Introduction

The term *computer graphics* describes any use of computers to create and manipulate images. This book introduces the algorithmic and mathematical tools that can be used to create all kinds of images—realistic visual effects, informative technical illustrations, or beautiful computer animations. Graphics can be two- or three-dimensional; images can be completely synthetic or can be produced by manipulating photographs. This book is about the fundamental algorithms and mathematics, especially those used to produce synthetic images of three-dimensional objects and scenes.

Actually doing computer graphics inevitably requires knowing about specific hardware, file formats, and usually a graphics API (see Section 1.3) or two. Computer graphics is a rapidly evolving field, so the specifics of that knowledge are a moving target. Therefore, in this book we do our best to avoid depending on any specific hardware or API. Readers are encouraged to supplement the text with relevant documentation for their software and hardware environment. Fortunately, the culture of computer graphics has enough standard terminology and concepts that the discussion in this book should map nicely to most environments.

This chapter defines some basic terminology, and provides some historical background as well as information sources related to computer graphics.

1.1 Graphics Areas

Imposing categories on any field is dangerous, but most graphics practitioners would agree on the following major areas of computer graphics:
• **Modeling** deals with the mathematical specification of shape and appearance properties in a way that can be stored on the computer. For example, a coffee mug might be described as a set of ordered 3D points along with some interpolation rule to connect the points and a reflection model that describes how light interacts with the mug.

• **Rendering** is a term inherited from art and deals with the creation of shaded images from 3D computer models.

• **Animation** is a technique to create an illusion of motion through sequences of images. Animation uses modeling and rendering but adds the key issue of movement over time, which is not usually dealt with in basic modeling and rendering.

There are many other areas that involve computer graphics, and whether they are core graphics areas is a matter of opinion. These will all be at least touched on in the text. Such related areas include the following:

• **User interaction** deals with the interface between input devices such as mice and tablets, the application, feedback to the user in imagery, and other sensory feedback. Historically, this area is associated with graphics largely because graphics researchers had some of the earliest access to the input/output devices that are now ubiquitous.

• **Virtual reality** attempts to immerse the user into a 3D virtual world. This typically requires at least stereo graphics and response to head motion. For true virtual reality, sound and force feedback should be provided as well. Because this area requires advanced 3D graphics and advanced display technology, it is often closely associated with graphics.

• **Visualization** attempts to give users insight into complex information via visual display. Often there are graphic issues to be addressed in a visualization problem.

• **Image processing** deals with the manipulation of 2D images and is used in both the fields of graphics and vision.

• **3D scanning** uses range-finding technology to create measured 3D models. Such models are useful for creating rich visual imagery, and the processing of such models often requires graphics algorithms.

• **Computational photography** is the use of computer graphics, computer vision, and image processing methods to enable new ways of photographically capturing objects, scenes, and environments.
1.2 Major Applications

Almost any endeavor can make some use of computer graphics, but the major consumers of computer graphics technology include the following industries:

- **Video games** increasingly use sophisticated 3D models and rendering algorithms.
- **Cartoons** are often rendered directly from 3D models. Many traditional 2D cartoons use backgrounds rendered from 3D models, which allows a continuously moving viewpoint without huge amounts of artist time.
- **Visual effects** use almost all types of computer graphics technology. Almost every modern film uses digital compositing to superimpose backgrounds with separately filmed foregrounds. Many films also use 3D modeling and animation to create synthetic environments, objects, and even characters that most viewers will never suspect are not real.
- **Animated films** use many of the same techniques that are used for visual effects, but without necessarily aiming for images that look real.
- **CAD/CAM** stands for computer-aided design and computer-aided manufacturing. These fields use computer technology to design parts and products on the computer and then, using these virtual designs, to guide the manufacturing process. For example, many mechanical parts are designed in a 3D computer modeling package and then automatically produced on a computer-controlled milling device.
- **Simulation** can be thought of as accurate video gaming. For example, a flight simulator uses sophisticated 3D graphics to simulate the experience of flying an airplane. Such simulations can be extremely useful for initial training in safety-critical domains such as driving, and for scenario training for experienced users such as specific fire-fighting situations that are too costly or dangerous to create physically.
- **Medical imaging** creates meaningful images of scanned patient data. For example, a computed tomography (CT) dataset is composed of a large 3D rectangular array of density values. Computer graphics is used to create shaded images that help doctors extract the most salient information from such data.
- **Information visualization** creates images of data that do not necessarily have a “natural” visual depiction. For example, the temporal trend of the
price of ten different stocks does not have an obvious visual depiction, but clever graphing techniques can help humans see the patterns in such data.

1.3 Graphics APIs

A key part of using graphics libraries is dealing with a graphics API. An application program interface (API) is a standard collection of functions to perform a set of related operations, and a graphics API is a set of functions that perform basic operations such as drawing images and 3D surfaces into windows on the screen.

Every graphics program needs to be able to use two related APIs: a graphics API for visual output and a user-interface API to get input from the user. There are currently two dominant paradigms for graphics and user-interface APIs. The first is the integrated approach, exemplified by Java, where the graphics and user-interface toolkits are integrated and portable packages that are fully standardized and supported as part of the language. The second is represented by Direct3D and OpenGL, where the drawing commands are part of a software library tied to a language such as C++, and the user-interface software is an independent entity that might vary from system to system. In this latter approach, it is problematic to write portable code, although for simple programs it may be possible to use a portable library layer to encapsulate the system specific user-interface code.

Whatever your choice of API, the basic graphics calls will be largely the same, and the concepts of this book will apply.

1.4 Graphics Pipeline

Every desktop computer today has a powerful 3D graphics pipeline. This is a special software/hardware subsystem that efficiently draws 3D primitives in perspective. Usually these systems are optimized for processing 3D triangles with shared vertices. The basic operations in the pipeline map the 3D vertex locations to 2D screen positions and shade the triangles so that they both look realistic and appear in proper back-to-front order.

Although drawing the triangles in valid back-to-front order was once the most important research issue in computer graphics, it is now almost always solved using the z-buffer, which uses a special memory buffer to solve the problem in a brute-force manner.

It turns out that the geometric manipulation used in the graphics pipeline can be accomplished almost entirely in a 4D coordinate space composed of three tra-
ditional geometric coordinates and a fourth *homogeneous* coordinate that helps with perspective viewing. These 4D coordinates are manipulated using $4 \times 4$ matrices and 4-vectors. The graphics pipeline, therefore, contains much machinery for efficiently processing and composing such matrices and vectors. This 4D coordinate system is one of the most subtle and beautiful constructs used in computer science, and it is certainly the biggest intellectual hurdle to jump when learning computer graphics. A big chunk of the first part of every graphics book deals with these coordinates.

The speed at which images can be generated depends strongly on the number of triangles being drawn. Because interactivity is more important in many applications than visual quality, it is worthwhile to minimize the number of triangles used to represent a model. In addition, if the model is viewed in the distance, fewer triangles are needed than when the model is viewed from a closer distance. This suggests that it is useful to represent a model with a varying *level of detail* (LOD).

### 1.5 Numerical Issues

Many graphics programs are really just 3D numerical codes. Numerical issues are often crucial in such programs. In the “old days,” it was very difficult to handle such issues in a robust and portable manner because machines had different internal representations for numbers, and even worse, handled exceptions in different and incompatible ways. Fortunately, almost all modern computers conform to the *IEEE floating-point* standard (IEEE Standards Association, 1985). This allows the programmer to make many convenient assumptions about how certain numeric conditions will be handled.

Although IEEE floating-point has many features that are valuable when coding numeric algorithms, there are only a few that are crucial to know for most situations encountered in graphics. First, and most important, is to understand that there are three “special” values for real numbers in IEEE floating-point:

1. **infinity** ($\infty$). This is a valid number that is larger than all other valid numbers.
2. **minus infinity** ($-\infty$). This is a valid number that is smaller than all other valid numbers.
3. **not a number** (NaN). This is an invalid number that arises from an operation with undefined consequences, such as zero divided by zero.

The designers of IEEE floating-point made some decisions that are extremely convenient for programmers. Many of these relate to the three special values
above in handling exceptions such as division by zero. In these cases an exception is logged, but in many cases the programmer can ignore that. Specifically, for any positive real number \( a \), the following rules involving division by infinite values hold:

\[
\begin{align*}
+a/(+\infty) &= +0 \\
-a/(+\infty) &= -0 \\
+a/(-\infty) &= -0 \\
-a/(-\infty) &= +0
\end{align*}
\]

Other operations involving infinite values behave the way one would expect. Again for positive \( a \), the behavior is:

\[
\begin{align*}
\infty + \infty &= +\infty \\
\infty - \infty &= \text{NaN} \\
\infty \times \infty &= \infty \\
\infty/\infty &= \text{NaN} \\
\infty/a &= \infty \\
\infty/0 &= \infty \\
0/0 &= \text{NaN}
\end{align*}
\]

The rules in a Boolean expression involving infinite values are as expected:

1. All finite valid numbers are less than \(+\infty\).
2. All finite valid numbers are greater than \(-\infty\).
3. \(-\infty\) is less than \(+\infty\).

The rules involving expressions that have NaN values are simple:

1. Any arithmetic expression that includes NaN results in NaN.
2. Any Boolean expression involving NaN is false.

Perhaps the most useful aspect of IEEE floating-point is how divide-by-zero is handled; for any positive real number \( a \), the following rules involving division by zero values hold:

\[
\begin{align*}
+a/0 &= +\infty \\
-a/0 &= -\infty
\end{align*}
\]

IEEE floating-point has two representations for zero, one that is treated as positive and one that is treated as negative. The distinction between \(-0\) and \(+0\) only occasionally matters, but it is worth keeping in mind for those occasions when it does.
There are many numeric computations that become much simpler if the programmer takes advantage of the IEEE rules. For example, consider the expression:

\[ a = \frac{1}{\frac{1}{b} + \frac{1}{c}}. \]

Such expressions arise with resistors and lenses. If divide-by-zero resulted in a program crash (as was true in many systems before IEEE floating-point), then two if statements would be required to check for small or zero values of \( b \) or \( c \). Instead, with IEEE floating-point, if \( b \) or \( c \) is zero, we will get a zero value for \( a \) as desired. Another common technique to avoid special checks is to take advantage of the Boolean properties of NaN. Consider the following code segment:

\[
\begin{align*}
a &= f(x) \\
\text{if } (a > 0) \text{ then} \\
& \quad \text{do something}
\end{align*}
\]

Here, the function \( f \) may return “ugly” values such as \( \infty \) or NaN, but the if condition is still well-defined: it is false for \( a = \text{NaN} \) or \( a = -\infty \) and true for \( a = +\infty \). With care in deciding which values are returned, often the if can make the right choice, with no special checks needed. This makes programs smaller, more robust, and more efficient.

1.6 Efficiency

There are no magic rules for making code more efficient. Efficiency is achieved through careful tradeoffs, and these tradeoffs are different for different architectures. However, for the foreseeable future, a good heuristic is that programmers should pay more attention to memory access patterns than to operation counts. This is the opposite of the best heuristic of two decades ago. This switch has occurred because the speed of memory has not kept pace with the speed of processors. Since that trend continues, the importance of limited and coherent memory access for optimization should only increase.

A reasonable approach to making code fast is to proceed in the following order, taking only those steps which are needed:

1. Write the code in the most straightforward way possible. Compute intermediate results as needed on the fly rather than storing them.
2. Compile in optimized mode.
3. Use whatever profiling tools exist to find critical bottlenecks.