a simple and classical illumination model that is adopted in early versions of OpenGL. This model is just empirical, not based on physics, and does not even preserve basic physical assumption such preserving energy. Nonetheless, it has been commonly used thanks to its simplicity.

The Phong illumination model has the following three components:

- **Ambient term.** The ambient term represents a kind of background illumination, and works as a constant value (Fig. 8.5). Specifically, for computing the reflected ambient illumination, $I_{r,a}$, it multiplies an ambient reflectance coefficient, k_a, to an incoming ambient illumination, $I_{i,a}$, i.e., $I_{r,a} = k_a I_{i,a}$. Intuitively, this is a drastic

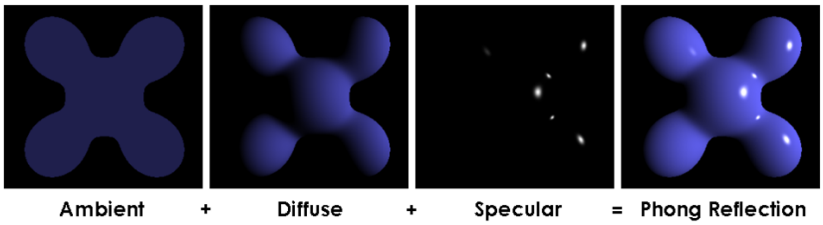


Figure 8.5: This shows different terms of the Phong model. This figure is made by Brad Smith.

simplification of complex inter-reflection between lights and materials. Simulating this term well is a critical component of global illumination, while it is drastically ignored in the Phong model.

- **Diffuse term.** Most objects can be seen in any view directions, and this indicates that they are diffuse. The diffuse term aims to support this visual phenomenon.

- **Specular term.** Certain objects such as metals show strong highlights in a particular viewing direction. The specular term simulates this feature.

Before we discuss diffuse and specular terms in a detailed manner, let's first discuss light sources, which are also mentioned when we explain the ambient term. For the ambient term, we use an ambient light that virtually emits light energy to every location of triangles. We discuss point and direction light sources, followed by briefly mentioning other types of light sources.

Point and directions light sources. The light direction plays an important role on computing illumination. A point light source emits light energy from a single point, p_l. The point light may seem too crude approximation compared to light sources that we encounter in real life. Nonetheless, we can approximate them by using a set of point light sources.

The light direction, $\vec{L}$, on a point, p, on a surface is then computed as the following:

$$\vec{L} = \frac{p_l - p}{|p_l - p|}. \tag{8.1}$$

Note that the light direction varies depending on the location of the surface p.

Unlike the point light, the directional light is located far away from the observer, and thus the light direction is considered as a constant, irrespective of observing locations. The directional light can be thought as a point light source whose location is set at infinity.

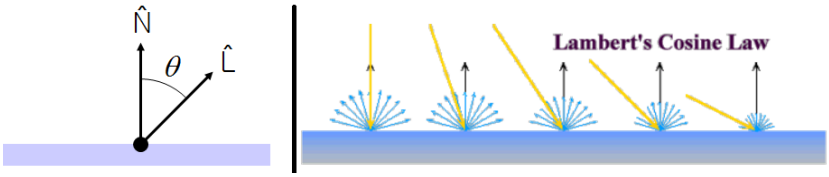


Figure 8.7: The left shows the configuration of the Lambert's cosine law, and its effects are shown in the right.

For example, sun is located far away from us, and thus we use the directional light to represent the sun light source.

Area light sources are a common type of light sources. The area light has a certain light shape with an area and thus can generate soft shadows (Fig. 8.6). Directly considering area lights is more complex than working with point lights. A simple approximation to an area light is to generate a set of point lights on the area light. The number of generated point lights defines illumination levels of soft shadow.

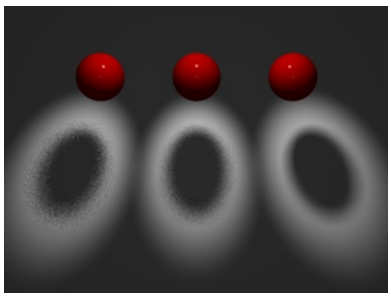


Figure 8.6: Area lights.

Diffuse term. Many objects have the diffuse property, and that is why we can see them! Here we assume the ideal diffuse material; the chalk is close to such a material. The ideal diffuse material reflects an incoming energy into all the possible directions with the same amount of energy. As a result, the reflection becomes view-independent. This diffuse property is caused by a rough surface in a microscopic level, and is perceived as the uniform distribution in the space in the macroscopic level.

The diffuse term is explained by the Lambert's cosine law.

The Lambert's cosine law explains how an incoming energy is reflected depending on the configuration between the surface and the light direction. Suppose that we want to compute the reflected energy I_r on a surface having a normal $\vec{N}$ given the light direction $\vec{L}$. The reflected energy I_r is then computed as the following:

$$\begin{aligned} I_r &= I_i \cos\theta \\ &= I_i(\vec{N} \cdot \vec{L}), \end{aligned} \tag{8.2}$$

where θ is the angle between two vectors of $\vec{L}$ and $\vec{N}$. Fig. 8.7 shows the configuration of these vectors.

Note that as the reflected energy becomes the highest, when the light direction is aligned with the surface normal. Fig. 8.7 also shows how the reflected energy behaves as we have different light directions. When we have a material-dependent, diffuse coefficient, k_d, the reflected energy of the diffuse term is $I_{r,d} = k_d I_i(\vec{N} \cdot \vec{L})$.

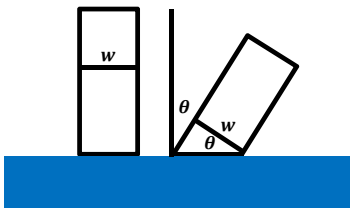


Figure 8.8: The configuration for the Lambert's cosine law.

Proving the cosine law. Let us see how to prove the Lambert's cosine law. Suppose that we have a beam of light with a width of w and energy of I (Fig. 8.8). In this case, the light density per unit area

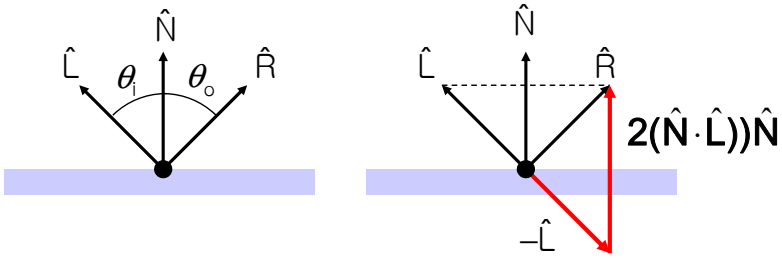


Figure 8.9: This shows how the reflected light direction is computed for the perfect specular object.

is $\frac{I}{w}$. We then lean the beam as an amount of θ. The area receiving the light energy become larger, and takes $\frac{w}{\cos\theta}$. Its light density is then $\frac{I\cos\theta}{w}$. As a result, we can see that the light density reduces as an amount of $\cos\theta$.

Specular term. Let us first consider a perfect mirror-like object. In this case, the reflected light angle is same to the incoming light angle (Fig. 8.9). This is explained by the Snell's law, which is described in a detailed manner in Ch. 10.1. Under the ideal specular material, the reflected light direction, $\vec{R}$ is computed as $2(\vec{N}\cdot\vec{L})\vec{N}-\vec{L}$.

The ideal specular material rarely exists in practice. More common objects are glossy materials, which have highlight along a particular direction and spreads its energy out from the direction. Specifically, when the viewing direction, $\vec{V}$, is on the ideal reflected direction $\vec{R}$, the viewer sees the highest illumination. We then reduce the energy as the viewing direction is away from $\vec{R}$. To capture this observation, the Phong illumination uses the following specular term:

$$\begin{aligned} I_{r,s} =& k_s I_s(\cos\phi)^{n_s} \\ =& k_s I_s(\vec{V}\cdot\vec{R})^{n_s}, \end{aligned} \tag{8.3}$$

where k_s, I_s, n_s are material-dependent specular coefficient, intensity for the specular component of a light, and specular exponent, respectively. Fig. 8.10 shows example results of the specular term.

The Phong model describes materials by treating them to have ambient, diffuse, and specular properties together.

The final Phong illumination is computed by summing these different terms, ambient, diffuse, and specular terms, of different lights (Fig. 8.5). Note that most common objects are described by combining these terms, not a single term.

Local and global illumination. While the Phong illumination is not a physically-based model, it has been widely used for its simplicity and efficiency. Nonetheless, it has a fundamental issue, a local illumination model. The Phong illumination achieves its

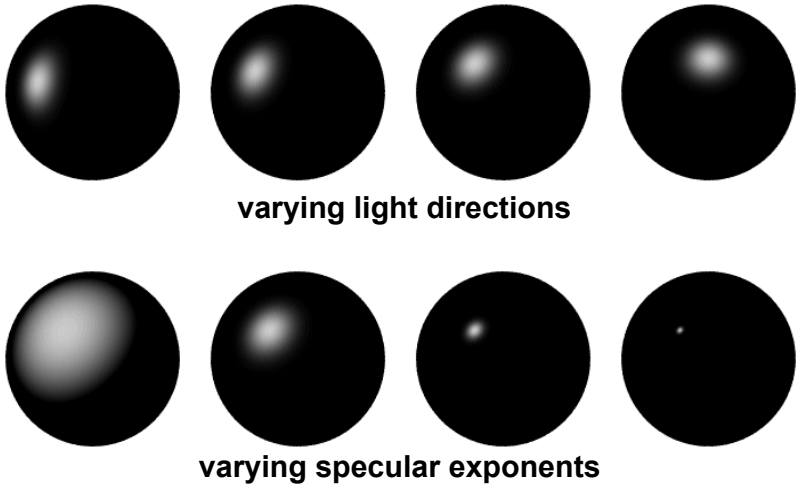


Figure 8.10: This shows how the illumination changes as a function of a viewing direction and specular exponents of the specular term of the Phong illumination model.

efficiency by considering only the local information such as the surface normal, viewing information, and the light information.

The Phong model and rasterization considers the rendering process locally for efficiency.

When there is a blocker occluding the light energy, it, however, generates shadows, which is not considered at all at the Phong illumination. To support such effects, we need to consider the reflected energy or blocked energy from other geometry. This requires us to access global information, which slows down the overall processing. Rasterization is designed in a way to reduce such global access for achieving a high performance and thus the Phong illumination has been well suited for rasterization. We discuss how to generate shadow within rasterization in Ch. 9.4.1. A more physically based approach is to use ray tracing techniques, which are explained in Part II.

8.4 Shading

Shading is a process of adjusting a color of a primitive based on various information such as the normal of the primitive and its angle to the light or view direction. Shading can refer to the illumination process and cover broader approaches including various effects (e.g., lens flare), which are implemented by shader programs. While shading is a broad topic, we discuss how to compute colors within the primitive (e.g., triangle) in this section.

Common shading (or interpolation) methods are as the following:

- **Flat shading.** For flat shading, we use only a single color for the primitive. As a result, each triangle in this approach looks to be constant, i.e., flat, in the image space. To perform flat shading, we use a normal of a triangle and perform the Phong illumination

model or other illumination models to compute the single color. This is the simplest and fastest approach.

- **Gouraud shading.** This approach provides a smooth rendering result by computing different colors for vertices of a primitive and interpolating them within the primitive. While this approach shows smooth rendering results, it comes with performing three independent illumination for vertices and interpolation.
- **Phong shading.** This approach interpolates normals of vertices and computes colors with them within the primitive by evaluating an illumination model with interpolated normals. Since we interpolate normals of vertices, we can generate highlight within the triangle, even though we do not have such highlight in each vertex. When we use the Gouraud shading in this case, we cannot generate the highlight within the primitive, since we interpolate colors of vertices. Fig. 8.11 shows difference between Gouraud and Phong shading methods.

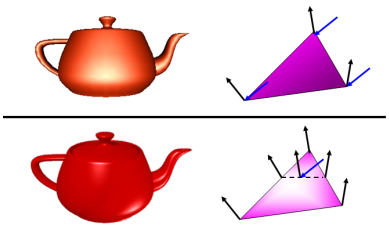


Figure 8.11: This shows Gouraud (top) and Phong (bottom) shading methods. Blue arrows indicate the locations of evaluating an illumination model.

8.5 Common Questions

We have learned that we compute an illumination value for each vertex. For smooth objects (like the teapot model shown in the slide), it should be okay. But, when we draw a box, then the box may look smooth, not showing different and discontinuous colors between neighboring faces of the box. If we use a normal for each vertex of the box, we may get such smooth rendering result, which is not correct one. Instead, we use multiple vertices for each point of the box. For example, we use different vertex data for each point in different faces of the box. Note that these vertices should have the same positional value, but they can have different normals, which can generate discontinuous colors between different faces. Refer to the slide of "Decoupling Vertex and Face Attributes via Indirection" in the lecture slide of "Interacting with a 3D world".

Is there any techniques that can show better quality than the Phong Illumination and can be used in interactive games? I want to know techniques that can show near physically-based illumination that can be used in games? Ambient occlusion has been proposed as an approximation for physically-based global illumination. It can be pre-computed and used quite quickly at runtime, leading to be suitable for interactive games. Moreover, in some CG movies, this technique has been used.

9
Texture

Achieving higher realism has been one of main goals of computer graphics. For this goal, we have developed many modeling techniques by using more triangles, lights, and materials. Unfortunately, using additional resources (e.g., triangles and lights) come with sides effects such as additional running time and memory overheads.

Texture mapping adds additional details, without much overheads.

Since achieving the interactive performance has been another main goal of computer graphics, various approximation rather than directly relying upon additional geometry and lights has been studied. Among various techniques, texture mapping has been one of main approximation techniques (Fig. 9.1).

One thing that we need to understand is that while textures are originally designed for representing complex shapes of geometry, they can be utilized for various other purposes. At a high level, a texture is simply a 2D array, which is one of the most simplest data representations in computer architectures, and can be readily precomputed and used at runtime. Note that the Z-buffer used for visibility determination can be considered as a type of a texture. Thanks to these nice properties, textures have been widely used.

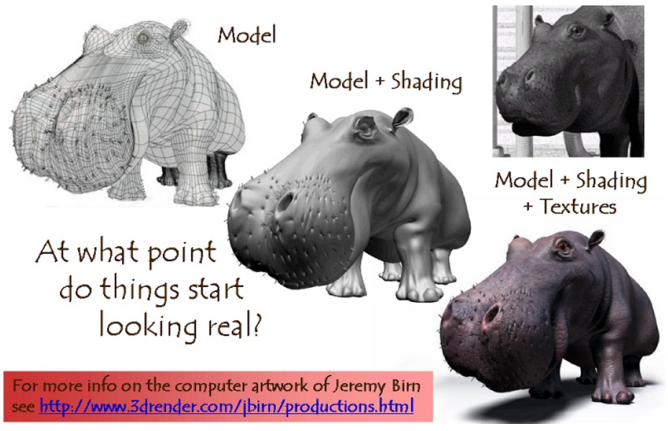


Figure 9.1: Texture mapping adds a lot of details to the geometry illuminated by lights, enabling higher realism without adding much overheads.

We first explain the main purpose of using texture mapping, followed by its various applications.

9.1 *Texture Mapping*

A texture is a 2 D (or 3 D and even a higher dimensional) buffer, whose each element represents a pixel color or some other values. Commonly a texture refers to a 2D image. Texture mapping indicates a mapping from a part of the texture to a part of a model. Fig. 9.3 shows an example of a 2D texture mapping.

We use 2 D texture coordinates, commonly (u, v), to locate a particular location of a 2D texture. We link the texture location to a particular location or a vertex of a mesh by using 2D or higher geometry coordinates (e.g., 3D coordinate (x, y, z) of a vertex). Fig. 9.2 shows that we map (u, v) coordinates of multiple texture locations to a 2D mesh, i.e., quad in the 2D space, represented by (x, y) coordinates. You may also recall that we use vt token to specify (u, v) texture coordinates to a vertex for an obj file format (Ch. 5.1).

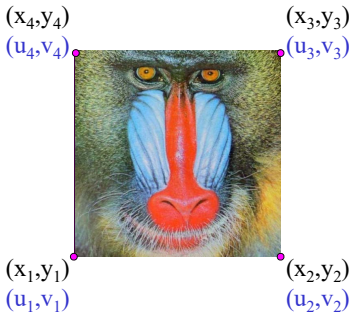


Figure 9.2: Texture mapping.

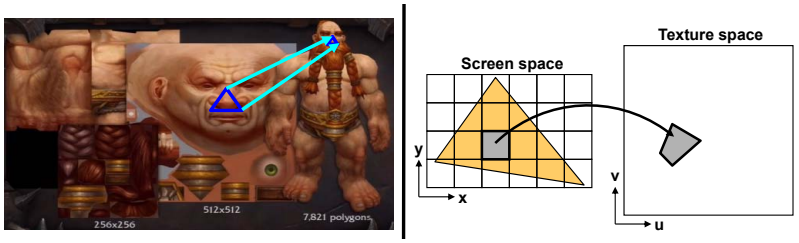


Figure 9.3: The left figure shows a mapping from a texture to a triangle representing a part of a character. We use a few texture maps with different resolutions. The image is excerpted from a MMO-champion site. To perform texture mapping, we compute a representative color of a pixel within the triangle from the texture space, as shown in the right figure. Note that a pixel of the triangle maps to an arbitrarily shaped quadrilateral in the texture space.

To apply the texture mapping, we compute a texture coordinate of a fragment of a triangle, while rasterizing the triangle. We compute the texture coordinate by interpolating texture coordinates associated with vertices of a triangle, as we did for other attributes (e.g., color) (Ch. 7.3).

Once we compute the texture coordinate, we compute the 2D indices of the corresponding texture pixel, known as texel, and use the color of the texel for illumination or other purposes.

Perspective-correct interpolation. Note that a naive interpolation of various vertex attributed in the image space does not provide the expected results that are supposed to be computed by the object-space interpolation. To achieve the correct result even in the image-space interpolation, the perspective-correct interpolation is developed.

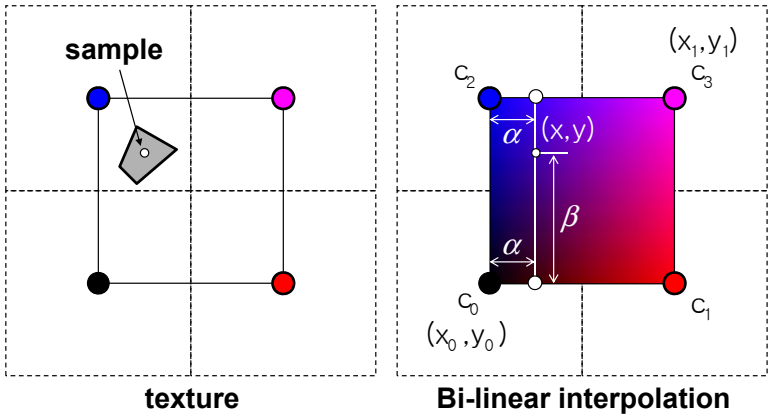


Figure 9.4: The left image shows the case of oversampling. The sampling in the texture space is too small compared to the texture resolution. The right image shows the bi-linear interpolation to address the oversampling problem.

9.2 *Oversampling of Textures*

Let's take a look at the right image of Fig. 9.3. A box-shaped fragment generated for rasterizing a triangle maps to a quadrilateral, i.e., a polygon with four sides, instead of the uniform box shape. This phenomenon occurs due to various transformations (e.g., modeling and projective transformations) and the angle of the triangle against the view direction.

Oversampling occurs when we zoom in the triangles.

Since the pixel in the image space does not match with that in the texture space, we have two cases: oversampling and undersampling cases. Oversampling refers to the case where the sampling resolution in the texture space is smaller than the available resolution of the texture. Fig. 9.4 shows this oversampling case. The quadrilateral in the texture space is even contain in a texel of the texture. The oversampling issue occurs when we magnify the geometry.

We aim to reconstruct the original signal out of available texture samples and compute the signal value at the sampled texture location.

Please take a moment to think about how to compute the representation color for the quadrilateral. Surprisingly, this kind of issues is quite common in computer graphics, image processing, etc. A simple method is to identify the nearest neighbor texel center and use its color for the quadrilateral. In the case of the left image of Fig. 9.4, the blue pixel is the closest to the quadrilateral, more exactly, the sampling point location. Note that during the rasteriation, we compute colors or other attributes based on center positions of pixels.

While this nearest neighbor approach is quite fast, its visual quality is poor, especially along the boundary of texels. In other words, when two sampling locations are very close, but are in different texels, they get different colors, resulting in visual gaps in the image space (Fig. 9.5).

Another approach is to use linear interpolation. Given the sam-

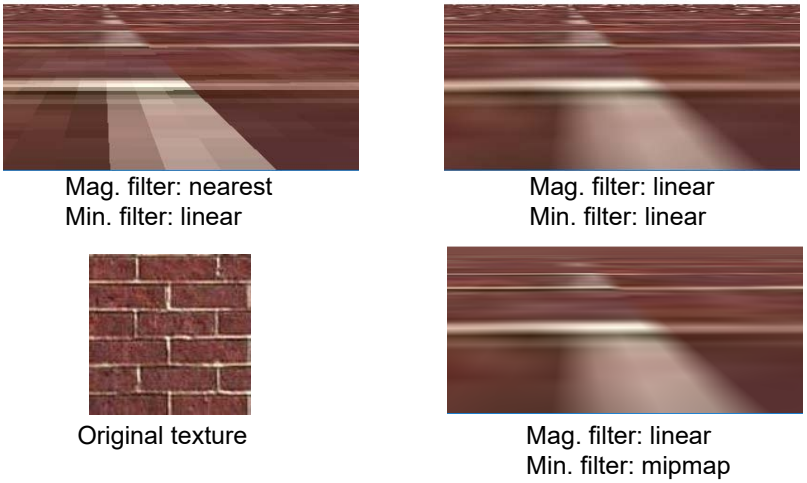


Figure 9.5: Different rendering results with different mag. and min. filters for texture mapping.

pling location, we identify four nearby texels, e.g., c_0 to c_3 in the right image of Fig. 9.4. We then perform the linear interpolation along the U and V texture directions, thus named as bi-linear interpolation. Let us define α and β to be a blending factor along X and Y directions, respectively. They are then defined as the following:

$$\alpha = \frac{x - x_0}{x_1 - x_0}, \beta = \frac{y - y_0}{y_1 - y_0}. \tag{9.1}$$

The color, c, at the sampling location under the bi-linear interpolation is computed as follows:

$$c = (1 - \beta)\left((1 - \alpha)c_0 + \alpha c_1\right) + \beta\left((1 - \alpha)c_2 + \alpha c_3\right). \tag{9.2}$$

The effect of using the bi-linear interpolation over the nearest neighbor one is shown in Fig. 9.5. We can see that boundary shapes of texels were smoothened. Nonetheless, we can also see that the edge information inherent in the original texture was filtered out too. We can thus see that there are trade-off in terms of filtering unnecessary edges and preserving original edges. This boils down to the classical reconstruction and sampling problem.

9.3 *Under-sampling of Textures*

Under-sampling arises when we zoom out the geometry, and thus a fragment of a triangle maps to a large area in the texture space.

Let us know discuss the other sampling issue, undersampling. Undersampling rises when we zoom out from the geometry and thus each triangle become small in the image space. Therefore, a pixel of a triangle maps to a large quadrilateral area in the texture space. The problem is thus to compute a representative color out of many texels covered by the quadrilateral.

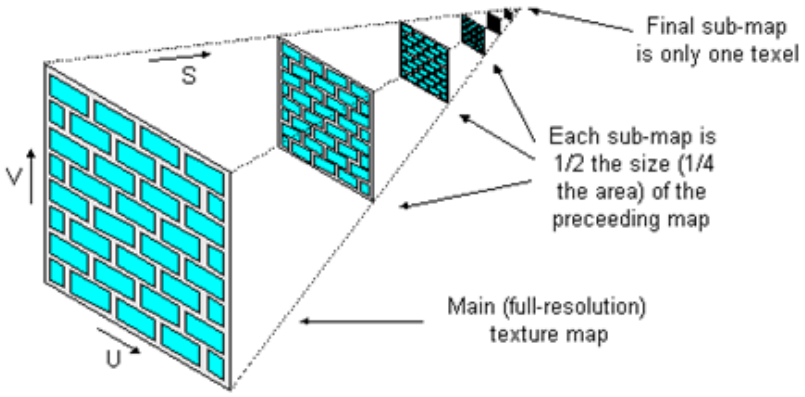


Figure 9.6: This shows a mipmap or image pyramid of an image.

A naive approach to the undersampling is to compute all those texels under the quadrilateral and compute a representative color value, e.g., the average value of them. This approach, unfortunately, slows, since it requires us to access many texels and computations. Instead of this on-demand approach, pre-filtering that pre-filters the original texture in a way to efficiently handle undersampling has been more widely used and studied. In this section, we discuss two approaches, mip mapping and summed area table, for the undersampling problem.

Mipmap or mipmapping is a multi-scale representation for a texture (or any other types of images) to efficiently handle the undersampling issue. Given an input image, a mipmap is composed of a sequence of images whose U and V resolutions are reduced half over its higher resolution (Fig. 9.6). As a result, mipmap is also called an image pyramid. Each low-resolution image is a pre-filtered version of its higher one.

At runtime when we use the mipmap, we pick a particular image level among the available image resolutions of the mipmap given the required texture resolution. If necessary, we can also perform interpolation between two image resolutions, resulting in tri-linear interpolation for computing a color for the sampling location. In whatever cases, we access only a few samples on the mipmap and get pre-filtered texture values, resulting in faster and better visual quality.

The memory requirement of using a mipmap is $\frac{1}{3}$, about 33%, since the total size of using the mipmap is computed as the following:

$$\sum_{0}^{\infty} \left(\frac{1}{4}\right)^{i} = \frac{1}{1-\frac{1}{4}} = \frac{4}{3}. \tag{9.3}$$

Fig. 9.5 shows different rendering results w/ linear filtering or

mipmap. By using the mipmap, we get smoother image results over linear filtering for far-away regions where we minify the geometry. Again, the mipmap is a fast way of handling the undersampling problem, but can remove the original edge information.

A reason why the mipmap produces a over-smoothed result is that the mipmap computes its image pyramid only based on an isotropic filtering shape, i.e., square shapes. As a result, when we have have an very elongated quadrilateral shape in the texture space, we cannot find filtering resolutions along both U and V directions. A solution to this case of anisotropic filtering is the summed area table.

Summed-area table. A summed-area table is proposed to support anisotropic filtering, specifically, a rectangular shape, not the squared shape, on the texture space. Given a texture, $T(u, v)$, its summed-area table, $S(u, v)$, is computed by summing all the elements whose elements are smaller than u or v:

$$S(u, v) = \sum_{i \leq u \wedge j \leq v} T(u, v). \tag{9.4}$$

We then compute the average color value, c_a, on a rectangular regions, e.g., the blue region given by $[u_0, u_1] \times [v_0, v_1]$ as shown in Fig. 9.7, as the following:

$$c_a = \frac{T(u_1, v_1) - T(u_1, v_0) - T(u_0, v_1) + T(u_0, v_0)}{(u_1 - u_0)(v_1 - v_0)}. \tag{9.5}$$

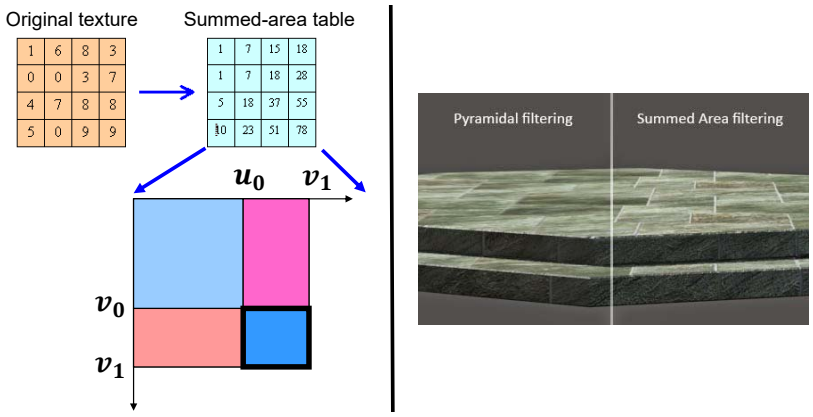


Figure 9.7: The left images show a configuration of the summed area table, while the right image shows rendering results of the summed area table and mipmapping. The right image is created by Denny.

Fig. 9.7 also compares the rendering results computed by the mipmap and summed-area table. The summed-area table shows better quality, since it provides anisotropic filtering. Nonetheless, it has additional runtime and memory overheads over the mipmap.

9.4 *Approximating Lights*

It is easy to paint on images and capture images than constructing geometry, and thus textures have been widely used for various applications. In this section, we discuss two techniques, shadow mapping and environment mapping, of using textures for approximating complex lights.

Before we move on to them, let us first discuss light maps. Light maps are images that contain light intensity. We then use these light maps as textures for adjusting colors of triangles. A simple method of computing colors with a light map is to multiply the intensity contained in the light map with the color computed by illumination or other functions. Fig. 9.8 shows an example of using textures and light maps. Creating complex lighting effects requires high computation, and thus pre-computing, also called baking, them in light maps are still commonly used in many interactive applications.

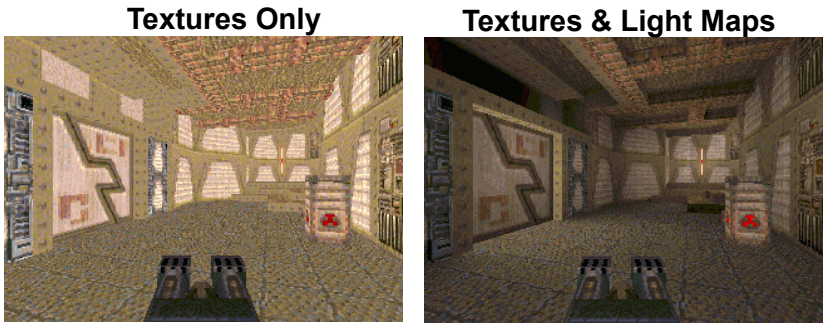


Figure 9.8: This shows results only with textures and both with textures and light maps used for a game called Quake.

9.4.1 *Shadow Mapping*

Shadow is one of fundamental lighting effects that we can see in daily life, and provide various 3D depth cues. While providing shadows is important, it is not that easy to efficiently and correctly generate shadows in rasterization. This problem has been studied for many decades and shadow mapping as a type of texture mapping is proposed for creating realistic rendering results within the rasterization framework.

Please recall our discussion on the Phong illumination model (Ch. 8.3). The model has three components of ambient, diffuse, and specular terms. Unfortunately, diffuse and specular terms do not consider any other objects that block lights from light sources, while the ambient term is a drastic simplification by using a constant for considering inter-reflection. Essentially, the Phong illumination model does not consider the case of having shadows, i.e., the existence of

other objects that block the light. This is mainly attributed since the Phong illumination model, more importantly, rasterization itself, is a local model that mainly aims for high efficiency.

Our challenge is to generate shadows within the rasterization framework. While considering shadows itself requires us to access other objects, resulting in global access on various data, we approach this problem as a two-pass algorithm using shadow mapping. Its main concept is shown in Fig. 9.9.

Shadow mapping is a two pass rendering method to generate shadows without global and random access on other objects.

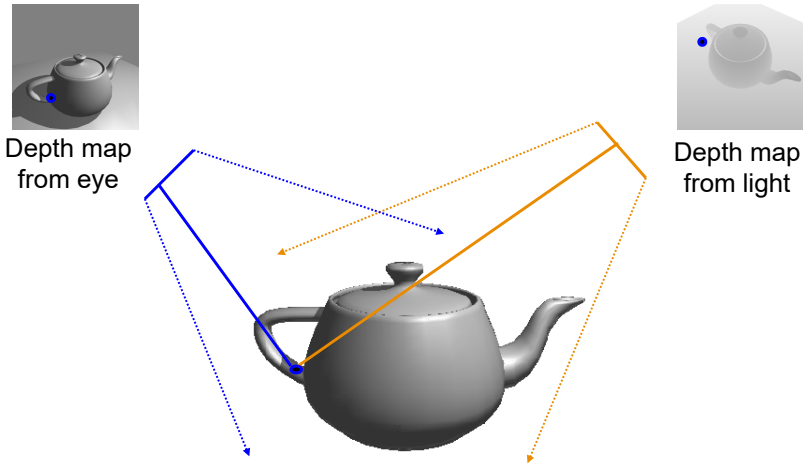


Figure 9.9: This visualizes the process of using shadow mapping to generate shadows on the rendering result seen by the eye.

The problem of the rasterization process is that when we perform an illumination on a fragment of a triangle, we cannot know whether the fragment can receive the light energy from a light source. When we do not receive the light energy due to a blocking object, we add only the ambient term, since the diffuse and specular terms become zero. To know whether a fragment can receive the light energy from a light source, we rasterize the whole scene at the position of the light source and treat its depth map as a shadow map for the light. This is the first-pass of generating shadows.

The depth map generated from the light position contains depth values of visible geometry from the light. We then raster the whole scene at the viewer's position, similar to the regular rasterization process. This is the second pass of our method. A difference in this second pass compared to the regular rasterization is that we check whether we can receive the light energy on a fragment that we generate at the second pass.

To know whether the fragment receives the light energy or not, we compute its depth from the light position, d_l. When d_l is bigger than the stored depth value, d, of the shadow map, we determine that the fragment cannot receive the light energy and we thus give only the

ambient term to the fragment, not the diffuse nor specular terms.

While we explain shadow mapping in a concise manner above, there are a lot of technical issues. Most of them are related to the oversampling and undersampling that we discussed for texture mapping; note that the shadow map is another type of textures and thus inherits issues of texture mapping. Nonetheless, it is very important to understand how we address a kind of global illumination, shadow generation, through a texture, the shadow map.

Shadows maps are just a type of textures and thus inherits pros. and cons. of texture mapping.

9.4.2 *Environment Mapping*

In the prior section, we discussed how to generate shadows using shadow mapping. Another common rendering effect is to support reflection on mirrors or other metal-like objects. For those models, we see other objects reflected on such reflecting objects. In other words, supporting this effect belongs to a type of global illumination requiring the access to other objects.

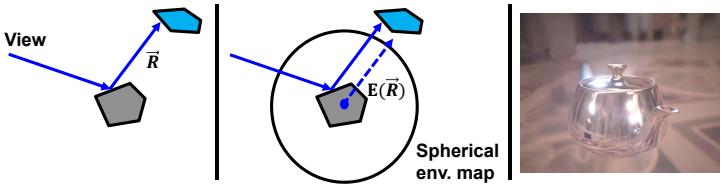


Figure 9.10: Two images in the left visualize how we use the spherical environment mapping, while the right image shows an example of using the environment map to simulate the reflection effect.

The rasterization framework also relies upon using another type of texture mapping, environment mapping, for this reflection effect. Suppose that we have a view direction on a reflecting object shown in the gray color in Fig. 9.10. When the object is the specular object, the reflected ray, $\vec{R}$, is computed by the Snell's law (Ch. 8.3). We then need to access an object along the reflected ray, $\vec{R}$. Unfortunately, this is a ray tracing process, and is not efficiently adopted for rasterization.

To enable the reflection efficiently, we introduce environment mapping, which captures colors of the surrounding environment in a texture. For environment mapping, we can use different types of geometry capturing the environment. Examples include sphere, cube maps, etc. In this chapter, we explain environment mapping based on a sphere for the sake of the simplicity.

An environment map captures surrounding geometry or lights, and can be used as a texture to approximate them at runtime.

As shown in the middle image of Fig. 9.10, we place a sphere at the center of the reflection object. We map the sphere into a 2D texture space; since we can represent the sphere with two angles, θ and ϕ, the 2D texture space can be constructed by these two angles. We then generate a ray starting from the center of the sphere to each texel of the sphere and encode the color of the ray at that texel; we

use a projection instead of ray tracing for efficiently building the map.

At runtime, when we raster a triangle of the reflecting model, we know the viewing direction, and thus identify a texel ID that the reflection ray from the center of the sphere, $E(\vec{R})$, will access. Unfortunately, since the environment map is generated at the center of the object, not each location that we have reflection, there are visual gaps between the computed one and the ground truth. Nonetheless, we can support an approximate reflection by using an additional texture.

The environment map is also used to encode complex types of lights and used for providing realistic lighting for rasterization.

9.5 *Approximating Geometry*

Textures are also used to approximate complicated geometry. Especially, when we have many geometry, it requires long running computation time with high memory requirement. A single or multiple textures are effective ways of approximating them with reduced running and memory overheads.

Figure 9.11: We use the bump map (shown in the middle) to adjust normals of the geometry during the rasterization, to enrich the appearance of the model (shown in the right.) Since we do not change the actual geometry, we can see that the geometry is unchanged at its silhouette.

Bump and normal mapping. Bump mapping modifies normals of geometry, not the actual geometry. The texture used for bump mapping encodes an amount of changes to normals of the geometry (Fig. 9.11). This is an approximate, yet effective way of enriching the geometry. Nonetheless, we can observe that the actual geometry is not aligned with the adjusted normals, especially when we look at the silhouette of the object. Normal mapping is similar to bump mapping, but the normal map directly gives the normal that we use on top of a simple geometry (Fig. 9.12).

Displacement mapping. Unlike bump and normal mapping, displacement mapping adjusts the actual geometry based on a provided displacement map. A common usage of displacement mapping is to encode a height change on the displacement map and adjust the geometry along its normal direction according to the height. Adjusting the geometry requires tessellation, subdividing the geometry into

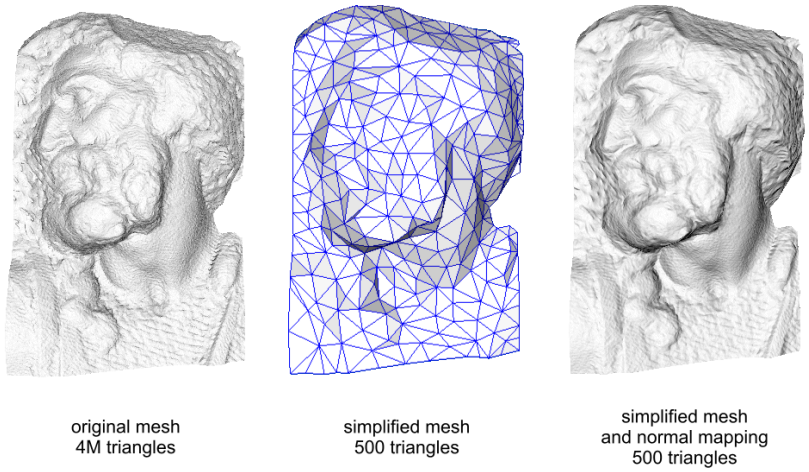


Figure 9.12: We can provide detailed look on a simple geometry by using normal mapping. The image is created by Paolo Cignoni.

smaller patches and adjusting them to accommodate the given height (Fig. 9.13).

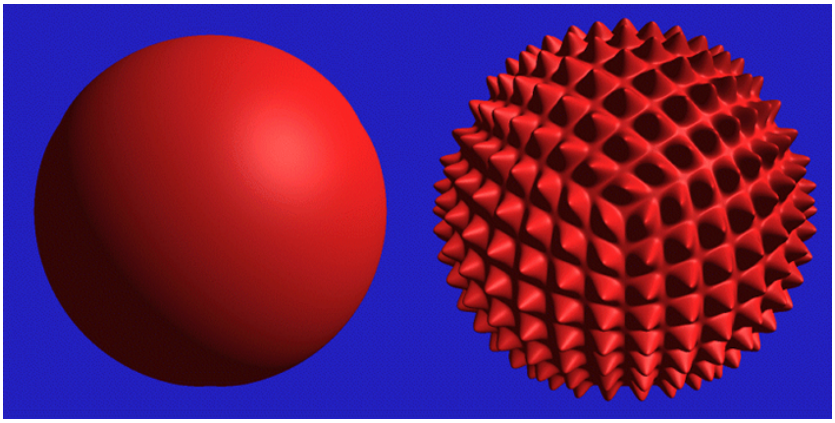


Figure 9.13: Displacement mapping changes the actual geometry according to its map unlike bump mapping. To enable displacement mapping, we tessellate the initial geometry into smaller ones.

We covered only a few examples of approximating geometry. Other notable examples include 3D or solid textures representing 3D shapes and billboards, which are a set of 2D textures representing complex geometry (e.g., trees).

Part II

Physically-based Rendering

In Part I, we discussed rasterization techniques. While the raterization technique provides the efficient performance based on rendering pipeline utilizing modern GPUs, its fundamental approach is not based on the physical interaction between lights and materials. Another large stream of rendering methods are based on such physical interactions and thus are known as physically-based rendering.

In this part, we discuss two different approaches, ray tracing and radiosity, of physically based rendering methods. Ray tracing and radiosity are two main building blocks of many interactive or physically based rendering techniques. We first discuss ray tracing in this chapter, followed by radiosity (Ch. 11. We then study radiometric quantities (Ch. 12) to measure different energy terms to describe the physical interaction, known as the rendering equation (Ch. 13.1).

The rendering equation is a high dimensional integral problem, and thus its analytic solutions in many cases are not available. As an effective solution to solving the equation, we study the Monte Carlo technique, a numerical approach in Ch. 14, and its integration with ray tracing in Ch. 15. In many practical problems, such Monte Carlo approaches are slow to converge to noise-free images. We therefore study importance sampling techniques in Ch. 14.3.

9.6 Available Tools

Physically based rendering has been studied for many decades, and many useful resources are available. Some of them are listed here:

- Physically Based Rendering: From Theory to Implementation [1]. This book also known as pbrt comes with concepts with their actual implementations. As a result, readers can get understanding on those concepts and actual implementation that they can play with. Since this book discusses such implementation, we strongly recommend you to play with their source codes, which are available at github.
- Embree [2] and Optix [3]. Embree and Optix are interactive ray tracing kernels that run on CPUs and GPUs, respectively. While source codes of Optix are unavailable, Embree comes with their source codes.
- Instant Radiosity. Instant radiosiy is widely used in many games, thanks to its high quality rendering results with reasonably fast performance. Unfortunately due to its importance in recent game industry, mature library or open source projects are not available. One of useful open source projects are from my graphics lab. It is available at: `http://sglab.kaist.ac.kr/~sungeui/ICG/student_presentations.html`.

[1] Matt Pharr and Greg Humphreys. *Physically Based Rendering, Second Edition: From Theory To Implementation*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2010b. ISBN 0123750792, 9780123750792

[2] Ingo Wald, Sven Woop, Carsten Benthin, Gregory S Johnson, and Manfred Ernst. Embree: A kernel framework for efficient cpu ray tracing. *ACM Trans. Graph.*, 2014

[3] Steven G. Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, and Martin Stich. Optix: a general purpose ray tracing engine. *ACM Trans. Graph.*, 29:66:1–66:13, 2010

10
Ray Tracing

Ray casting and tracing techniques have been introduced late 70's and early 80's to the computer graphics field as rendering techniques for achieving high-quality images.

Ray casting [1] shoots a ray from the camera origin to a pixel and compute the first intersection point between the ray and objects in the scene. Ray casting then computes the color from the intersection point and use it as the color of the pixel. It computes a direct illumination that has one bounce from the light to the eye. Its result is same to those of the basic rasterization considering only the direct illumination.

[1] Arthur Appel. Some techniques for shading machine renderings of solids. In *AFIPS 1968 Spring Joint Computer Conf.*, volume 32, pages 37–45, 1968

Ray tracing [2] is an recursive version of the ray casting. In other words, once we have the intersection between the initial ray and objects, ray tracing generates another ray or rays to simulate the interaction between and the light and objects. A ray can be considered as a photon traveling in a straight line, and by simulating many rays in a physically correct way, we can achieve physically correct images. While the algorithm is extremely simple, we can support various effects by generating different rays (Fig. 10.1).

[2] Turner Whitted. An improved illumination model for shaded display. *Commun. ACM*, 23(6):343–349, 1980

10.1 Basic algorithm

Ray tracing simulates how a photon interacts with objects.

The basic ray tracing algorithm is very simple, as shown in Algorithm 1. We first generate a ray from the eye to the scene. While a photon travels from a light source, we typically perform ray tracing in backward from the eye (Fig. 10.2). We then identify the first intersection point between the ray and the scene. This has been studied well, especially around the early stage of developing this technique. At this point, we simply assume that we can compute such intersection points and this is discussed in Sec. 10.2.

Suppose that we identify such an intersection point between the ray and the scene. We can then perform various shading operations based on the Phong illumination (Sec. 8.3). To see whether the point

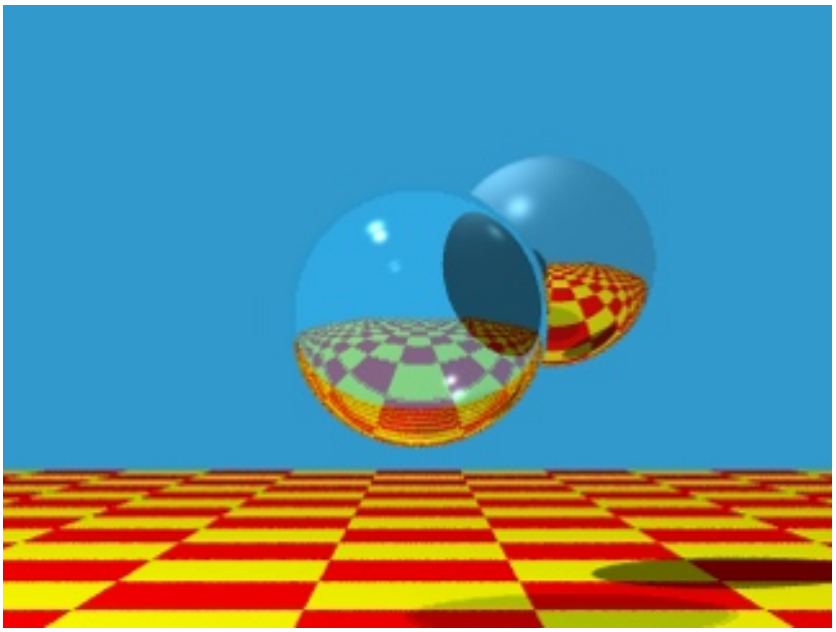


Figure 10.1: One of early images generated by ray tracing, i.e., Whitted style ray tracing. The image has reflection, refraction, and shadow effects. The image is excerpted from its original paper.

Algorithm 1 Basic ray tracing

```
Trace rays from the eye into the scene (backward ray tracing).
Identify the first intersection point and shade with it.
Generate additional, secondary rays needed for shading.
    Generate ray for reflections.
    Generate ray for refraction and transparency.
    Generate ray for shadows.
```

is under the shadow or not, we simple generate another ray, called shadow ray, to the light source (the bottom image of Fig. 10.2).

Reflection and refractions are handled in a similar manner by generating another secondary rays (Fig. 10.3). The main question that we need to address here is how we can construct the secondary rays for supporting reflection and refraction. For the mirror-like objects, we can apply the perfect-specular reflection and compute the reflection direction for the reflection ray, where the incoming angle is same to the outgoing angle. In other words, the origin of the reflection ray, R, is set to the hit point of the prior ray, and the direction of R is set as the reflection direction. Its exact equation is shown in Sec. 8.

Most objects in practice do not support such perfect reflection. For simple cases such as rays bending in glasses or water, we apply the Snell's law to compute the outgoing angle for refraction. The Snell's

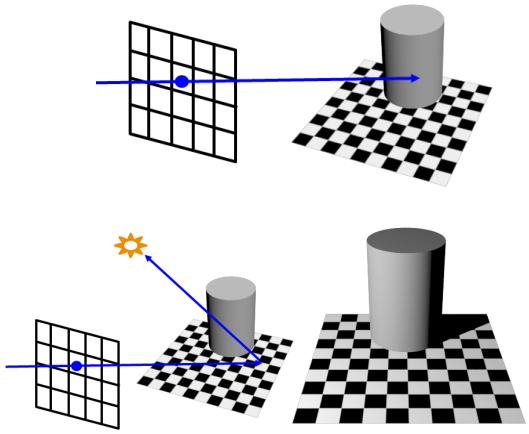


Figure 10.2: We generate a ray, primary ray, from the eye (top). To see whether the intersection point is in the shadow or not, we generate another ray, shadow ray, to the light source (bottom). These images are created by using 3ds Max.

law is described as follows:

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1}, \tag{10.1}$$

where θ_1 and θ_2 are incoming and outgoing angles given rays at the interface between two different objects (Fig. 10.4). n_1 and n_2 are refractive indices of those two objects. The refractive index of a material (e.g., water) is defined as $\frac{c}{v}$, where c is the velocity of the light in vacuum, while v is the speed of the light in that material. As a result, refractive indices of different materials are measured and can be used for simulating such materials within ray tracing.

Many objects used in practice consist of many different materials. As a result, the Snell's law designed for isotropic media may not be appropriate for such cases. For general cases, BRDF and BSSRDF have been proposed and are discussed in Ch. 12.

Physically based rendering techniques adopt many physical laws, as exemplified by adopting the Snell's law for computing refraction rays. This is one of main difference between rasterization and physically based rendering methods.

For various effects, ray tracing generate different types of rays, while rasterization adopts different types of texture maps.

Note that in rasterization techniques, to handle shadow, reflection, refraction, and many other rendering effects, we commonly generate some maps (e.g., shadow maps) accommodating such effects. As a result, handling texture mapping efficiently is one of key components for many rasterization techniques running on GPUs. On the other hand, ray tracing generates various rays for such effects, and handling rays efficiently is one of key components of ray tracing.

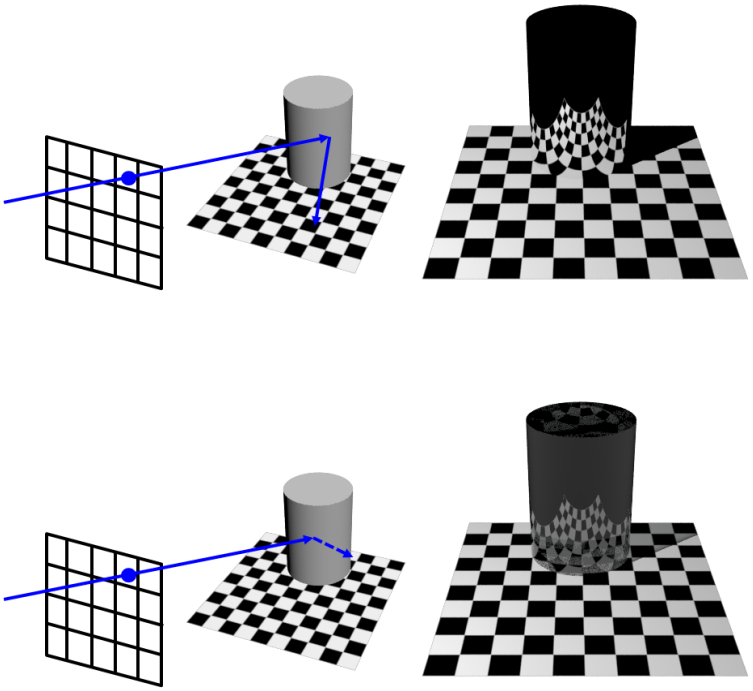


Figure 10.3: Handling reflection and refraction by generating secondary rays.

10.2 *Intersection Tests*

Performing intersection tests is one of main operations of ray tracing. Furthermore, they tend to become the main bottleneck of ray tracing and thus have been optimized for a few decades. In this section, we discuss basic ways of computing intersection tests between a ray and a few simple representations of a model.

Implicit forms of objects are commonly used for intersection tests.

Any points, $p(t)$, in a ray parameterized by a parameter t can be represented as follows:

$$p(t) = o + t\vec{d}, \tag{10.2}$$

where o and $\vec{d}$ are the origin and direction of the ray, respectively. A common way of approaching this problem is to first define an object in an implicit mathematical form, $f(p) = 0$, where p is any point on the object. We then compute the intersection point, t_i, satisfying $f(p(t_i)) = 0$.

We now look at a specific case of computing an intersection point between a ray and a plane. A well known implicit form of a plane is:

$$\vec{n}p - d = 0, \tag{10.3}$$

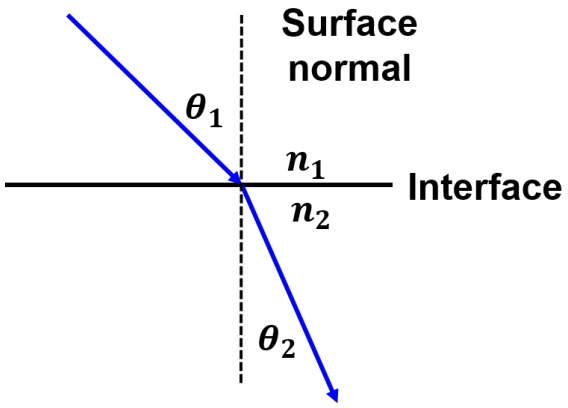


Figure 10.4: How a ray bends at an interface between simple objects, specifically, isotropic media such as water, air, and glass, is described by the Snell's law.

where $\vec{n}$ is a normalized normal vector of the plane and d is the distance from the origin to the plane. This implicit form of the plane equation is also known as the Hessian normal form [3].

[3] E. Weisstein. From mathworld–a wolfram web resource. URL http://mathworld.wolfram.com

By plugging the ray equation into the implicit of the plane equation, we get:

$$\begin{aligned} \vec{n}(o + t\vec{d}) - d &= 0 \\ t &= \frac{d - \vec{n}o}{\vec{n} \cdot \vec{d}}. \end{aligned} \tag{10.4}$$

We now discuss a ray intersection method against triangles, which are one of common representations of objects in computer graphics. There are many different ways of computing the intersection point with triangles. We approach the problem based on barycentric coordinates of points with a triangle.

Barycentric coordinates are computed based on non-orthogonal bases.

Barycentric coordinates are computed based on non-orthogonal bases unlike the Cartesian coordinate system, which uses orthogonal bases such as X, Y, and Z-axis. Suppose that p is an intersection point between a ray and a triangle consisting of three vertices, v_0, v_1, v_2 (Fig. 10.5). We can represent the point p as the following:

$$\begin{aligned} p =& v_0 + \beta(v_1 - v_0) + \gamma(v_2 - v_0) \\ =& (1 - \beta - \gamma)v_0 + \beta v_1 + \gamma v_2 \\ =& \alpha v_0 + \beta v_1 + \gamma v_2, \end{aligned} \tag{10.5}$$

where we use α to denote $1 - \beta - \gamma$. We can then see a constraint that $\alpha + \beta + \gamma = 1$, indicating that we have two degrees-of-freedom, while there are three parameters.

Let's see in what ranges of these parameters the point p is inside the triangle. Consider edges along two vectors $v_0 - v_1$ and $v_2 - v_0$ (Fig. 10.5). Along those edges, β and γ should be in $[0, 1]$, when the

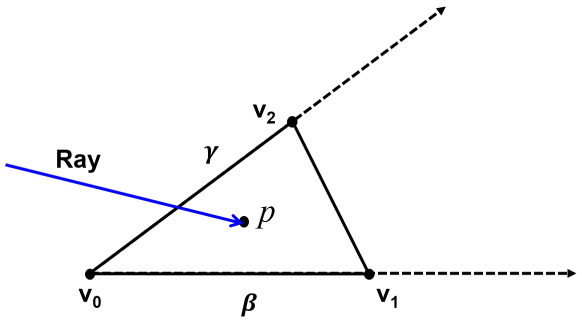


Figure 10.5: In the barycentric coordinate system, we represent the point p with β and γ coordinates with two non-orthogonal basis vectors, $v_1 - v_0$ and $v_2 - v_0$.

point is inside the triangle. Additionally, when we consider the other edge along the vector of $v_1 - v_2$, points on the edge satisfy $\gamma = 1 - \beta$[4]. When we plug the equation into the definition of α, we see α to be zero. On the other hand, on the point of v_0, β and γ should be zero, and thus α to be one. As a result, we have the following property:

$$0 \leq \alpha, \beta, \gamma \leq 1, \tag{10.6}$$

where these three coordinates are barycentric coordinates and $\alpha = 1 - \beta - \gamma$.

[4] When we consider a 2 D space whose basis vectors map to canonical vectors (e.g., X and Y axises) with β and γ coordinates, one can easily show that the relationship $\gamma = 1 - \beta$ is satisfied on the edge of $v_2 - v_1$.

Barycentric coordinates are also known as area coordinates, since they map to areas of sub-triangles associated with vertices.

There are many different ways of computing barycentric coordinates given points defined in the Cartesian coordinate system. An intuitive way is to associate barycentric coordinates with areas of sub-triangles of the triangle; as a result, barycentric coordinates are also known as area coordinates. For example, β associated with v_1 is equal to the ratio of the area of $\triangle p v_0 v_2$ to that of $\triangle v_0 v_1 v_2$.

Once we represent the intersection point p within the triangle with the barycentric coordinates, our goal is to find t of the ray that intersects with the triangle, denoted as the following:

$$o + t\vec{d} = (1 - \beta - \gamma)v_0 + \beta v_1 + \gamma v_2, \tag{10.7}$$

where unknown variables are t, β, γ. Since we have three different equations with X, Y, and Z coordinates of vertices and the ray, we can compute those three unknowns.

10.3 *Bounding Volume Hierarchy*

We have discussed how to perform intersection tests between a ray and implicit equations representing planes and triangles. Common models used in games and movies have thousands of or millions of triangles. A naive approach of computing the first intersection point between a ray and those triangles is to linearly scan those triangles and test the ray-triangle intersection tests. It, however, has a linear

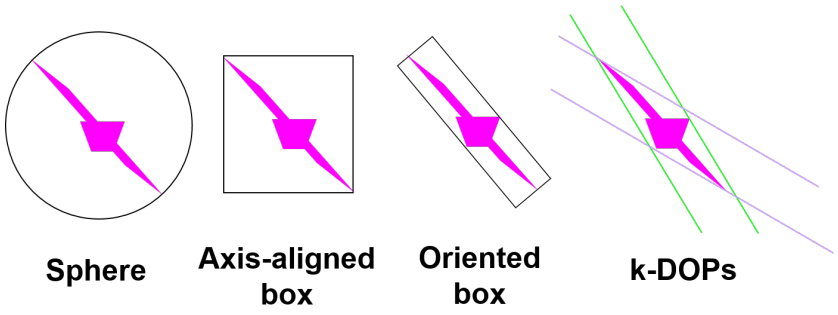


Figure 10.6: This figure shows different types of Bounding Volumes (BVs).

time complexity as a function of the number of triangles, and thus can take an excessive amount of computation time.

Many acceleration techniques have been proposed to reduce the time spent on ray intersection tests. Some of important techniques include optimized ray-triangle intersection tests using Barycentric coordinates [5]. In this section, we discuss an hierarchical acceleration technique that can improve the linear time complexity of the naive linear scan method.

Two hierarchical techniques have been widely used for accelerating the performance of ray tracing. They are kd-trees and bounding volume hierarchies (BVHs). kd-trees are constructed by partitioning the space of a scene and thus are classified as spatial partitioning trees. On the other hand, BVHs are constructed by partitioning underlying primitives (e.g., triangles) and thus known as object partitioning trees. They have been demonstrated to work well in most cases [6]. We focus on explaining BVHs in this chapter thanks to its simplicity and wide acceptance in related fields such as collision detection.

[5] Tomas Möller and Ben Trumbore. Fast, minimum storage ray-triangle intersection. *J. Graph. Tools*, 1997

Bounding volume hierarchies are simple to use and have been widly adopted in related applications including collision detection.

[6] Ingo Wald, Sven Woop, Carsten Benthin, Gregory S Johnson, and Manfred Ernst. Embree: A kernel framework for efficient cpu ray tracing. *ACM Trans. Graph.*, 2014

10.3.1 Bounding Volumes

We first discuss bounding volumes (BVs). A BV is an object that encloses triangles. Also, the BV should be efficient for performing an intersection test between a ray and the BV. Given this constraint, simple geometric objects have been proposed. BVs commonly used in practice are sphere, Axis-Aligned Bounding Box (AABB), Oriented Bounding Box (OBB), k-DOPs (Discrete Oriented Polytopes), etc. (Fig. 10.6).

Spheres and AABBs are fast for checking intersection tests against a ray. Furthermore, constructing these BVs can be done quite quickly. For example, to compute a AABB from a soup of triangles, we just need to traverse those triangles and compute mim and max values of x, y, and z coordinates of triangles. We then compute the AABB

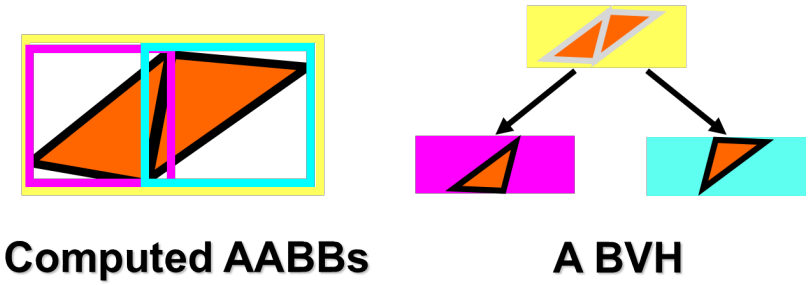


Figure 10.7: This figure shows a BVH with its nodes and AABBs given a model consisting of three triangles. Note that two child AABBs have a spatial overlap, while their nodes have different triangles. As a result, BVHs are classified into an object partitioning tree.

out of those computed min and max values. Since many man made artifacts have box-like shapes, AABB works well for those types. Nonetheless, spheres and AABBs may be too lose BVs, especially when the underlying object is not aligned into such canonical directions or is elongated along a non-canonical direction (Fig. 10.6).

A single BV type is not always better than others, but AABBs work reasonably well and are easy to use.

On the other hand, OBBs and k-DOPs tend to provide tighter bounding, but to require more complex and thus slow intersection tests. Given these trade-offs, an overhead of computing a BV, tightness of bounding, and time spent on intersection tests between a ray and a BV, it is hard to say which BV shows the best performance among all those BVs. Nonetheless, AABBs work reasonably well in models used for games and CAD industry, thanks to its simplicity and reasonable bounding power on those models.

10.3.2 Construction

Let's think about how we can construct a bounding volume hierarchy out of triangles. A simple approach is a top-down construction method, where we partition the input triangles into two sets in a recursive way, resulting in a binary tree. For simplicity, we use AABBs as BVs.

We first construct a root node with its AABB containing all the input triangles. We then partition those triangles into left and right child nodes. To partition those triangles associated with a current node, a simple method is to use a 2 D plane that partitions the longest edge of the current AABB of the node. Once we compute triangle sets for two child nodes, we recursively perform the process until each node has a fixed number of triangles (e.g., 1 or 2).

In the aforementioned method, we explained a simple partitioning method. More advanced techniques have been proposed including optimization techniques with Surface Area Heuristic (SAH) [7]. The SAH method estimates the probability that a BV intersects with random rays, and we can estimate the quality of a computed BVH. It

[7] C. Lauterbach, S.-E. Yoon, D. Tuft, and D. Manocha. RT-DEFORM: Interactive ray tracing of dynamic scenes using bvhs. In *IEEE Symp. on Interactive Ray Tracing*, pages 39–46, 2006

has been demonstrated that this kind of optimizations can be slower than the simple method, but can show shorter traversal time spent on performing ray-BVH intersection tests.

BVHs suits well for dynamic models, since it can be refitted or re-computed from scratch efficiently.

Dynamic models. Many applications (e.g., games) use dynamic or animated models. As a result, it is important to build or update BVHs of models as they are changing. This is one of main benefits of using BVHs for ray tracing, since it is easy to update the BVH of a model, as the model changes its positions or is animated.

One of the most simple methods is to refit the existing BVH in a bottom-up manner, as the model is changing. Each leaf node is associated with a few triangles. As they change their positions, we re-compute the min and max values of the node and update the AABB of the node. We then merge those re-computed AABBs of two child nodes for their parent node by traversing the BVH in a bottom-up manner. This process has the linear time complexity in terms of the number of triangle. Nonetheless, this refitting approach can result in a poor quality, when the underlying objects deform significantly.

To address those problems, many techniques have been proposed. Some of them is to build BVHs from scratch every frame by using many cores [8] and to selectively identify a sub-BVH with poor quality and rebuild only those regions, known as selective restructuring [9]. At an extreme case, the topology of models can change due to fracturing of models. BVH construction methods even for fracturing cases have been proposed [10].

[8] C. Lauterbach, M. Garland, S. Sengupta, D. Luebke, and D. Manocha. Fast bvh construction on gpus. *Computer Graphics Forum (EG)*, 28(2):375–384, 2009

[9] Sungeui Yoon, Sean Curtis, and Dinesh Manocha. Ray tracing dynamic scenes using selective restructuring. *Eurographics Symp. on Rendering*, pages 73–84, 2007

[10] Jae-Pil Heo, Joon-Kyung Seong, DukSu Kim, Miguel A. Otaduy, Jeong-Mo Hong, Min Tang, and Sung-Eui Yoon. FASTCD: Fracturing-aware stable collision detection. In *SCA '10: Proceedings of the 2010 ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, 2010

10.3.3 Traversing a BVH

Once we build a BVH, we now traverse the BVH for ray-triangle intersection tests. Since an AABB BVH provides AABBs, bounding boxes, on the scene in a hierarchical manner, we traverse the BVH in the hierarchical manner.

Given a ray, we first perform an intersection test between the ray and the AABB of the root node. If there is no intersection, it guarantees that there are no intersections between the ray and triangles contained in the AABB. As a result, we skip traversing its sub-tree. If there is an intersection, we traverse its sub-trees by accessing its two child nodes. Among two nodes, it is more desirable to access a node which is located closer to the ray origin, since we aim to identify the first intersection point along the ray starting from the ray origin.

Suppose that we decide to access the left node first. We then store the right node in a stack to process it later. We continue this process until we reach a leaf node containing primitives (e.g., triangles). Once we reach a leaf node, we perform ray-triangle intersection

tests for identifying an intersection point. If it is guaranteed that the intersection point is the closest to the ray origin, we terminate the process. Otherwise, we contribute to traverse the tree, by fetching and accessing nodes in the stack.

Many types of BVHs do not provide a strict ordering between two child nodes given a ray. This characteristic can result in traversing many parts of BVHs, leading to lower performance. Fortunately, this issue has been studied, and improvements such as identifying near and far child nodes have been proposed [11].

[11] C. Lauterbach, S.-E. Yoon, D. Tuft, and D. Manocha. RT-DEFORM: Interactive ray tracing of dynamic scenes using bvhs. In *IEEE Symp. on Interactive Ray Tracing*, pages 39–46, 2006

10.4 Visibility Algorithms

In this chapter, we discussed different aspects of ray tracing. At a higher level, ray casting, a module of ray tracing, is one type of visibility algorithms, since it essentially tells us whether we can see a triangle or not given a ray. In this section, we would like to briefly discuss other visibility algorithms.

The Z-buffer method, an fundamental technique for rasterization (Part I), is another visibility algorithm. The Z-buffer method is an image-space method, which identifies a visible triangle at each pixel of an image buffer by considering the depth value, i.e., Z values of fragments of triangles (Ch. 7.4). Many different visibility or hidden-surface removal techniques have been proposed. Old, but well-known techniques have been discussed in a famous survey [12]. Interestingly, the Z-buffer method was mentioned as a brute-force method in the survey, because of its high memory requirement. Nonetheless, it has been widely adopted and used for many graphics applications, thanks to its simple method, resulting in an easy adoption in GPUs.

While the Z-buffer method was considered as a brute-force method, it is the de-factor standard in the rasterization method thanks to its adoption in modern GPU architectures.

[12] Ivan E. Sutherland, Robert F. Sproull, and Robert A. Schumacker. A characterization of ten hidden-surface algorithms. *ACM Comput. Surv.*, 6(1):1–55, 1974

Compared with the Z-buffer, ray casting and ray tracing is much slower, since it uses a hierarchical data structure, and has many incoherent memory access. Ray casting based approaches, however, become more widely accepted in movies and games, because modern GPUs allow to support such complicated operations, and many algorithmic advances such as ray beams utilizing coherence have been prosed. It is hard to predict future exactly, but ray casting based approaches will be supported more and can be adopted as an interactive solution at some point in future.

11
Radiosity

In the last chapter, we discussed ray tracing techniques. While ray tracing techniques can support various rendering effects such as shadow and transparency, their performance was identified too slow to be used for interactive graphics applications. Some of issues of ray tracing is that we generate many rays whenever we change viewpoints. Furthermore, processing those rays take high computation time, and they tend to have random access patterns on underlying data structures (e.g., meshes and bounding volume hierarchy), resulting in high cache misses and lower computational performance.

On the other hand, radiosity emerges as an alternative rendering method that works for special cases with high performance [1]. While radiosity is not designed for handling various rendering effects, it has been widely used to complement other rendering techniques, since radiosity shows high rendering performance of specific material types such as diffuse materials. In other words, radiosiy as well as ray tracing are two common building blocks of designing other advanced rendering techniques, and we thus study this technique in this chapter.

[1] Cindy M. Goral, Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile. Modelling the interaction of light between diffuse surfaces. In *Computer Graphics (SIGGRAPH '84 Proceedings)*, volume 18, pages 212–22, July 1984

11.1 Two Assumptions

Radiosity has two main assumptions (Fig. 11.1):

- **Diffuse material.** We assume that the material type we handle for radiosiy is diffuse or close to the diffuse materials. The ideal diffuse material reflects incoming light into all the possible outgoing directions with the equal amount of light energy, i.e., the same radiance, which is one of radiometric quantity discussed in Sec. 12. Thanks to this diffuse material assumption, any surface looks the same and has the same amount of illumination level given the view point. This in turn simplifies many computations.

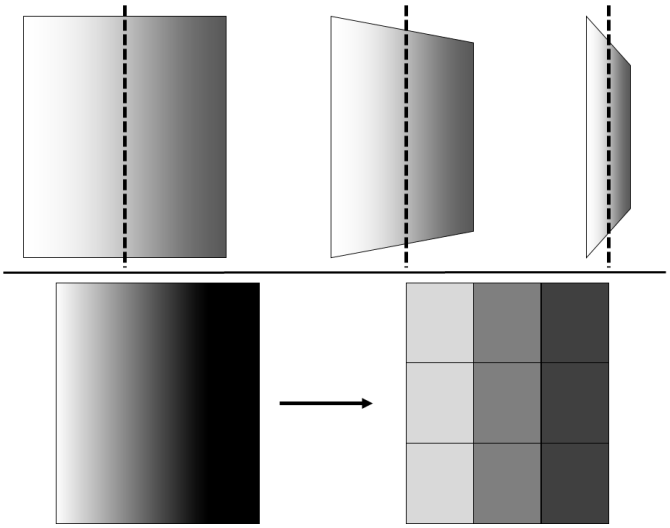


Figure 11.1: Radiosity has the diffuse material assumption (top) and constant illumination per surface element (bottom).

- **Constant radiance per each surface element.** Take a look at a particular surface (e.g., a wall or a desk in your room). The illumination level typically varies smoothly depending on the configuration between a point in the surface and position of light sources. To support this phenomenon, radiosity treats that each surface is decomposed into surface elements such as triangles. It then assumes for simplicity that each surface element has a single value related to the illumination level, especially, radiosity value (Ch. 12). Simply speaking, radiosity is the total incoming (or outgoing) energy arriving in a unit area in a surface.

We will see how these assumptions lead to a simple solution to the rendering problem.

Relationship with finite element method (FEM). As you will see, radiosity can generate realistic rendering results with an interactive performance, while dealing only with diffuse materials and light sources. This was excellent results, when radiosity was proposed back at 1984. Furthermore, approaches and solution for radiosity were novel at the graphics community at that time. Nonetheless, those techniques were originally introduced for simulating heat transfers and have been well established as Finite Element Methods (FEM). FEM was realizing its potential benefits around 1960s and 70s, and was applied even to a totally different problem, physically based rendering. This is a very encouraging story to us. By studying and adopting recently developing techniques into our own problem, we can design very creative techniques in our own field!

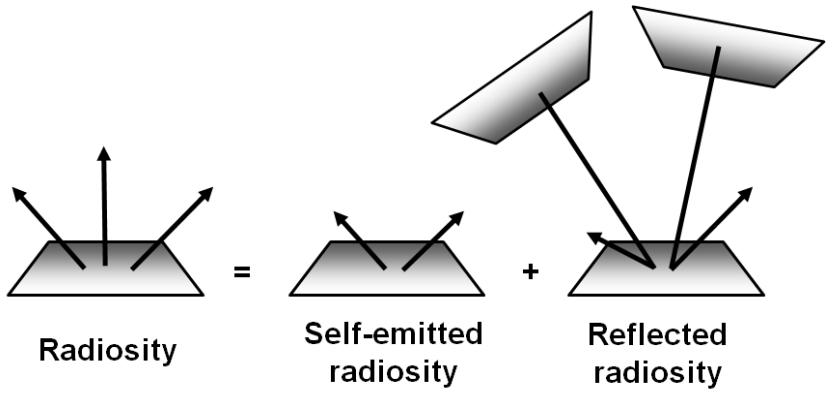


Figure 11.2: The radiosity of a patch is computed by the sum of the self-emitted radiosity from itself and the radiosity reflected and received from other patches.

11.2 *Radiosity Equation*

An input scene to radiosity is commonly composed of triangles. We first subdivide the scene into smaller triangles such that our assumption of the constant radiance per each subdivided triangle is valid. Suppose that there are n different surface elements. We use B_i to denote radiosity of a patch i. Some of such patches can be light sources and thus emit some energy. Since we also assume the light sources to be diffuse emitters, we also use radiosity for such self-emitting patches, and their emitting energy is denoted by $B_{e,i}$.

Intuitively speaking, the radiosity of the patch i is the sum of the self-emitting energy from the patch itself, $B_{e,i}$, and the energy reflected from the patch i by receiving energy from all the other patches (Fig. 11.2. We can then model the interaction between the patch i and different patches as the following:

$$B_i = B_{e,i} + \rho_i \sum_j B_j F(i \to j), \tag{11.1}$$

where j is another index to access all the surface elements in the scene, $F(i \to j)$ is a form factor that describes how much the energy from the patch i arrives at another patch j, and ρ_i is a reflectivity of the patch i.

$B_{e,i}$ and ρ_i are input parameters to the equation and given by a scene designer. The form factor is a term that we can compute depending on the geometric configuration between two patches i and j. The form factor can be understood by the area integration of the rendering equation, which is more general than the radiosity equation. This is discussed in Sec. 13.2. As a result, the unknown terms of the equation is the radiosity B_i of n different patches. Our goal is then to compute such unknown terms. We discuss them in the next section, followed by the overall algorithm of the radiosity

rendering method.

11.3 Radiosity Algorithm

Given the radiosity equation (Eq. 11.1), the unknown term is the radiosity, B_i, per each patch, resulting in n different unknown radiosity values for n patches. Since we can setup n different equation for each patch based on the radiosity equation, overall we have n different equations and unknowns. When we represent such n different equations, we have the following matrix representation:

$$\begin{bmatrix} 1-\rho_1 F(1\to 1) & -\rho_1 F(1\to 2) & \dots & -\rho_1 F(1\to n) \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_n F(n\to 1) & -\rho_n F(n\to 2) & \dots & 1-\rho_n F(n\to n) \end{bmatrix} \begin{bmatrix} B_1 \\ \vdots \\ B_n \end{bmatrix} = \begin{bmatrix} B_{e,1} \\ \vdots \\ B_{e,n} \end{bmatrix} \tag{11.2}$$

The above matrix has the form of $AX = B$, where $X = [B_1 \dots B_n]^T$ is a 1 by n matrix containing unknowns.

To compute the unknown X, we can apply many matrix inversion algorithms including Gaussian elimination that has $O(n^3)$ time complexity [2]. This approach, however, can be very expensive to be used for interactive applications, since the number of surface elements can be hundreds of thousands in practice.

[2] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, Cambridge, England, 2nd edition, 1993

Instead of using exact approaches of computing the linear equations, we can use other numerical approaches such as Jacobi and Gauss-Seidel iteration methods. Jacobi iteration works as the following:

- **Initial values.** Start with initial guesses on radiosity values to surface patches. For example, we can use the direct illumination results using Phong illumination considering the light sources as the initial values for surface patches.
- **Update step.** We plug those values, i.e., old values, into the right term of the radiosity equation (Eq. 11.1), and get new values on B_i. We perform this procedure to all the other patches.
- **Repeat until converge.** We repeat the update step until radiosity values converge.

The Jacobi iteration method has been studied well in numerical analysis, and its properties related to convergence have been well known [3].

[3] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, Cambridge, England, 2nd edition, 1993

One numerical iteration simulates one bounce of the light energy from a patch to another patch.

Effects of numerical iteration. Instead, we discuss how it works in the context of rendering. While performing the update step of the

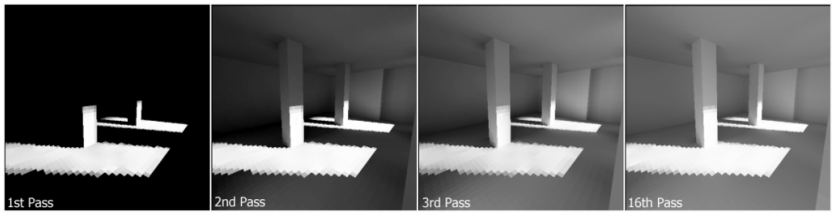


Figure 11.3: This shows a sequence of images computed by different updates, i.e., light bounces, during the radiosity iteration process. This is the courtesy of the wikipedia.

Jacobi iteration, we compute a new radiosity value for each patch from old values. In this process, we compute the new radiosity value received and reflected from other patches. Intuitively, the update step supports one bounce of the light energy from a patch to another patch.

Fig. 11.3 visualizes how radiosity values change as we have different number of update steps, i.e., passes. While only surface elements that are directly visible from the light source are lit in the first pass, other surface elements get brighter as we perform multiple update steps and thus multiple bounces. In a way, this also visualizes how the incoming light energy is distributed across the scene. In the end, we see only the converged result, which is the equilibrium state of the light and material interaction described in the radiosity equation.

Overall algorithm. In summary, we subdivide triangles of the input scene into smaller surface elements, i.e., patches. We then compute radiosity values per each patch by solving the linear equations given by the radiosity equation. For static models, we perform this process only a single time. At runtime, when a viewer changes a view point, we then project those triangles whose color. This projection process is efficiently performed by using the rasterization process in GPUs.

Radiosity is commonly accelerated by adopting the rasterization method

So far, we did not consider view points given by users while computing radiosity values. This is unnecessary, because we do not need to consider view-dependent information for radiosity computation process; note that radiosity algorithm assumes the diffuse materials and emitters and thus we get the same radiance value for any view directions. This is one of the main features of the radiosity algorithm, leading to its strength and weakness of the method.

The basic radiosity method does not support glossy materials.

Drawbacks of the basic radiosity method. The main benefit of the basic radiosity method is that we can re-use the pre-computed radiosity values, even though the user changes the viewpoint. Nonetheless, it has also drawbacks. First of all, the radiosity assumes different materials and emitters, while various scenes have other materials such as glossy materials. Also, when we have dynamic models, we

cannot re-use pre-computed radiosity values and thus re-compute them.

11.4 Light Path Expressions

The radiosity method does support light reflections between diffuse materials, but does not support interactions between glossy materials. Can we represent such light paths that the radiosity method supports?

Regular expressions are used to denote different types of light paths.

Heckbert proposed to use the regular expression to characterize light paths [4]. This approach considers light paths starting from the eye, noted E, to the light, denoted, L. Diffuse, specular, and glossy materials are denoted as D, S, and G, respectively. We also adopt various operations of regular expressions such as | (or), * (zero or more), and + (one or more).

The light paths that radiosity method are then characterized by LD^*E. On the other hand, the classic ray tracing method (Ch. 10) supports $L(DS^*)E$, since it generates secondary rays when a ray hits specular or refractive objects.

[4] Paul S. Heckbert. Adaptive radiosity textures for bidirectional ray tracing. In Forest Baskett, editor, *Computer Graphics (SIGGRAPH '90 Proceedings)*, volume 24, pages 145–154, August 1990

12
Radiometry

One of important aspects of physically-based rendering is to simulate physical interactions between lights and materials in a correct manner. To explain these physical interactions, we discuss various physical models of light in this chapter. Most rendering effects that we observe can be explained by a simple, geometric optics. Based on this simple light model, we then explain radiometric quantities that are important for computing colors. Finally, we explain basic material models that are used for simulating the physical interaction with lights.

12.1 Physics of Light

Understanding light has drawn major human efforts in physics and resulted in many profound progress on optics and related fields. Light or visible light is a type of electromagnetic radiations or waves that we can see through our eyes. The most general physical model is based on quantum physics and explains the duality of wave and particle natures of light.

While the quantum physics explains the mysterious wave-particle duality, it is rather impossible to simulate the quantum physics for making our applications, i.e., games and movies, at the current computing hardware. One of simpler light models is the wave model that treats light like sound. Such wave characteristics become prominent, when the wavelength of light is similar to sizes of interacting materials, and diffraction is one of such phenomena. For example, when we see sides of CD, we can see rainbow-like color patterns, which are created by small features of the CD surface.

The most commonly used light model used in computer graphics so far is the geometric optics, which treats light propagation as rays. This model assumes that object sizes are much bigger than the wavelength of light, and thus wave characteristics disappear mostly. This geometric optics can support reflection, refraction, etc.

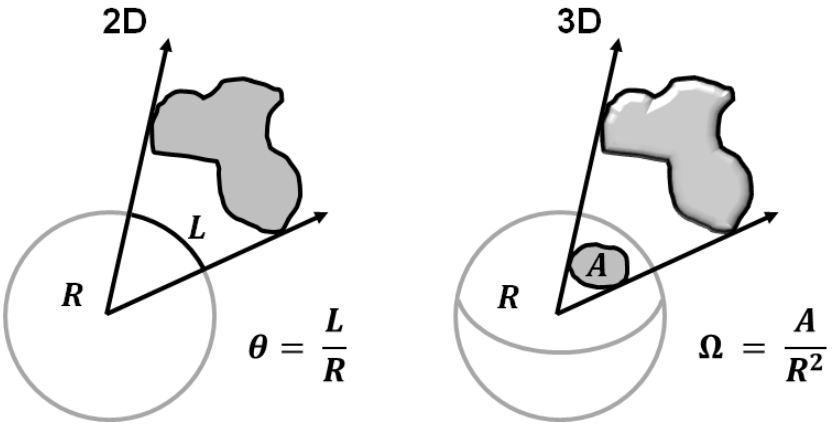


Figure 12.1: Solid angles in 2 D and 3 D cases.

Many rendering methods based on ray tracing assumes the geometric optics, and we also assume this model unless mentioned otherwise.

Our goal is then to measure the amount of energy that a particular ray carries or that a particular location receives from. Along this line, we commonly use a hemisphere, specifically, hemispherical coordinates, to parameterize rays that can arrive at a particular location in a surface. We discuss hemispherical coordinates before we move on to studying radiometry.

Solid angles. We use the concept of solid angles for various integration on the hemisphere. The solid angle is used to measure how much an objected located in 3 D space affects a point in a surface. This metric is very useful for computing shadow and other factors related to visibility. In the 2 D case (the left figure of Fig. 12.1), a solid angle, Ω, of an object is measured by $\frac{L}{R}$, where L is the length of the arc, where the object is projected to in the 2 D hemisphere (or sphere). R is the radius of the sphere; we typically use a unit sphere, where $R = 1$. The unit of the solid angle in the 2 D case is measured by radians. The solid angle mapping to the full circle is 2π radians.

The solid angle in the 3 D case is computed by $\frac{A}{R^2}$, whose unit is steradians (the right figure of Fig. 12.1). A indicates the area subtended by the 3 D object in the hemisphere. For example, the full sphere has 4π steradians.

Hemispherical coordinates. A hemisphere is two dimensional surface and thus we can represent a point on the hemisphere with two parameters such as latitude, θ, and longitude, φ (Fig. 12.2), where $\theta \in [0, \frac{\pi}{2}]$ and $\varphi \in [0, 2\pi]$. Now let's see how we can compute the differential area, dA, on the hemisphere controlled by $d\phi$ and $d\theta$.

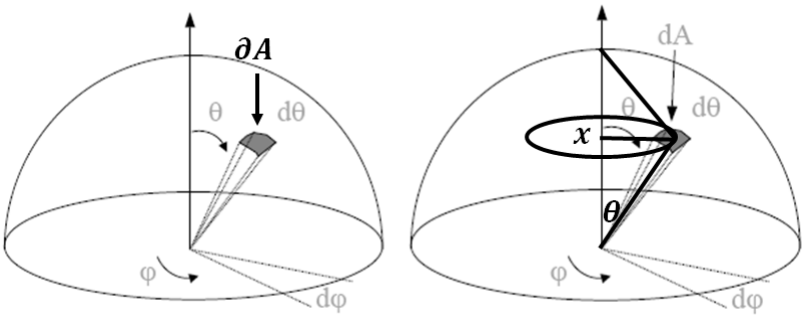


Figure 12.2: Hemispherical coordinates (θ, φ). These images from slides of Kavita Bala.

In infinitely small differential angles, we can treat that the area is approximated by a rectangular shape, whose area can be computed by multiplying its height and width. Its height is given by $d\theta$. On the other hand, its width varies depending on θ; its largest and minimum occur at $\theta = \pi/2$ and $\theta = 0$, respectively.

To compute the width, we consider a virtual circle that touches the rectangular shape of the hemisphere. Let x be the radius of the ch ircle. The radius is then compute by $\sin\theta = \frac{x}{r}, x = r\sin\theta$, where r is the radius of the hemisphere. The width is then computed by applying the concept of the solid angle, and is $r\sin\theta d\phi$. We then have the following differentials:

$$dA = (r\sin\theta d\phi)(rd\theta). \tag{12.1}$$

Based on this equation, we can easily derive differential solid angles, dw:

$$dw = \frac{dA}{r^2} \tag{12.2}$$

$$= \sin\theta d\phi d\theta. \tag{12.3}$$

We use these differential units to define the rendering equation (Ch. 13.1).

12.2 *Radiometry*

In this section, we study various radiometric quantities that are important for rendering. Human perception on brightness and colors depends on various factors such as the sensitivity of photoreceptor cells in our eyes. Nonetheless, those photoreceptor cells receive photons and trigger biological signals. As a result, measuring photons, i.e., energy, is the first step for performing the rendering process.

Power or flux. Power, P, is a total amount of energy consumed per unit time, denoted by dW/dt, where W indicates watt. In our

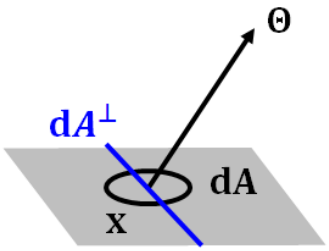


Figure 12.3: Radiance is measured per unit projected area, $dA^{\perp}$, while we receive the energy on the surface A.

rendering context, it is the total amount of energy arriving at (or passing through) a surface per unit time, and also called radiant flux. Its unit is Watt, which is joules per second. For example, we say that a light source emits 50 watts of radiant power or 20 watts of radiant power is incident on a table.

Irradiance or radiosity. Irradiance is power or radiant flux arriving at a surface per unit area, denoted by dW/dA with the unit of W/m^2. Radiant exitance is the radiant flux emited by a surface per unit area, while radiosity is the radiant flux emitted, reflected, or transmitted from a surface per unit area; that is why the radiosity algorithm has its name (Ch. 11). For example, when we have a light source emitting $100W$ of area $0.1m^2$, we say that the radiant existance of the light is $1000W/m^2$.

Radiance is one of the most important radiometric quantity used for physically-based rendering.

Radiance. In terms of computing rendering images, computing the radiance for a ray is the most important radiometric measure. The radiance is radiant flux emitted, reflected, or received by a surface per unit solid angle and per unit projected area, $dA^{\perp}$, whose normal is aligned with the center of the solid angle (Fig. 12.3):

$$L(x \to \Theta) = \frac{d^2P}{d\Theta dA^{\perp}} \tag{12.4}$$

$$= \frac{d^2P}{d\Theta dA\cos\theta}. \tag{12.5}$$

$\cos\theta$ is introduced for considering the projected area.

Diffuse emitter. Suppose that we have an ideal diffuse emitter that emits the equal radiance, L, in any possible direction. Its irradiance on a location is measured as the following:

$$\begin{aligned} E &= \int_{\Theta} L\cos\theta dw_{\Theta}, \\ &= \int_0^{2\pi}\int_0^{\frac{\pi}{2}} L\cos\theta\sin\theta d\theta d\phi = \int_0^{2\pi} d\phi \int_0^{\frac{\pi}{2}} L\cos\theta\sin\theta d\theta \\ &= 2\pi L\frac{1}{2} = L\pi. \end{aligned} \tag{12.6}$$

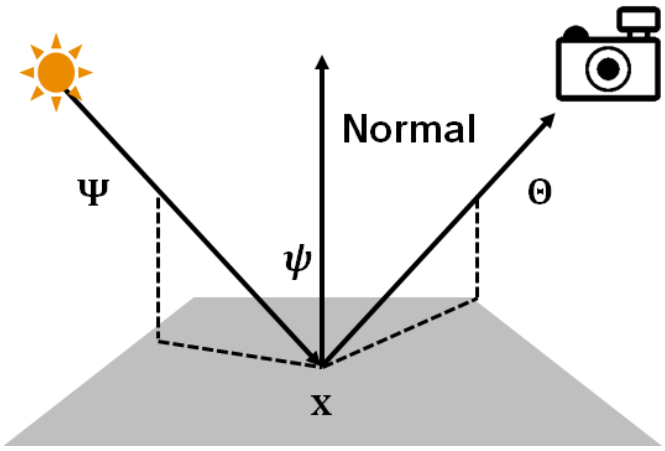


Figure 12.4: A configuration setting for measuring the BRDF is shown. Ψ and Θ are incoming and outgoing directions, while ψ is the angle between the surface normal and Ψ.

where Θ is the hemispherical coordinates, (θ, ϕ).

12.3 *Materials*

We discussed the Snell's law to support the ideal specular (Sec. 10.1. Phong illumination supports ideal diffuse and a certain class of glossy materials (Ch. 8). However, some materials have complex appearances that are not captured by those ideal specular, ideal diffuse, and glossy materials. In this section, we discuss Bidirectional Reflectance Distribution Function (BRDF) that covers a wide variety of materials.

Our idea is to measure an appearance model of a material and to use it within physically based rendering methods. Suppose the light and camera settings shown in Fig. 12.4. We would like to measure how the material reflects incoming radiance with a direction of Ψ into outgoing radiance with a direction of Θ. As a result, BRDF, $f_r(x, \Psi \rightarrow \Theta)$, at a particular location x is a four dimensional function, defined as the following:

$$f_r(x, \Psi \rightarrow \Theta) = \frac{dL(x \rightarrow \Theta)}{dE(x \leftarrow \Psi)} = \frac{dL(x \rightarrow \Theta)}{L(x \leftarrow \Psi) \cos \psi dw_\Psi}, \tag{12.7}$$

where ψ is the angle between the normal of the surface at x and the incoming direction Ψ, and dw_Ψ is the differential of the solid angle for the light. The main reason why we use differential units, not non-differential units, is that we want to cancel existing light energy in addition to the light used for measuring the BRDF.

The BRDF satisfies the following properties:

1. Reciprocity. Simply speaking, when we switch locations of the camera and light, we still get the same BRDF. In other words, $f_r(x, \Psi \rightarrow \Theta) = f_r(x, \Theta \rightarrow \Psi)$.

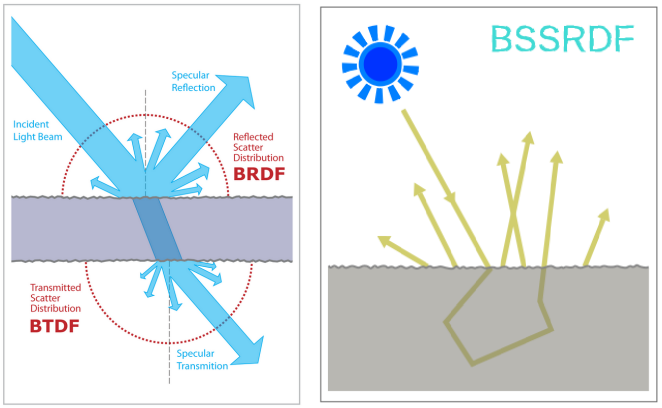


Figure 12.5: These images show interactions between the light and materials that BRDF, BTDF, and BSSRDF. These images are excepted from Wiki.

2. Energy conservation. $\int_\Theta f_r(x, \Psi \rightarrow \Theta) \cos\theta dw_\Theta \leq 1$.

To measure a BRDF of a material, a measuring device, called gonioreflectometer, is used. Unfortunately, measuring the BRDF takes long time, since we have to scan different incoming and outgoing angles. Computing BRDFs in an efficient manner is an active research area.

Material appearance varies depending on wavelengths of lights. To support such material appearance depending on wavelengths of lights, we can measure BRDFs as a function of wavelengths, and use a BRDF given a wavelengths of the light.

12.3.1 Other Distribution Functions

So far, we mainly considered BRDF. BRDF, however, cannot support many other rendering effects such as subsurface scattering.

BRDF considered reflection at a particular point, x. For translucent models, lights can pass through the surface and are reflected in the other side of the surface. To capture such transmittance, BTDF (Bi-direction Transmittance Distribution Function) is designed (Fig. 12.5). Furthermore, light can be emitted from points other than the point x that we receive the light. This phenomenon occurs as a result of transmittance and reflection within a surface of translucent materials. BSSRDF (Bidirectional Surface Scattering Reflection Distribution Function) captures such complex phenomenon. Capturing and rendering these complex appearance models is very important topics and still an active research area.

13
Rendering Equation

In this chapter, we discuss the rendering equation that mathematically explains how the light is reflected given incoming lights. The radiosity equation (Ch. 11) is a simplified model of this rendering equation assuming diffuse reflectors and emitters.

Nonetheless, the rendering equation does not explain all the light and material interactions. Some aspects that the rendering equation does not capture include subsurface scattering and transmissions.

13.1 Rendering Equation

The rendering equation explains how the light interacts with materials. In particular, it assumes geometric optics (Sec. 12.1) and the light and material interaction in an equilibrium status.

The inputs to the rendering equation are scene geometry, light information, material appearance information (e.g., BRDF), and viewing information. The output of the rendering equation is radiance values transferred, i.e., reflected and emitted, from a location to a particular direction. Based on those radiance values for primary rays generated from the camera location, we can compute the final rendered image.

Suppose that we want to compute the radiance, $L(x \rightarrow \Theta)$, from a location x in the direction of Θ [1]. To compute the radiance, we need to sum the emitted radiance, $L_e(x \rightarrow \Theta)$, and the reflected radiance, $L_r(x \rightarrow \Theta)$ (Fig. 13.1). The emitted radiance can be easily given by the input light configurations. To compute the reflected radiance, we need to consider incoming radiance to the location x and the BRDF of the object at the location x. The incoming radiance can come to x in any possible directions, and thus we introduce an integration with the hemispherical coordinates. In other words, the reflected radiance is computed as the following:

$$L_r(x \rightarrow \Theta) = \int_{\Psi} L(x \leftarrow \Psi) f_r(x, \Psi \rightarrow \Theta) \cos\theta_x dw_{\Psi}, \tag{13.1}$$

[1] For simplicity, we use a vector Θ for representing a direction based on the hemispherical coordinates.

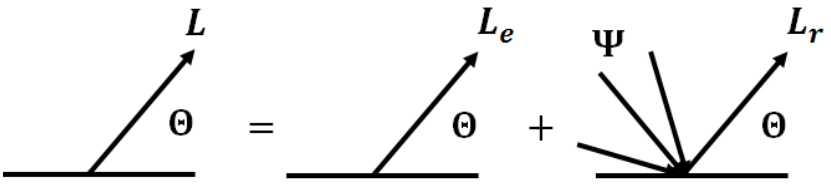


Figure 13.1: The radiance, $L(x \rightarrow \Theta)$, is computed by adding the emitted radiance, $L_e(x \rightarrow \Theta)$, and the reflected radiance, $L_r(x \rightarrow \Theta)$.

where $L(x \leftarrow \Psi)$ is a radiance arriving at x from the incoming direction, Ψ, $\cos\theta_x$ is used to consider the angle between the incoming direction and the surface normal, and the BRDF $f_r(\cdot)$ returns the outgoing radiance given its input.

The rendering equation can be represented in different manners including hemispherical or area integration.

We use the hemispherical coordinates to derive the rendering equation shown in Eq. 13.1, known as hemispherical integration. In some cases, a different form of the rendering equation, specifically area integration, is used. We consider the area integration of the rendering equation in the following section.

13.2 *Area Formulation*

To derive the hemispherical integration of the rendering equation, we used differential solid angles to consider all the possible incoming light direction to the location x. We now derive the area integration of the rendering equation by considering a differential area unit, in a similar manner using the differential solid angle unit.

Let us introduce a visible point, y, given the negated direction, $-\Psi$, of an incoming ray direction, Ψ, from the location x (Fig. 13.2). We can then have the following equation thanks to the invariance of radiance:

$$L(x \leftarrow \Psi) = L(y \rightarrow -\Psi). \tag{13.2}$$

Our intention is to integrate any incoming light directions based on y. To do this, we need to substitute the differential solid angle by the differential area. By the definition of the solid angle, we have the following equation:

$$dw_\Psi = \frac{dA\cos\theta_y}{r_{xy}^2}, \tag{13.3}$$

where θ_y is the angle between the differential area dA and the orthogonal area from the incoming ray direction, and r_{xy} is the distance between x and y.

When we plug the above two equations, we have the following equation:

$$L_r(x \rightarrow \Theta) = \int_y L(y \rightarrow -\Psi) f_r(x, \Psi \rightarrow \Theta) \frac{\cos\theta_x \cos\theta_y}{r_{xy}^2} dA, \tag{13.4}$$

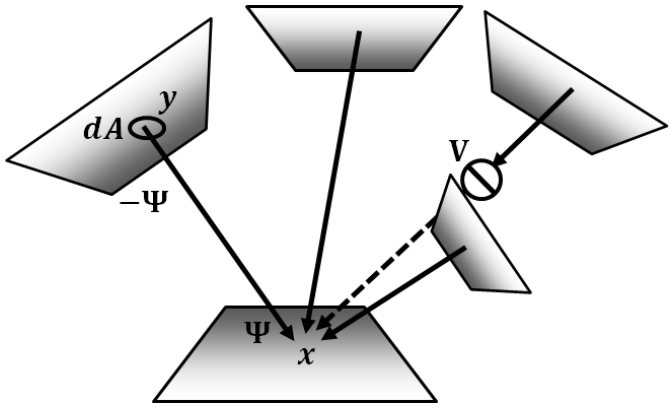


Figure 13.2: This figure shows a configuration for deriving the area formulation of the rendering equation.

where y is any visible area on triangles from x. In the above equation, we need to first compute visible areas from x on triangles. Instead, we would like to integrate the equation on any possible area while considering visibility, $V(x, y)$, which is 1 when y is visible from x, and 0 otherwise. We then have the following area integration of the rendering equation:

$$L_r(x \to \Theta) = \int_A L(y \to -\Psi) f_r(x, \Psi \to \Theta) \frac{\cos\theta_x \cos\theta_y}{r_{xy}^2} V(x, y) dA, \tag{13.5}$$

where A indicates any area on triangles.

Form factor. The radiosity algorithm requires to compute form factors that measure how much light from a patch is transferred to another patch (Sec. 11.2). The area integration of the rendering equation (Eq. 13.5) is equivalent to a form factor between a point on a surface and any points on another surface, while a diffuse BRDF is used in the equation. For the form factor between two surfaces, we simply perform one more integration over the surface.

14
Monte Carlo Integration

In this chapter, we study Monte Carlo integration to evaluate complex integral functions such as our rendering equation. In the next chapter, we will discuss Monte Carlo based ray tracing techniques that are specialized techniques for evaluating the rendering equations.

The rendering equation (Eq. 13.1) is a complex integration function. First of all, to compute a radiance for a ray starting from a surface point x, we need to integrate all the incoming radiances that arrive at x. Moreover, evaluating those incoming radiances requires us to evaluate the same procedure in a recursive way. Since there could be an infinite number of light paths starting from a light source to the eye, it is almost impossible to find an analytic solution for the rendering equation, except simple cases.

Rendering equations can be high dimensional, since we need to consider motion blur and many other effects with time and complex camera lens.

Second, the rendering equation can be high dimensional. The rendering equation shown in Eq. 13.1 is two dimensional. In practice, we need to support the motion blur for dynamic models and moving cameras. Considering such motion blur, we need to integrate radiance over time in each pixel, resulting in three dimensional rendering equation. Furthermore, supporting realistic cameras requires two or more additional dimensions on the equation. As a result, the equation for generating realistic images and video could be five or more dimensional.

Due to these issues, high dimensionality and infinite number of possible light paths, deriving analytic solutions and using deterministic approaches such as quadrature rules are impossible for virtually all of rendering environments that we encounter. Monte Carlo integration was proposed to integrate such high-dimensional functions based on random samples.

Overall, Monte Carlo (MC) integration is a numerical solution to integrate high complex and high-dimensional function. Since it uses sampling, it has stochastic errors, commonly quantified as Mean Squared Error (MSE). Fortunately, MC integration is unbiased,

indicating that it gives us a correct solution with an infinite number of samples on average.

14.1 MC Estimator

Suppose that we have the following integration, whose solution is I:

$$I = \int_a^b f(x)dx. \tag{14.1}$$

The goal of MC integration is to take N different random samples, x_i, that follow the same probability density function, $p(x_i)$. We then use the following estimator:

$$\hat{I} = \frac{1}{N}\sum_i \frac{f(x_i)}{p(x_i)}. \tag{14.2}$$

We now discuss how the MC estimator is good. One of measures for this goal is Mean Squared Error (MSE), measuring the difference between the estimated values, $\hat{Y}_i$, and observed, real values, Y_i:

$$MSE(\hat{Y}) = E[(\hat{Y} - Y)^2] = \frac{1}{N}\sum_i (\hat{Y}_i - Y_i)^2. \tag{14.3}$$

MSE can be decomposed into bias and variances terms as the following:

$$\begin{aligned} MSE(\hat{Y}) &= E\left[\left(\hat{Y} - E[\hat{Y}]\right)^2\right] + \left(E(\hat{Y}) - Y\right)^2 && (14.4)\\ &= Var(\hat{Y}) + Bias(\hat{Y}, Y)^2. && (14.5)\end{aligned}$$

The bias term $Bias(\hat{Y}, Y)$ measures how much the average value of the estimator $\hat{Y}$ is away from its ground-truth value Y. On other hand, the variance term $Var(\hat{Y})$ measures how the estimator values are away from its average values. We would like to discuss bias and variance of the MC estimator (Eq. 14.2).

Bias of the MC estimator. The MC estimator is unbiased, i.e., on average, it returns the correct solution, as shown in below:

$$\begin{aligned} E[\hat{I}] &= E\left[\frac{1}{N}\sum_i \frac{f(x_i)}{p(x_i)}\right] \\ &= \frac{1}{N}\int \sum_i \frac{f(x_i)}{p(x_i)} p(x)dx \\ &= \frac{1}{N}\sum_i \int \frac{f(x)}{p(x)} p(x)dx, \because x_i \text{ samples have the same } p(x) \\ &= \frac{N}{N}\int f(x)dx = I. \end{aligned} \tag{14.6}$$

Variance of the MC estimator. To derive the variance of the MC estimator, we utilize a few properties of variance. Based on those properties, and Independent and Identically Distributed samples (IID) of random samples, the variance of the MC estimator can be derived as the following:

$$\begin{aligned}
Var(\hat{I}) &= Var(\frac{1}{N}\sum_i \frac{f(x_i)}{p(x_i)}) \\
&= \frac{1}{N^2} Var(\sum_i \frac{f(x_i)}{p(x_i)}) \\
&= \frac{1}{N^2}\sum_i Var(\frac{f(x_i)}{p(x_i)}), \because x_i \text{ samples are independent from each other.} \\
&= \frac{1}{N^2} N Var(\frac{f(x)}{p(x)}), \because x_i \text{ samples are from the same distribution.} \\
&= \frac{1}{N} Var(\frac{f(x)}{p(x)}) = \frac{1}{N}\int \left(\frac{f(x)}{p(x)} - E\left[\frac{f(x)}{p(x)}\right]\right)^2 p(x)dx. \quad (14.7)
\end{aligned}$$

As can be in the above equations, the variance of the MC estimator decreases as a function of $\frac{1}{N}$, where N is the number of samples.

Simple experiments with MC estimators. Suppose that we would like to compute the following, simple integration:

$$I = \int_0^1 4x^3 dx = 1. \quad (14.8)$$

We know its ground truth value, 1, for the integration. We can now study various properties of the MC estimator by comparing its result against the ground truth. When we use the uniform sampling on the integration domain, the MC estimator is defined as the following:

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} 4x_i^3, \quad (14.9)$$

where $p(x_i) = p_x = 1$, since the sampling domain is $[0, 1]$, and the integration of uniform sampling on the domain has to be one, $\int_0^1 p_x = 1$.

Fig. 14.1 shows how the MC estimator behaves as we have more samples, N. As can be seen, MC estimators approach to its ground truth value, as we have more samples. Furthermore, when we measure the mean and variance of different MC estimators that have different random numbers given the same MC estimator equation (Eq. 14.9), their mean and variance shows the expected behaviors; its mean is same to the ground truth and the variance decreases as a function of $\frac{1}{N}$, respectively.

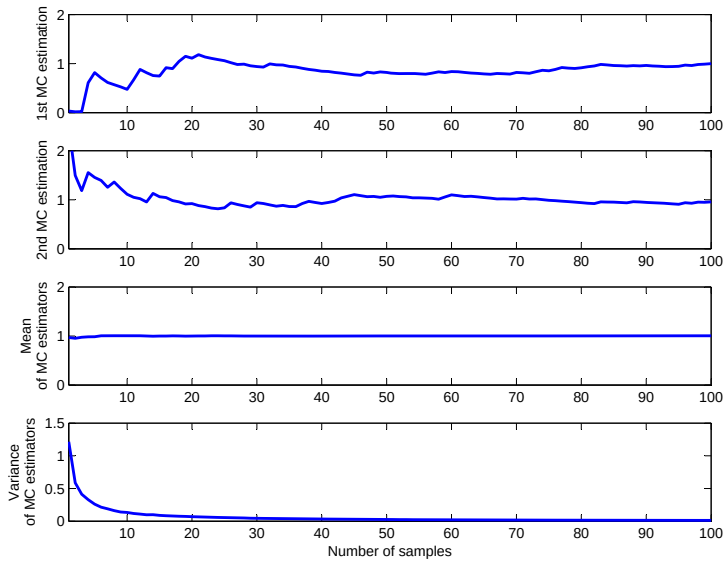


Figure 14.1: The top two subfigures show the first and second MC estimators of $\int_0^1 4x^3 dx$, whose ground truth value is 1. These MC estimators approach to their ground-truth, as we have more number of samples. While these individual MC estimators have up and down depending on their randomly generated values, their mean and variance measured with 600 estimators show the expected behavior, as theoretically predicted in Sec. 14.1. Its source code, mc_int_ex.m, is available.

14.2 *High Dimensions*

Suppose that we have an integration with higher dimensions than one:

$$I = \int \int f(x,y) dx dy. \tag{14.10}$$

Even in this case, our MC estimator is extended straightforwardly to handle such an two-dimensional integration (and other higher ones):

$$\hat{I} = \frac{1}{N} \sum \frac{f(x_i, y_i)}{p(x_i, y_i)}, \tag{14.11}$$

where we generate N random samples following a two dimensional probability density function, $p(x,y)$. We see how to generate samples according to pdf in Sec. 14.4. This demonstrates that MC integration supports well high dimensional integrations including the rendering equation with many integration domains, e.g., image positions, time, and lens parameters.

In addition, MC integration has the following characteristics:

- **Simplicity.** We can compute MC estimators based only on point sampling. This results in very convenient and simple computation.
- **Generality.** As long as we can compute values at particular points of functions under the integration, we can use MC estimations. As a result, we can compute integrations of discontinuous functions, high dimensional functions, etc.

Example. Suppose that we would like to compute the following integration defined over a hemisphere:

$$I = \int_{\Theta} f(\Theta) dw_{\Theta}, \tag{14.12}$$

$$= \int_0^{2\pi} \int_0^{\frac{\pi}{2}} f(\theta, \phi) \sin\theta d\theta d\phi. \tag{14.13}$$

where Θ is the hemispherical coordinates, (θ, ϕ).

The MC estimator for the above integration can be defined as follows:

$$\hat{I} = \frac{1}{N} \sum \frac{f(\theta_i, \phi_i) \sin\theta}{p(\theta_i, \phi_i)}, \tag{14.14}$$

where we generate (θ_i, ϕ_i) following $p(\theta_i, \phi_i)$.

Now let's get back to the irradiance example mentioned in Sec. 12.2. The irradiance equation we discussed in the irradiance example is to use $L_s \cos\theta$ for $f(\theta, \phi)$. In this case, the MC estimator of Eq. 14.14 is transformed to:

$$\hat{I} = \frac{1}{N} \sum \frac{L_s \cos\theta \sin\theta}{p(\theta_i, \phi_i)}. \tag{14.15}$$

One can use different pdf $p(\theta, \phi)$ for the MC estimator, but we can use the following one:

$$p(\theta_i, \phi_i) = \frac{\cos\theta \sin\theta}{\pi}, \tag{14.16}$$

where the integration of the pdf in the domain is one: i.e., $\int_0^{2\pi} \int_0^{\frac{\pi}{2}} \frac{\cos\theta \sin\theta}{\pi} = 1$. Plugging the pdf into the estimator of Eq. 14.14, we get the following:

$$\hat{I} = \frac{\pi}{N} \sum L_s. \tag{14.17}$$

14.3 *Importance Sampling*

In this section, we see how different pdfs affect variance of our MC estimators. As we see in Sec. 14.1, our MC estimator is unbiased regardless of pdf employed, i.e., its mean value becomes the ground truth of the integration. Variances, however, vary depending on chosen pdf.

Let's see the example integration, $I = \int_0^1 4x^3 dx = 1$, again. In the following, we test three different pdfs and see their variance:

- $p(x) = 1$. As the simplest choice, we can use the uniform distribution on the domain. The variance of our MC estimator, $\hat{I} = \frac{1}{N} \sum_i 4x_i^3$ is $\frac{36}{28N} \approx \frac{1.285}{N}$, according to the variance equation (Eq. 14.7).

- $p(x) = x$**.** The variance of this MC estimator, $\frac{1}{N}\sum_i 4x^2$, is $\frac{14}{12N} \approx \frac{1.666}{N}$. Its variance is reduced from the above, uniform pdf!
- $p(x) = 4x^3$**.** The shape of this pdf is same to the underlying function under the integration. In this case, its variance turns out to be zero.

The variance of an MC estimator goes to zero, when the shape of its pdf is same to the underlying function under the integration. We, however, do not know such a shape of the rendering equation!

As demonstrated in the above examples, the variance of a pdf decreases, as the distribution of a pdf gets closer to the underlying function $f(x)$. Actually, when the pdf $p(x)$ is set to be $\frac{f(x)}{\int f(x)dx} = \frac{f(x)}{I}$, the ideal distribution, we get the lowest variance, zero. This can be shown as the following:

$$\begin{aligned} Var(\hat{I}) &= \frac{1}{N}\int (\frac{f(x)}{p(x)} - I)^2 p(x)dx \\ &= \frac{1}{N}\int (I - I)^2 p(x)dx \\ &= 0. \end{aligned} \tag{14.18}$$

Unfortunately, in some cases, we do not know the shape of the function under the integration. Especially, this is the case for the rendering equation. Nonetheless, the general idea is to generate more samples on high values on the function, since this can reduce the variance of our MC estimator, as demonstrated in aforementioned examples. In the same reason, when the pdf is chosen badly, the variance of our MC estimator can even go higher.

The main idea of importance sampling is to generate more samples on high values on the function.

This is the main idea of importance sampling, i.e., generate more samples on high values on the underlying function, resulting in a lower variance.

Fortunately, we can intuitively know which regions we can get high values on the rendering equation. For example, for the light sources, we can get high radiance values, and we need to generate rays toward such light sources to reduce the variance in our MC estimators. Technical details on importance sampling are available in Ch. 14.3.

14.4 Generating Samples

We can use any pdf for the MC estimator. In the case of the uniform distribution, we can use a random number generator, which generates random numbers uniformly given a range.

The question that we would like to ask in this section is how we can generate samples according to the pdf $p(x)$ different from the uniform pdf.

Fig. 14.2 shows a pdf and its cdf (cumulative distribution function) in a discrete setting. Suppose that we would like to generate samples

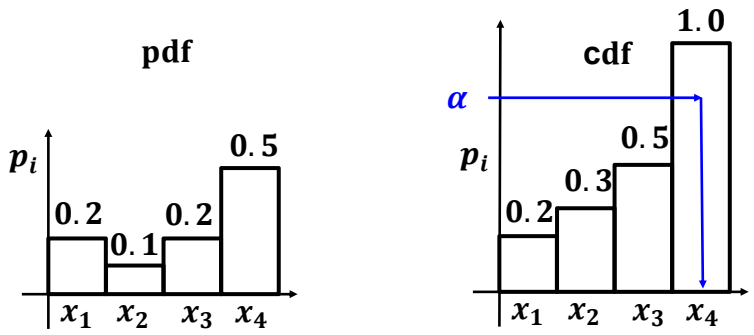


Figure 14.2: This figure shows a pdf and its cdf. Using the inverse cumulative distribution function generates samples according to the pdf by utilizing its cdf.

according to the pdf. In this case, x_1, x_2, x_3, x_4 are four events, whose probabilities are $0.2, 0.1, 0.2, 0.5$, respectively. In other words, we would like to generate those events with the pre-defined pdf.

We can use an inverse cumulative distribution function to generate samples according to a pdf.

A simple method of generating samples according to the pdf is to utilize its cdf (Fig. 14.2). This is known to be inverse cumulative distribution function. In this method, we first generate a random number α uniformly in the rnage of [0, 1). When the random number α is in the range $[\sum_0^{i-1} p_i, \sum_0^i p_i)$, we return a sample of x_i.

Let's see the probability of generating a sample x_i in this way to be p_i, as the following:

$$\begin{aligned} p(x_i) &= p(\alpha \in [\sum_0^{i-1} p_i, \sum_0^i p_i]) \\ &= p(\sum_0^i p_i) - p(\sum_0^{i-1} p_i) \\ &= p_i, \end{aligned} \tag{14.19}$$

where p_0 is set to be zero. So far, we see the discrete case, and we now extend it to the continuous case.

Continuous case. Suppose that we have a pdf, $p(x)$. Its cdf function, $F_X(x)$, is defined as $F_X(x) = p(X < x) = \int_{-\infty}^{x} p(x)dx$. We then generate a random number α uniformly in a range $[0, 1]$. A sample, y, is generated as $y = F_X^{-1}(\alpha)$.

Example for the diffuse emitter. Let's consider the following integration of measuring the irradiance with the diffuse emitter and our

sampling pdf:

$$I = \frac{1}{\pi}\int_{\Theta} dw_{\Theta},$$
$$= \frac{1}{\pi}\int_0^{2\pi}\int_0^{\frac{\pi}{2}} \sin\theta\cos\theta d\theta d\phi. \quad (14.20)$$
$$p(\theta,\phi) = \frac{\sin\theta\cos\theta}{\pi}, \quad (14.21)$$

where $\int\int p(\theta,\phi)d\theta d\phi = 1$.

Our goal is to generate samples according to the chosen pdf. We first compute its cdf, $CDF(\theta,\phi)$, as the following:

$$CDF(\theta,\phi) = \int_0^{\phi}\int_0^{\theta} \frac{\sin\theta\cos\theta}{\pi} d\theta d\phi$$
$$= (1-\cos^2\theta)\frac{\phi}{2\pi} = F(\theta)F(\phi), \quad (14.22)$$

where $F(\theta)$ and $F(\pi)$ are $(1-\cos^2\theta)$ and $\frac{\phi}{2\pi}$, respectively. Since the pdf is two dimensional, we generate two random numbers, α and β. We then utilize inverse function of those two separated functions of $F(\theta)$ and $F(\phi)$:

$$\theta = F^{-1}(\alpha) = \cos^{-1}\sqrt{1-\alpha},$$
$$\phi = F^{-1}(\beta) = 2\pi\beta. \quad (14.23)$$

The aforementioned, the inverse CDF method assumes that we can compute the inverse of the CDF. In some cases, we cannot compute the inverse of CDFs, and thus cannot use the inverse CDF method. In this case, we can use the rejection method.

In the rejection method, we first generate two random numbers, α and β. We accept β, only when $\alpha \leq p(\beta)$ (Fig. 14.3). In the example of Fig. 14.3, the ranges of α and β are $[0,1]$ and $[a,b]$. In this approach, we can generate random numbers β according to the pdf $p(x)$ without using its cdf. Nonetheless, this approach can be inefficient, especially when we do not accept and thus reject samples. This inefficiency occurs when the value of $p(x)$ is smaller than the upper bound, which we generate such random numbers up to. The upper bound of α in our example is 1.

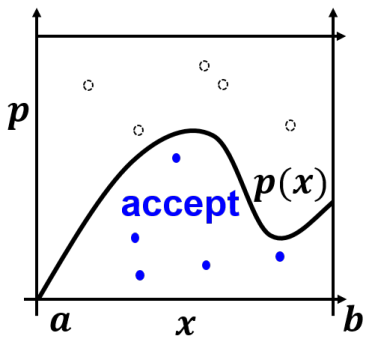


Figure 14.3: In the rejection method, we generate random numbers and accept numbers only when those numbers are within the pdf $p(x)$.

15
Monte Carlo Ray Tracing

In the prior chapters, we have discussed the rendering equation, which is represented in a high dimensional integral equation (Ch. 13.1), followed by the Monte Carlo integration method, a numerical approach to solve such equations (Ch. 14). In this chapter, we discuss how to use the Monte Carlo integration method to solve the rendering equation. This algorithm is known as a Monte Carlo ray tracing method. Specifically, we discuss the path tracing method that connects the eye and the light with a light path.

15.1 Path Tracing

The rendering equation shown below is a high dimensional integration equation defined over a hemisphere. The radiance that we observe from a location x to a direction Θ, $L(x \rightarrow \Theta)$, is defined as the following:

$$\begin{aligned} L(x \rightarrow \Theta) &= L_e(x \rightarrow \Theta) + L_r(x \rightarrow \Theta), \\ L_r(x \rightarrow \Theta) &= \int_{\Psi} L(x \leftarrow \Psi) f_r(x, \Psi \rightarrow \Theta) \cos\theta_x dw_{\Psi}, \end{aligned} \tag{15.1}$$

where $L_e(\cdot)$ is a self-emitted energy at the location x, $L_r(x \rightarrow \Theta)$ is a reflected energy, $L(x \leftarrow \Psi)$ is a radiance arriving at x from the incoming direction, Ψ, $\cos\theta_x$ is used to consider the angle between the incoming direction and the surface normal, and the BRDF $f_r(\cdot)$ returns the outgoing radiance given its input. Fig. 15.1 shows examples of the reflected term and its incoming radices.

$L(x \rightarrow \Theta)$ of Eq. 15.1 consists of two parts, emitted and reflected energy. To compute the emitted energy, we check whether the hit point x is a part of a light source. Depending whether it is in a light source or not, we compute its self-emitted energy.

The main problem of computing the radiance is on computing the reflected energy. It has several computational issues:

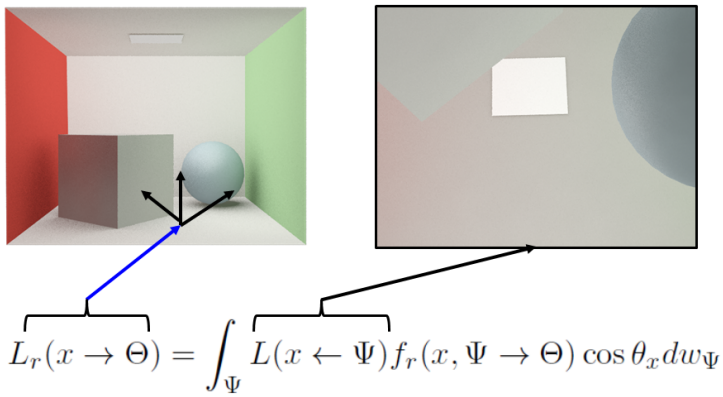


Figure 15.1: This figure shows graphical mapping between terms of the rendering equation and images. The right image represents the incoming radiance passing through the hemisphere.

1. Since the rendering equation is complex, its analytic solution is not available.
2. Computing the reflected energy requires us to compute the incoming energy $L(x \leftarrow \Psi)$, which also recursively requires us to computer another incoming energy. Furthermore, there are an infinite number of light paths from the light sources and to the eye. It is virtually impossible to consider all of them.

Since an analytic approach to the rendering equation is not an option, we consider different approaches, especially numerical approaches. In this section, we discuss the Monte Carlo approach (Ch. 14) to solve the rendering equation. Especially, we introduce path tracing, which generates a single path from the eye to the light based on the Monte Carlo method.

15.2 MC Estimator to Rendering Equation

Given the rendering equation shown in Eq. 15.1, we omit the self-emitting term $L_e(\cdot)$ for simplicity; computing this term can be done easily by accessing the material property of the intersecting object with a ray.

To solve the rendering equation, we apply the Monte Carlo (MC) approach, and the MC estimator of the rendering equation is defined as the following:

$$\hat{L}_r(x \to \Theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{L(x \leftarrow \Psi_i) f_r(x, \Psi_i \to \Theta) \cos\theta_x}{p(\Psi_i)}, \tag{15.2}$$

where Ψ_i is a randomly generated direction over the hemisphere and N is the number of random samples generated.

To evaluate the MC estimator, we generate a random incoming direction Ψ_i, which is uniformly generated over the hemisphere. We

Figure 15.2: Top: computing the outgoing radiance from x requires us to compute the radiance from y to x, which is also recursively computed by simulating additional bounce to y. Bottom: this sequence visualizes rendering results by considering the direct emission and single, double, and triple bounces, adding more energy to the image. Images are excerpted from slides of Prof. Bala.

then evaluate BRDF $f_r(\cdot)$ and the cosine term. The question is how to compute the radiance we can observe from the incoming direction $L(x \leftarrow \Psi_i)$. To compute the radiance, we compute a visible point, y, from x toward Ψ_i direction and then recursively use another MC estimator. This recursion process effectively simulates an additional bounce of photon (Fig. 15.2), and repeatedly performing this process can handle most light transports that we observe in our daily lives.

The aforementioned process uses the recursion process and can simulate various light transport. The recursion process terminates when a ray intersects with a light source, establishing a light path from the light source to the eye. Unfortunately, hitting the light source can have a low probability and it may require an excessive amount of recursion and thus computational time.

Many heuristics are available to break the recursion. Some of them uses a maximum recursion depth (say, 5 bounces) and uses some thresholds on radiance difference to check whether we go into a more recursion depth. These are easy to implement, but using these heuristics and simply ignoring radiances that we can compute with additional bounces results in bias in our MC estimator. To terminate the recursion process without introducing a bias, Russian roulette is introduced.

Russian roulette. Its main idea is that we artificially introduce a case where we have zero radiance, which effectively terminate recursion process. The Russian roulette method realizes this idea without introducing a bias, but with an increased variance. Suppose that we aim to keep the recursion P percentage (e.g., 95%), i.e., cancel the

recursion $1 - P$ percentage. Since we lose some energy by terminating the recursion, we increase the energy when we accept the recursion, in particular, $\frac{1}{P}$, to compensate the lost energy.

In other words, we use the following estimator:

$$\hat{I}_{roulette} = \begin{cases} \frac{f(x_i)}{P} & \text{if } x_i \leq P, \\ 0 & \text{if } x_i > P. \end{cases} \tag{15.3}$$

One can show its bias to be zero, but also show that the original integration is reformulated as the following with a substitute, $y = Px$:

$$I = \int_0^1 f(x)dx = \int_0^P \frac{f(y/P)}{P} dy. \tag{15.4}$$

While the bias of the MC estimate with the Russian roulette is zero, its variance is higher than the original one, since we have more drastic value difference, zero value in a region, while bigger values in other regions, on our sampling.

A left issue is how to choose the constant of P. Intuitively, P is related to the reflectance of the material of a surface, while $1 - P$ is considered as the absorption probability. As a result, we commonly set P as the albedo of an object. For example, albedo of water, ice, and snow is approximately about 7%, 35%, and 65%, respectively.

Branching factor. We can generate multiple ray Samples Per Pixel (SPP). For each primary ray sample in a pixel, we compute its hit point x and then need to estimate incoming radiance to x. The next question is how many secondary rays we need to generate for estimating the incoming radiance well. This is commonly known as a branching factor. Intuitively, generating more secondary rays, i.e., having a higher branching factor, may result in better estimation of incoming radiance. In practice, this approach turns out to be less effective than having a single branching factor, generating a single secondary ray. This is because while we have many branching factors, their importance can be less significant than other rays, e.g., primary ray. This is related to importance sampling (Ch. 14.3) and is discussed more there.

Path tracing is one of simple MC ray tracing for solving the rendering equation. Since it is very slow, it is commonly used for generating the reference results compared to other advanced techniques.

Path tracing. The rendering algorithm with a branching factor of one is called path tracing, since we generate a light path from the eye to the light source. To perform path tracing, we need to set the number of ray samples per pixel (SPP), while the branching factor is set to be one. Once we have N samples per each pixel, we apply the MC estimator, which is effectively the average sum of those N sample values, radiance.

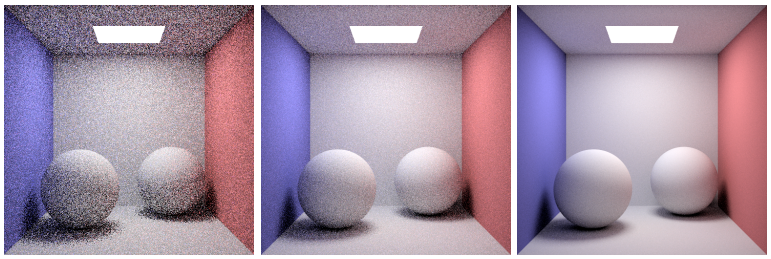


Figure 15.3: This figure shows images that are generated with varying numbers of samples per each pixel. Note that direct illumination sampling, generate a ray toward the light (Sec. 16.1), is also used. From the left, 1 spp (sample per pixel), 4 spp, and 16 spp are used.

Fig. 15.3 shows rendering results with different number of ray samples per pixel. As we use more samples, the variance, which is observed as noise, is reduced.

The theory tells us that as we generate more samples, the variance is reduced more, but it requires a high number of samples and long computational time. As a result, a lot of techniques have been developed to achieve high-quality rendering results while reducing the number of samples.

Programming assignment. It is very important to see how the rendering results vary as a function of ray samples and a different types of sampling methods. Fortunately, many ray tracing based rendering methods are available. Some of well known techniques are Embree, Optix, and pbrt (Sec. 9.6). Please download one of those softwares and test the rendering quality with different settings. In my own class, I ask my students to download pbrt and test uniform sampling and an adaptive sampling method that varies the number of samples. Also, measuring its error compared to a reference image is important to analyze different rendering algorithms in a quantitative manner. I therefore ask to compute a reference image, which is typically computed by generating an excessive number of samples (e.g., 1 k or 10 k samples per pixel), and measure the mean of squared root difference between a rendering result and its reference. Based on those computed errors, we can know which algorithm is better than the other.

15.2.1 Stratified Sampling

We commonly use a uniform distribution or other probability density function to generate a random number. For the sake of simple explanation, let assume that we use a uniform sampling distribution on a sampling domain. While those random numbers in a domain, say, $[0, 1)$, are generated in a uniform way, some random numbers can be arbitrarily close to each other, resulting in noise in the estimation.

A simple method of ameliorating this issue is to use stratified sampling, also known as jittered sampling. Its main idea is to partition

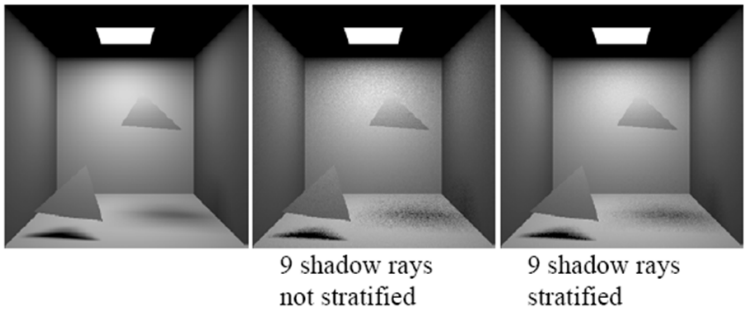


Figure 15.4: The reference image is shown on the leftmost, while images with and without stratified sampling are shown on the right. Images are excerpted from slides of Prof. Bala.

the original sampling domains into multiple regions, say, $[0, 1/2)]$ and $[1/2, 1)$, and perform sampling in those regions independently.

While this approach cannot avoid a close proximity of those random samples, it has been theoretically and experimentally demonstrated to reduce the variance of MC estimators. Fig. 15.4 shows images w/ and w/o using stratified sampling. We can observe that the image with stratified sampling shows less noise.

Theoretically, stratified sampling is shown to reduce the variance over the non-stratified approach. Suppose X to be a random variable representing values of our MC sampling. Let k to be the number of partitioning regions of the original sampling domain, and Y to be an event indicating which region is chosen among k different regions. We then have the following theorem:

Theorem 15.2.1 (Law of total variance)**.** $Var[X] = E(Var[X|Y]) + Var(E[X|Y])$.

Proof.

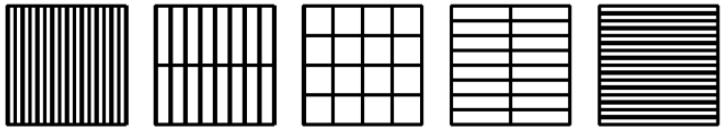
$$\begin{aligned} Var[X] &= E[X^2] - E[X]^2 \\ &= E[E[X^2|Y]] - E[E[X|Y]]^2, \because \text{Law of total expectation} \\ &= E[Var[X|Y]] + E[E[X|Y]^2] - E[E[X|Y]]^2, \\ &= E[Var[X|Y]] + Var(E[X|Y]). \end{aligned} \tag{15.5}$$

□

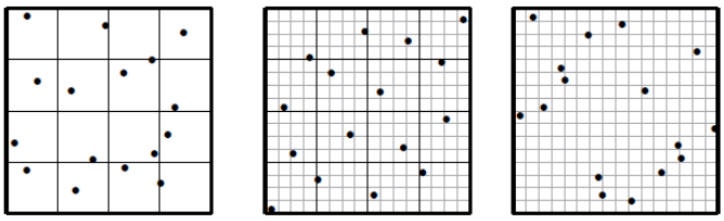
According to the law of total variance, we can show that the variance of the original random variance is equal to or less than the variance of the random variance in each sub-region.

$$Var[X] \geq E(Var[X|Y]) = \frac{1}{k} k Var[X|Y_r] = Var[X|Y_r], \tag{15.6}$$

where Y_r is an event indicating that random variances are generated given each sub-region, and we assume iid for those sub-regions.



(a) All the elementary intervals with the volume of $\frac{1}{16}$.



(b) This figure shows sampling patterns of jittered, Sobol, and N-Rooks samplings, respectively from the left.

Figure 15.5: These images are excerpted from the cited paper.

N-Rooks sampling. N-Rooks sampling or Latin hypercube sampling is a variant of stratified sampling with an additional requirement that has only a single sample in each row and column of sampling domains. An example of N-Rooks sampling is shown in Fig. 15.5. For stratified sampling, we generate N^d samples for a d-dimensional space, where we generate N samples for each space. On the other hand, since it generates only a single sample per each column and row, we can arbitrary generate N samples when we create N columns and rows for high dimensional cases.

Sobol sequence. Sobol sequence is designed to maintain additional constraints for achieving better uniformity. It aims to generate a single sample on each elementary interval. Instead of giving its exact definition, we show all the elementary intervals having the volume of $\frac{1}{16}$ in the 2 D sampling space in Fig. 15.5; images are excerpted from [1].

[1] Thomas Kollig and Alexander Keller. Efficient multidimensional sampling. *Comput. Graph. Forum*, 21(3):557–563, 2002

15.3 *Quasi-Monte Carlo Sampling*

Quasi-Monte Carlo sampling is another numerical tool to evaluate integral interactions such as the rendering equation. The main difference over MC sampling is to use deterministic sampling, not random sampling. While quasi-Monte Carlo sampling uses deterministic sampling, those samples are designed to look random.

The main benefit of using quasi-Monte Carlo sampling is that we can have a particular guarantee on error bounds, while MC methods do not. Moreover, we can have a better convergence to Monte Carlo sampling, especially, when we have low sampling dimensions and need to generate many samples [2].

Specifically, the probabilistic error bound of the MC method

[2] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, 1992

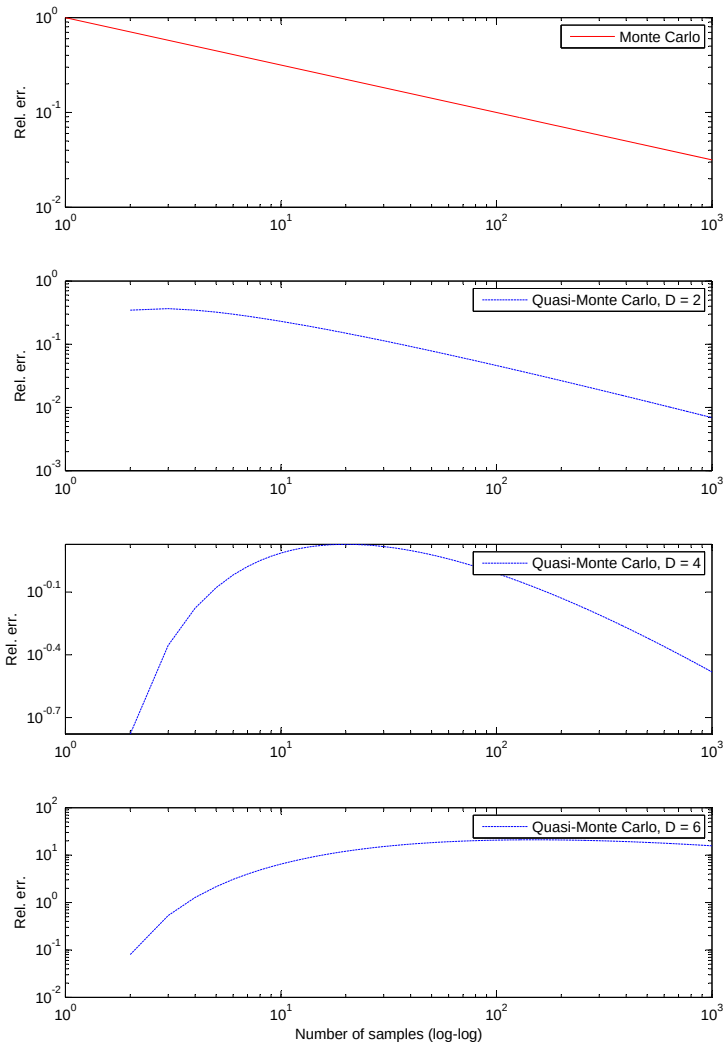


Figure 15.6: This figure shows error behavior of MC and quasi-Monte Carlo methods. They are not aligned in the same error magnitude. As a result, only shapes of these curves are meaningful. The basic quasi-Motel Carlo shows better performance than MC on low dimensional spaces (e.g, two).

reduces $O(\frac{1}{\sqrt{N}})$. On the other hand, the quasi-Monte Carlo can provide a deterministic error bound of $O(\frac{\log N^{D-1}}{N})$ for a well chosen set of samples and for integrands with a low degree of regularity, where D is the dimensionality. Better error bounds are also available for integrands with higher regularity.

Fig. 15.6 shows shapes of two different error bounds of Monte Carlo and quasi-Monte Carlo. Note that they are not aligned in the same error magnitude, and thus only their shapes are meaningful. Furthermore, the one of MC is a probabilistic bound, while that of quasi-Monte Carlo is a deterministic bound. The quasi-Monte Carlo has demonstrated to show superior performance than MC on low dimensional sample space (e.g., two). On the other hand, for a high dimensional case, say six dimensional case, the quasi-Monte Carlo is

not effectively reducing its error on a small number of samples.

The question is how to construct such a deterministic sampling pattern than looks like random and how to quantify such pattern? A common approach for this is to use a discrepancy measure that quantifies the gap, i.e. discrepancy, between the generated sampling and an ideal uniform and random sequence. Sampling methods realizing low values for the discrepancy measure is low-discrepancy sampling.

Various stratified sampling techniques such as Sobol sequence is also used as a low-discrepancy sampling even for the quasi-Monte Carlo sampling, while we use pre-computed sampling pattern and do not randomize during the rendering process. In additional to that, other deterministic techniques such as Halton and Hammersley sequences are used. In this section, we do not discuss these techniques in detail, but discuss the discrepancy measure that we try to minimize with low-discrepancy sampling.

For the sake of simplicity, suppose that we have a sequence of points $P = \{x_i\}$ in a one dimensional sampling space, say $[0, 1]$. The discrepancy measure, $D_N(P, x)$, can be defined as the following:

$$D_N(P, x) = |x - \frac{n}{N}|, \tag{15.7}$$

where $x \in [0, 1]$ and n is the number of points that are in $[0, x]$. Intuitively speaking, we can achieve uniform distribution by minimizing this discrepancy measure. Its general version is available at the book of Niederreiter [3];see pp. 14.

[3] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, 1992

Randomized quasi-Monte Carlo integration. While quasi-Monte Carlo methods have certain benefits over Monte Carlo approaches, it also has drawbacks. Some of them include 1) it shows better performance over MC methods when we have smaller dimensions and the number of samples are high, and 2) its deterministic bound are rather complex to compute. Also, many other techniques (e.g., reconstruction) are based on stochastic analysis and thus the deterministic nature may result in lose coupling between different rendering modules.

To address the drawbacks of quasi-Monte Carlo approaches, randomization on those deterministic samples by permutation can be applied. This is known as randomized quasi-Monte Carlo techniques. For example, one can permute cells of 2 D sample patterns of the Sobol sequence and can generate a randomized sampling pattern. We can then apply various stochastic analysis and have an unbiased estimator. Fig. 15.7 shows error reduction rates of different sampling methods; images are excepted from [4].

[4] Thomas Kollig and Alexander Keller. Efficient multidimensional sampling. *Comput. Graph. Forum*, 21(3):557–563, 2002

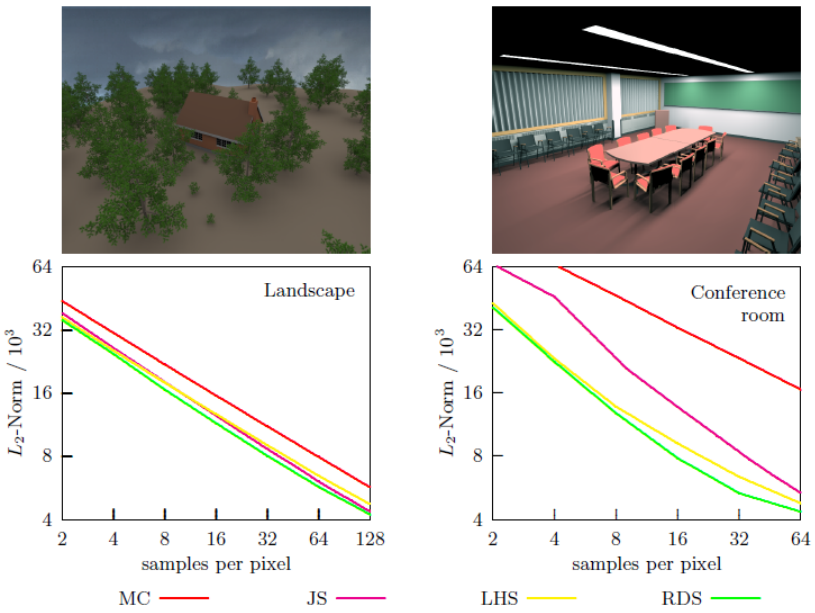
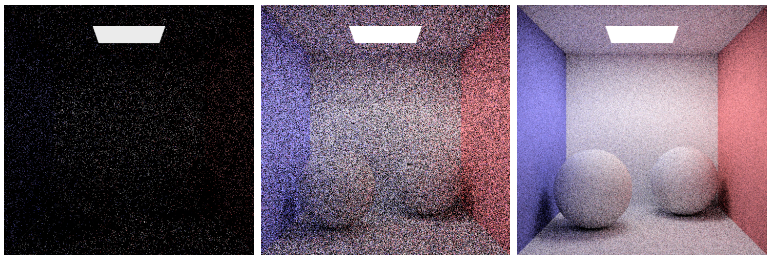


Figure 15.7: These graphs show different error reduction rates of Monte Carlo (MC), jittered (JS), Latin hypercube (LHS), and randomized Sobol sequence (RDS). These techniques are applied to four dimensional rendering problems with direct illumination.

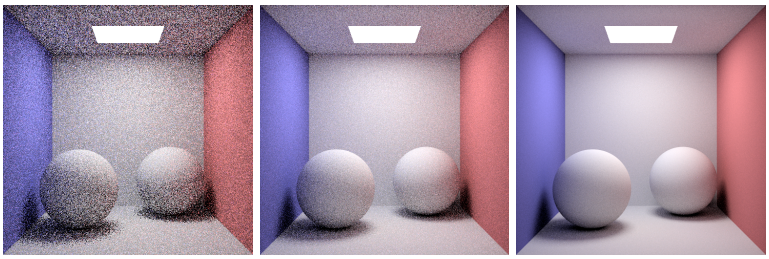
16
Importance Sampling

In the last chapter, we discussed Monte Carlo (MC) ray tracing, especially, path tracing that generates a light path from the camera to the light source. While it is an unbiased estimator, it has significant variance, i.e., noise, when we have a low ray samples per pixel. To reduce the noise of MC generated images, we studied quasi-Monte Carlo technique in Sec. 15.3.

In this chapter, as an effective way of reducing the variance, we discuss importance sampling. We first discuss an importance sampling method considering light sources, called direct illumination method. We then discuss other importance sampling methods considering various factors of the rendering equation.



(a) Results w/o direct illumination. From the left, 1 spp, 4 spp, and 16 spp are used.



(b) Results w/ direct illumination.

Figure 16.1: These images are generated by path tracer w/ and w/o direct illumination. They are created by using a path tracer created by Ritchie et al. `http://web.stanford.edu/~dritchie/path/index.html`.

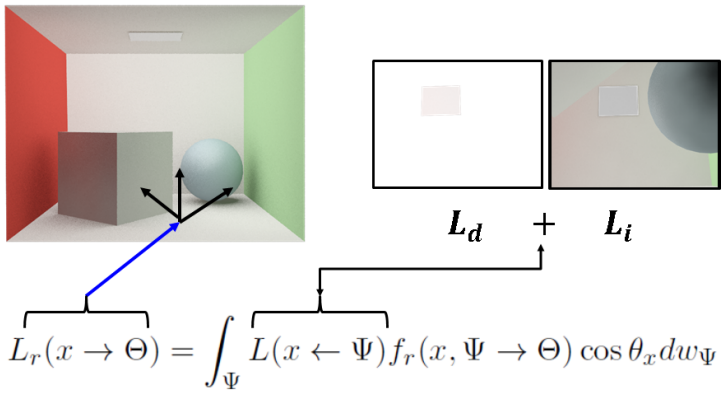


Figure 16.2: This figure illustrates the factorization of the reflected radiance into direct and indirect illumination terms.

16.1 *Direct Illumination*

Fig. 16.1 show rendering results w/ and w/o direct illumination. The first row shows rendering results w/o direct illumination under 1, 4, and 16 spp. In this scene, we adopt path tracing and observe severe noise even when we use 16 spp. This noise is mainly from the variance of the MC estimator. Note that we use random sampling on the hemisphere to generate a reflected ray direction, and it can keep bounce unless arriving at the light source located at the ceiling of the scene. Furthermore, since we are using the Russian roulette, some rays can be terminated without carrying any radiance, resulting in dark colors.

A better, yet intuitive approach is to generate a ray directly toward the light source, since we know that the light source is emitting energy and brightens the scene. The question is how we can accommodate this idea within the MC estimation framework! If we just generate a ray toward the light source, it will introduce a bias and we may not get a correct result, even when we generate an infinite number of samples.

Let's consider the rendering equation that computes the radiance $L(x \to \Theta)$, from a location x in the direction of Θ [1]. The radiance is composed of the self-emitted energy and reflected energy (Fig. 13.1):

[1] This notation is introduced in Sec. 13.1

$$L(x \to \Theta) = L_e(x \to \Theta) + L_r(x \to \Theta). \tag{16.1}$$

For the reflected term $L_r(\cdot)$, we decompose it into two terms: direct illumination term, $L_d(\cdot)$, and indirect illumination term, $L_i(\cdot)$:

$$L_r(x \to \Theta) = L_d(x \to \Theta) + L_i(x \to \Theta). \tag{16.2}$$

Fig. 16.2 illustrates an example of this decomposition.

Once we decomposed the radiance term into the direct and indirect illumination terms, we apply two separate MC estimators for

those two terms. For the direct illumination term, we cannot use the hemispherical integration described in Sec. 13.1, since we need to generate rays to the light source. For generating rays only to the light source, we use the area formulation, Eq. 13.5 explained in Sec. 13.2.

Rays corresponding to the direct illumination should be not duplicated considered for indirect illumination.

For estimating the indirect illumination, we use the hemispherical integration. The main difference to the regular hemispherical integration is that a ray generated from the hemispherical integration should not accumulate energy directly from the light source. In other words, when the ray intersects with the light source, we do not transfer the energy emitted from the light source, since the ray in this case is considered in the direct illumination term, and thus its energy should not be considered for the indirect illumination to avoid duplicate computation.

Many light problems. We discussed a simple importance sampling with the direct illumination sampling to reduce the variance of MC estimators. What if we have so many lights? In this case, generating rays to many lights can require a huge amount of time. In practice, simulating realistic scenes with complex light setting may require tens or hundreds of thousands of point light sources. This problem has been known as the many light problem. Some of simple approaches are to generate rays to those lights with probabilities that are proportional to their light intensity.

16.2 Multiple Importance Sampling

In the last section, we looked into direct illumination sampling as an importance sampling method. While it is useful, it cannot be a perfect solution, as hinted in our theoretical discussion (Sec. 14.3)

There are many other different terms in the rendering equation. Some of them are incoming radiance, BRDF, visibility, cosine terms, etc. The direct illumination sampling is a simple heuristic to consider the incoming radiance, while there could be many other strong indirect illuminations such as strong light reflection from a mirror. BRDF of an intersected object and cosine terms are available, and thus we can design importance sampling methods considering those factors. Nonetheless, these different importance sampling methods are designed separately and may work well in one case, but not in other cases.

Multiple importance sampling (MIS) is introduced to design a combined sampling method out of separately designed estimators. Suppose that there are n different sampling methods, and we allocate n_i samples for each sampling method. Given the total number of samples N, $n_i = c_i N$ with independent $X_{i,j}$ samples. The whole

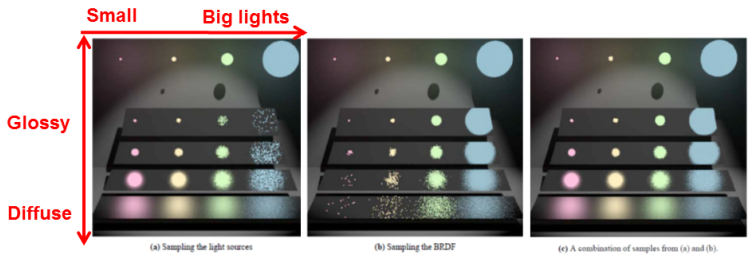


Figure 16.3: These figures show rendering results with different sampling methods. From the left, we use sampling light sources, BRDF, and both of them w/ multiple importance sampling.

distribution, $\bar{p}(x)$, combined with those n different methods, is defined as the following:

$$\bar{p}(x) = \sum_{i}^{n} c_i p_i(x), \tag{16.3}$$

where $p_i(x)$ is a i-th sampling distribution. $\bar{p}(x)$ is also called combined sample distribution [2], whose each sample $X_{i,j}$ has $1/N$ sampling probability.

By applying the standard MC estimator with the combined sampling distribution, we get the following estimator:

$$I = \frac{1}{N} \sum_{i} \sum_{n_i} \frac{f(X_{i,j})}{\bar{p}(X_{i,j})}. \tag{16.4}$$

This estimator is also derived by assigning the relative importance, i.e., probability, of a sampling method among others. In this perspective, this is also known as to be derived under balance heuristic. Surprisingly, this simple approach has been demonstrated to work quite well as shown in Fig. 16.3; these figures are excepted from the paper of Veach et al. [3]. A theoretical upper bound of the variance error of this approach is available in the original paper.

[2] Eric Veach and Leonidas J. Guibas. Optimally combining sampling techniques for monte carlo rendering. In *SIGGRAPH*, pages 419–428, 1995

[3] Eric Veach and Leonidas J. Guibas. Optimally combining sampling techniques for monte carlo rendering. In *SIGGRAPH*, pages 419–428, 1995

17
Conclusion

In this book, our discussions have revolved around two main topics: rasterization and ray tracing. These two techniques have their own pros and cons. For example, ray tracing is slower compared to rasterization, and is more natural to support a wide variety of rendering effects. We have mainly explains basic concepts on these topics, and there are many other advanced topics including scalable techniques and sub-surface scattering approaches. We plan to cover them in a coming edition.

Bibliography

Tomas Akenine-Möller, Eric Haines, and Naty Hoffman. *Real-Time Rendering 3rd Edition*. A. K. Peters, Ltd., 2008.

Arthur Appel. Some techniques for shading machine renderings of solids. In *AFIPS 1968 Spring Joint Computer Conf.*, volume 32, pages 37–45, 1968.

Philip Dutre, Kavita Bala, and Philippe Bekaert. *Advanced Global Illumination*. AK Peters, 2006.

Cindy M. Goral, Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile. Modelling the interaction of light between diffuse surfaces. In *Computer Graphics (SIGGRAPH '84 Proceedings)*, volume 18, pages 212–22, July 1984.

Paul S. Heckbert. Adaptive radiosity textures for bidirectional ray tracing. In Forest Baskett, editor, *Computer Graphics (SIGGRAPH '90 Proceedings)*, volume 24, pages 145–154, August 1990.

Jae-Pil Heo, Joon-Kyung Seong, DukSu Kim, Miguel A. Otaduy, Jeong-Mo Hong, Min Tang, and Sung-Eui Yoon. FASTCD: Fracturing-aware stable collision detection. In *SCA '10: Proceedings of the 2010 ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, 2010.

Thomas Kollig and Alexander Keller. Efficient multidimensional sampling. *Comput. Graph. Forum*, 21(3):557–563, 2002.

C. Lauterbach, S.-E. Yoon, D. Tuft, and D. Manocha. RT-DEFORM: Interactive ray tracing of dynamic scenes using bvhs. In *IEEE Symp. on Interactive Ray Tracing*, pages 39–46, 2006.

C. Lauterbach, M. Garland, S. Sengupta, D. Luebke, and D. Manocha. Fast bvh construction on gpus. *Computer Graphics Forum (EG)*, 28(2):375–384, 2009.

Tomas Möller and Ben Trumbore. Fast, minimum storage ray-triangle intersection. *J. Graph. Tools*, 1997.

H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics, 1992.

Steven G. Parker, James Bigler, Andreas Dietrich, Heiko Friedrich, Jared Hoberock, David Luebke, David McAllister, Morgan McGuire, Keith Morley, Austin Robison, and Martin Stich. Optix: a general purpose ray tracing engine. *ACM Trans. Graph.*, 29:66:1–66:13, 2010.

Matt Pharr and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation 2nd*. Morgan Kaufmann Publishers Inc., 2010a.

Matt Pharr and Greg Humphreys. *Physically Based Rendering, Second Edition: From Theory To Implementation*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2010b. ISBN 0123750792, 9780123750792.

William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, Cambridge, England, 2nd edition, 1993.

G. Sellers and J.M. Kessenich. *Vulkan Programming Guide: The Official Guide to Learning Vulkan*. Addison Wesley, 2016.

Peter Shirley and Steve Marschner. *Fundamentals of Computer Graphics*. A. K. Peters, Ltd., 3rd edition, 2009.

Peter Shirley and R. Keith Morley. *Realistic Ray Tracing*. AK Peters, second edition, 2003.

Ivan E. Sutherland, Robert F. Sproull, and Robert A. Schumacker. A characterization of ten hidden-surface algorithms. *ACM Comput. Surv.*, 6(1):1–55, 1974.

Eric Veach and Leonidas J. Guibas. Optimally combining sampling techniques for monte carlo rendering. In *SIGGRAPH*, pages 419–428, 1995.

Ingo Wald, Sven Woop, Carsten Benthin, Gregory S Johnson, and Manfred Ernst. Embree: A kernel framework for efficient cpu ray tracing. *ACM Trans. Graph.*, 2014.

E. Weisstein. From mathworld–a wolfram web resource. URL `http://mathworld.wolfram.com`.

Turner Whitted. An improved illumination model for shaded display. *Commun. ACM*, 23(6):343–349, 1980.

Sung-Eui Yoon, Brian Salomon, Russell Gayle, and Dinesh Manocha. Quick-VDR: Interactive View-dependent Rendering of Massive Models. In *IEEE Visualization*, pages 131–138, 2004.

Sung-Eui Yoon, Peter Lindstrom, Valerio Pascucci, and Dinesh Manocha. Cache-Oblivious Mesh Layouts. *ACM Transactions on Graphics (SIGGRAPH)*, 24(3):886–893, 2005.

Sungeui Yoon, Sean Curtis, and Dinesh Manocha. Ray tracing dynamic scenes using selective restructuring. *Eurographics Symp. on Rendering*, pages 73–84, 2007.

Index