IOP Concise Physics | A Morgan & Claypool Publication

The Physics and Art of Photography

Geometry and the nature of light

John Beaver

VOLUME ONE

IOP ebooks

The Physics and Art of Photography, Volume 1

Geometry and the nature of light

The Physics and Art of Photography, Volume 1

Geometry and the nature of light

John Beaver
University of Wisconsin Fox Valley, Menasha, WI, USA

Morgan & Claypool Publishers

Multimedia content is available for this book from http://iopscience.iop.org/book/978-1-64327-332-7.

ISBN 978-1-64327-332-7 (ebook)
ISBN 978-1-64327-329-7 (print)
ISBN 978-1-64327-330-3 (mobi)

DOI 10.1088/2053-2571/aae1b6

Version: 20181101

IOP Concise Physics
ISSN 2053-2571 (online)
ISSN 2054-7307 (print)

A Morgan & Claypool publication as part of IOP Concise Physics
Published by Morgan & Claypool Publishers, 1210 Fifth Avenue, Suite 250, San Rafael, CA, 94901, USA

IOP Publishing, Temple Circus, Temple Way, Bristol BS1 6HG, UK

For Valeria

Contents

Preface

Early drafts of this book were written for a course I first taught in the Fall of 2013 at the University of Wisconsin–Fox Valley, in Menasha, Wisconsin.

I assume no specific prior knowledge of the reader except for a very basic understanding of physical units, dimensions and scientific notation (these topics are reviewed in appendix D for readers unfamiliar with them). The mathematics presented in the text is rudimentary, with only the most basic of algebra (more detailed derivations or those that require calculus are relegated to the appendices). If you have little experience with photography, it is my goal that *The Physics and Art of Photography* will help form a useful foundation from which to learn about photography in whatever way that works best for you. If you are a seasoned pro, but looking to set off in a new direction, then I still hope that you will find much here that is fresh and inspiring, and it is my goal that the book will help to open new possibilities. *The Physics and Art of Photography* is in three volumes:

Volume 1: Geometry and the nature of light
- Part I: Some preliminary ideas
- Part II: The nature of light
- Part III: Geometry and two-dimensional design

Volume 2: Energy and color
- Part I: Energy and photography
- Part II: The art and science of color

Volume 3: Detectors and the meaning of digital
- Part I: The physics of light detectors
- Part II: Photography as an art and the meaning of digital

The Physics and Art of Photography covers some material that is typical of discussions that link physics and photography. But it is also personal; it is very much my own take on the two subjects. I would not say that my personal views regarding science and art are controversial, but they are perhaps somewhat unconventional. There are few details here that other artists and scientists are likely to strongly disagree with. It is, rather, what I have chosen to emphasize, what I have left out all together, and the particular connections I point to, that most shows my own personal likes and dislikes.

Since my formal training is in physics and astronomy, while I am essentially self-trained in art (with informal mentoring from many others), the science part of this book is perhaps more conventional and straightforward than is my portrayal of art. And so my choice of physics-related topics should give one a fairly balanced and conventional taste of that subject as it relates to photography. Regarding photography as an *art*, however, I am surely on shakier ground.

Certainly, I do not pretend to present a comprehensive or balanced overview of art photography; I am unqualified to attempt such a thing. But I do try to make a case that the particular thin slice that I present here has some merit and is worth spending a little time to consider, even if it turns out not to be your particular cup of

tea. This book is a bad place to get a sense of what are the hot topics in *ArtForum*, but I believe that it does at least point to important and interesting questions about art photography in general. And since it is my goal to get you thinking, it doesn't matter much whether you agree with me or not. Thus it is fitting that my discussion of art is more personal, since my own art is the wee bit for which I really do know about what I am talking.

And so one might complain that *The Physics and Art of Photography* is a very long artist's statement, justifying the value and relevance of my own art. That may be partly true, but I do try to approach it in a way that emphasizes broad *questions*, rather than the particular answers I try to give (tentatively) with my own art. And I hope this book does help a little to make you a better photographer, and as such I do spend time on some of the very basic technical aspects of photography that I find important. But in doing so, I try to use these technical issues as points of departure to consider the status of photography as an art, finally exploring some issues relating to this status in the digital age.

This book may also be read as a manifesto of sorts for the aspects of science that have always moved me the most. I am interested in science not for the technological gizmos it has produced, or for some notion of inevitable human 'progress.' Rather, science is, for me, part of *the study of nature*. My interest in Einstein's General Relativity, for example, is essentially the same as my interest in bird watching. Because I have spent some time learning a bit about birds, I can now walk through the woods free of binoculars, looking only at the ground at my feet, and a world is open to me just by the sounds I hear. And when I stumble on my way up the stairs, as a physicist I can take comfort in the idea that my shin in contact with the stair prevented me from following my normal straight-line path through four-dimensional spacetime.

You will find throughout the book illustrations from my own photography as examples. This is convenient, since I know my own pictures and the stories behind them, and I don't need permission to use them. But of course I also want you to look at other photography, and so I have included some examples from a few other artists whose work I admire.

A useful companion is *The Photography Book* (Phaidon Press, 2014), which presents hundreds of photographs, spanning the entire history of photography. Each has a short analysis, with cross references to other photographs that are related. The photographs, only one per photographer, are arranged in alphabetical order by photographer's name. Thus, the ordering of the pictures is thematically random, which often results in unusual juxtapositions on facing pages. I sometimes refer to pictures in *The Photography Book* as examples, and so it is useful to have it handy. But all of these pictures are famous and can easily be found online as well.

The reader will also find, scattered throughout the three volumes and their appendixes, details and examples from what I call *ephemeral process (EP) photography*. EP photography is my own invention—sort of—and I spend so much time on it because it is perfect for illustrating many of the concepts in *The Physics and Art of Photography* in a way that I believe goes directly to the heart of the matter. Furthermore, it is *accessible*. The materials and equipment are

inexpensive, it requires no specialized facilities (such as a darkroom), and it is surprisingly versatile. But most importantly, it is a lot of fun.

The larger concerns of *The Physics and Art of Photography* are to give the reader some background that is helpful for asking important questions about the nature of art and science. But the practice of photography is the point of departure for these bigger issues, and as such *The Physics and Art of Photography* does contain a lot of simply practical information as well. And so *The Physics and Art of Photography* has five basic goals:

1. To ask basic questions about how photography fits in as an *art*, and about the nature of art itself.
2. To ask basic questions about the nature of physics *as part of the study of the natural world*, and about the nature of science itself.
3. To gain some practical knowledge that will allow the reader to more easily learn technical aspects of photography, as they are needed.
4. To gain some practical knowledge that will help the reader more easily learn to be a better photographer.
5. To expose the reader to a set of interesting photographic processes and tools that are not usually covered in a beginning photography course.

One of the themes of this book is the meaning of digital technology and what it has to say regarding photography as an art form. This may seem like I am speaking out of turn here, since I have neither formal training in art, nor have I ever been a professional photographer using professional digital equipment. Nevertheless, there is a sense in which I am well-positioned to say something of interest about these issues.

My own photography is almost entirely devoid of the use of a digital camera. I often use equipment and old physical processes that are about as far removed from modern digital photography as one could imagine. But I use these in new ways that depend absolutely on the digital; many of my photographs could not exist without modern digital processing and scanning and printing. This kind of interplay between the old and new is one of the running themes of *The Physics and Art of Photography*.

And despite my collection of old cameras, I am not a knee-jerk hater of digital imaging technology. In fact, I am one of its early practitioners, having used digital cameras and sophisticated digital image processing long before most photographers. My formal training is in astronomy, and I was there (in graduate school) for the digital revolution as it transformed astronomy in the 1980s. The CCD digital detectors used in modern digital cameras were fairly new then, and still too expensive (and with insufficient resolution) to be of much practical use for photographers. I am the last person one would want to ask about the latest multi-thousand-dollar model of DSLR camera. But I do have a decades-long understanding of some of the most basic underlying principles of digital photography.

Acknowledgements

I thank Valeria Sapiain for all-around support and patience over the several years I spent working on this book. I received much valuable feedback, support and mentoring from Laura Andrews, Doug Fowler, Diana Ludwig, Dawn Patel, Teresa Patrick, Judith Waller and Frank Zetzman.

The software packages GIMP, Gnuplot, Inkscape, SciDAVis and OpticalRayTracer were used for many of the illustrations. All photographs and illustrations are by the author, except as noted below:

Figure 2.2: Copyright 2018 Google (detailed attribution in image).

Figure 3.11, Left: By Geek3 - Own work, CC BY 4.0. https://commons.wikimedia.org/wiki/File:Mplwp_dispersion_curves.svg.

Figure 3.15: By Haade, Wjh31, Quibik –Own work, CC BY 3.0. https://commons.wikimedia.org/wiki/File:Interference_of_two_waves.svg.

Figure 3.17: By Nicoguaro - Own work, CC BY 4.0. https://commons.wikimedia.org/w/index.php?curid=49324857.

Figure 3.19: By Arne Nordmann (norro) - Own illustration, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1944668.

Figure 3.20 Left: Public Domain. Right: Public Domain.

Figure 3.22: By P.wormer - Own work, CC BY 3.0. https://commons.wikimedia.org/wiki/File:Electromagnetic_wave.png.

Figure 4.2: Public Domain.

Figure 4.3: CC BY 3.0. https://en.wikipedia.org/wiki/File:Spectrum_of_halophosphate_type_fluorescent_bulb_(f30t12_ww_rs).png.

Figure 6.1: By SpoonSpa, Simon Viktória, CC BY 2.0 Generic. https://commons.wikimedia.org/wiki/File:Szakkad.jpg.

Figure 7.9: Photograph by Teresa Patrick, used by permission.

Figure 10.4: Public Domain.

Figure 11.1: Left: Public Domain. Right: NASA, J. Bell (Cornell U.) and M. Wolff (SSI).

Figure 11.8: Public Domain.

Figure B.1 Left: Public Domain. Right: By Chris 73, CC BY 3.0. https://commons.wikimedia.org/wiki/File:NautilusCutawayLogarithmicSpiral.jpg.

I thank the University of Wisconsin—Fox Valley Department of Physics and Astronomy for use of their near-infrared and thermal-imaging cameras. A portion of this work was carried out with the support of the University of Wisconsin Colleges sabbatical program, to which I extend my heartfelt thanks.

Author biography

John Beaver

For nearly 20 years, John Beaver has used old processes to make new negatives, often in ways that can only be realized as a print with digital scanning and printing. This includes his development of the cyanonegative process, innovative work (in collaboration with Teresa Patrick) with instant film, and most recently his development of an accelerated, unfixed printing-out process he calls (perhaps annoyingly) 'Ephemeral-Process photography.'

He is Professor of Physics and Astronomy at the University of Wisconsin – Fox Valley, where he teaches physics, astronomy, photography and interdisciplinary courses. He earned his BS in physics and astronomy in 1985 from Youngstown State University, and his PhD in astronomy in 1992 from Ohio State University. His published work in astronomy is on the topics of spectrophotometry of comets and gaseous nebulae, and multi-color photometry of star clusters.

He has exhibited photographs in many juried competitions in Wisconsin, Ohio, New York, Louisiana, Missouri, Oregon and Colorado, even occasionally winning an award or two (well, two actually). He has had several solo exhibitions, as well as joint shows with artists Judith Waller, Diana Ludwig, Dawn Patel and Teresa Patrick. Beaver has long been involved in art-science collaborations (many with artist Judith Baker Waller) in the classroom, at academic conferences, and in art galleries and planetaria.

Some of John Beaver's photography can be seen at http://www.JohnEBphotography.com.

Part I

Some preliminary ideas

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Chapter 1

What is science; what is art?

One purpose of this book is to use photography as a point of departure to learn a bit about both science and art, and how they relate to each other. And so it may be helpful to take a brief tour of the nature of each in turn, before really diving into the material. The difficulty is that there is much more to both art and science than what can be captured in any kind of short description. And so I will not really answer the questions 'what is science?' or 'what is art?' in any comprehensive way or with nice one-sentence definitions of these two odd and complex things that humans do.

I will say only a little about the nature of science in general, and then almost nothing about the overall nature of art. This is mostly because I am much more qualified to attempt such a 'big picture' description for science. I have been teaching science for a few decades now, mostly at an introductory level to non-scientists. And so I have had to think long and hard about just what it is that I am teaching and why I am doing it. My relationship with art on the other hand is less formal, more personal and more recent. Please note that in what follows I am talking about the *natural sciences*. The social sciences have much in common, but also much that is different, and I leave it to others to talk about them.

Much has been written about the nature of science, and it is a subject of ongoing debate. My brief and very-incomplete discussion follows the general approach of Sokal and Bricmont (1998, chapter 4) but also informed by Chomsky (2000, chapters 4 and 5); it should be taken as only one approach to this complex topic.

1.1 The coherence of our experience

One key aspect of science is that it favors explanations that economically tie together, *accurately*, a broad range of experience that otherwise would seem to be connected only by coincidence. What may seem on the surface to be a coincidental connection is instead evidence that there is a deeper understanding to be sought, in which all of the facts come together in a natural way.

doi:10.1088/2053-2571/aae1b6ch1

When a scientific explanation works well at connecting the seemingly unconnected, we are left to consider that either there is something to it, or the Universe is conspiring, as a sort of joke, to make it only *seem* as though the explanation is correct. It is tempting to give a simple, short example right here; the problem is that for any real example it is always a very long story.

But this is what happens when we find a compelling explanation in science: *many seemingly-disconnected facts all fit exactly into one simple scheme, called a theory. And this theory allows one to correctly predict or explain, again and again, new facts that one hasn't yet even looked for or considered.* And so we are then left with three choices:

1. It is simply a coincidence that our theory gets it so right, so often.
2. There is a conspiracy by some cosmic intelligence, either malevolent or at least with an odd sense of humor, to arrange things intentionally so that it only *looks* like our theory is correct, when actually the theory is completely wrong.
3. There is some kind of truth to our theory. It is at least on the right track, at the moment, although details will likely need to be modified as we learn more. And perhaps some day we will see it, not as something that stands alone, but rather as only a part of a larger, more complete theory.

Science chooses the third option whenever it gets to the point that the first option seems too unlikely to take seriously. In science we basically reject the second option out of hand, with neither argument nor apology. As a human being, one may very well accept option 2 in a particular case; science is only one of the many things we humans do. But whatever that is, it isn't science.

1.2 Truth in science

We never prove a theory true in science. When a scientist says (quoted in a news story, for example) that a particular theory has been 'proven,' they are being somewhat glib, and they do not mean it in the strictly mathematical sense.

Rather, they mean that the theory is *compelling*. That is, it is rational to believe the theory has at least a conditional and tentative truth. And more importantly, the evidence is great enough that it is, in a sense, *ir*rational to believe the theory is completely wrong.

Basically, to say that a scientific theory has been proven is to say that of the options in section 1.1 above, option 1 is, in the face of new evidence, too unlikely to take seriously. And so we are faced with either option 2 or option 3, and if we are acting as a scientist then we dismiss option 2 out of hand, and so choose option 3.

Now of course in real life it is never so clear-cut as I have laid out here. For it is always the case that evidence is limited and uncertain. This means that for any given theory that gets all of the available evidence correct—that 'gets it right' so to speak—*there will also be an infinite number of other theories that do the same thing*. So why then, in a given case, do we pick the particular theory we do?

Well, often, we don't. We disagree, at least for a while. But we do pretty much agree, as scientists, that theories that explain more, but still manage to do it with *less* complexity, are the theories that are more likely to be correct. This is sometimes called *Occam's razor*, and it serves as a rule of thumb for choosing between competing theories with equal explanatory power. But even given Occam's razor, it is not always clear, in a given case, which theory really is the least complex; there are always different ways to look at it. And so we often disagree about the details, in the short term, only approaching overall agreement in the long term as more is learned.

Clearly, this is not really a proof at all; any particular scientific theory, no matter how compelling, is never the *only* explanation possible for all of the evidence. And since there are other logically possible explanations, then the theory has not been proven in the mathematical sense. Rather, the scientist makes a much less-strong claim about their theory—simply that it seems *un*reasonable to *dis*believe it.

1.2.1 Proving a theory false

We can't really prove a scientific theory to be true, but it might seem that it should be easy to prove that an untrue theory is false. In theory (pun intended), a theory is *falsified* if it predicts something in particular to be true, and then we go out and check and find that the prediction was incorrect. So it would seem that it should be easy to prove a theory incorrect; simply find *one* fact that it gets wrong.

But surprisingly, even to *prove* a theory false is, in practice, rarely a straightforward process. For maybe the observed 'fact' is wrong, not the theory. Furthermore, real theories are usually subtle. That is, they use simple principles that when applied to particular circumstances predict extremely complex results. This means that, since the world is a complicated place, applying any given theory to a particular real-world situation can be surprisingly difficult. In fact, it is often too difficult to do exactly.

Newton's laws of gravity and motion are a good example. These laws are simple enough that, when expressed in the correct mathematical language, they can be written on the side of an envelope (see figure 1.1). Yet it can be maddeningly difficult

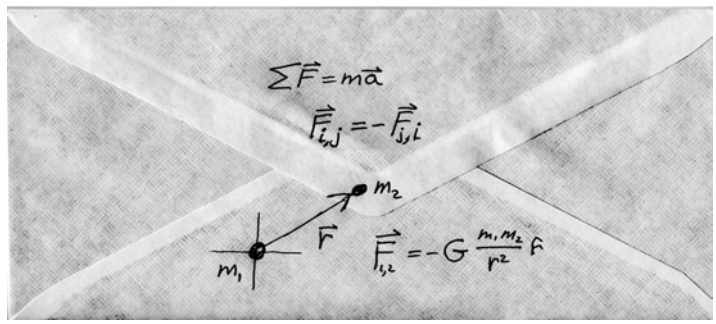

Figure 1.1. Newton's laws of motion and gravity are simple in a mathematical sense. When expressed in the right mathematical language, they can be written on an envelope. But applying them to real-world situations can be maddeningly complex.

to figure out exactly what they predict in even seemingly-simple real-world situations. It is so difficult that we are usually forced to make (hopefully) reasonable approximations. This leaves open the possibility, when a theory seems to give the wrong answer in a particular case, that it is not the theory that is incorrect. Rather, it may be that we have made invalid approximations when we *applied* the theory.

And so in the 19th century it was noticed that the application of Newton's laws to the orbit of the planet Mercury gives answers that are just slightly different from the planet's actual measured orbit. Faced with this, should 19th-century scientists have immediately rejected Newton's laws and looked for something better? Certainly not, for there were many imaginable explanations for the discrepancies in Mercury's orbit that did not conflict with Newton's laws (several were suggested at the time). And so, since Newton's laws worked so enormously well in so many zillions of cases, it would have been unreasonable to reject it because of this one discordant fact, when other explanations could easily be imagined.

I have chosen this example for its irony, because it turns out in retrospect that the discrepancy in Mercury's orbit really *is* caused by a failure of Newton's laws of gravity and motion. Einstein in 1915 superseded these laws with a more elaborate and complete theory called *General Relativity* (GR). One of the first things he demonstrated was that GR gives the same answer as Newton's laws whenever Newton's laws agree with experiment. But GR also gives the right answer for Mercury's orbit, whereas Newton's gravity gives a slightly incorrect answer.

A very tiny discrepancy in the orbit of one planet seems like a poor reason to discard Newton's laws, a wildly successful theory for 250 years, and replacing it with something much more complex and strange. But this was not Einstein's primary motivation. Furthermore, this successful prediction (*post*diction really) was only one of the many reasons (and not the most important one) why most other physicists fairly quickly adopted GR as the new best theory of gravity (Weinberg, 1992, chapter 5).

In short, Einstein provided no 'proof' that GR is correct. Instead, he argued that as a scientific theory it is *compelling*. So, what then *was* his argument? It is a long story![1]

But back to the 'failure' of Newton's laws regarding the orbit of Mercury; this example is the exception. Every day, facts are discovered that, if they are taken literally, conflict with some tried-and-true theory. But upon closer examination, it is almost always the *'facts'* that are wrong, not the tried-and-true theory. So it is prudent to be conservative and require extraordinary evidence for extraordinary claims.

1.3 Operational definitions

One necessary tool for the physical sciences is the *operational definition*, which is not at all like a dictionary definition. The dictionary appeals to our existing intelligence

[1] There are many excellent books on this topic, many of which are accessible to the lay reader; Rucker (1977) is a good example.

and knowledge, and most of the words defined have meanings that are somewhat fuzzy and vague by nature.

In the physical sciences we must define our terms so they are useful not just for describing our existing knowledge, but also for extending it in ways that we can't predict ahead of time. And the definition must mean exactly the same thing to everyone.

Fortunately, this is not quite so difficult as it sounds. Instead of tying ourselves up in knots writing a whole book just to carefully define a single word, we instead describe a *procedure*, a set of steps to follow. Basically, it is a recipe. 'Do this, now do that, attach this thing here, read the meter, plug that number into this equation. And that thing that comes out in the end—*that* is what I'm talking about.' And so a given physical quantity is defined by the very method that is used to measure it.

A good example is the quantity *mass*, which is defined by what is, in essence, instructions for performing a particular experiment. When this experiment is performed, the concept of mass naturally arises, and is thus named in context. Interestingly, it has long been known that there are actually *two* very different operational definitions for mass—one in terms of inertia and the other in terms of gravitation. Experimental evidence over 200 years showed that these two very-different definitions gave, essentially, the same answer. This was considered to be a strange and mysterious coincidence until Albert Einstein resolved the issue in 1915, in the context of developing his new theory of gravitation, GR.

The use of operational definitions means that terms in physics are defined much more precisely than are the words we use in our everyday speech. This means that, for the most part, if one particular word (velocity, for example) is used correctly in a given physical context, *then every other word used in the same context would be incorrect*. This is in contrast to the more vague and fuzzy meanings of the words we use in our everyday speech, where the meanings overlap somewhat; one word might be the 'best,' but often several other words are also non-wrong. This is, unfortunately, made more complicated by the fact that many words in physics are pronounced and spelled just like words we use in our everyday speech, and so it is tempting to wrongly apply those same vague meanings in the context of physics. Some common offenders are *force*, *energy* and *momentum*. These words have precise operational definitions in physics. In a given physical context, in a given sentence, only one is correct; the others are wrong.

1.4 Inspiration and perspiration

Einstein's famous phrase that '…genius is 99% perspiration and 1% inspiration' has much truth in it for both science and art. An artist needs many tools to draw from in order to act upon inspiration when it arises. It does no one any good if your genius is only in your own mind, and it can't get out.

A good example is *The Bird Cage*, which can be seen in figure 1.2. It received a little bit of attention, having been selected for a couple of national juried exhibitions, one in New York City. I certainly wouldn't say it is a work of 'genius,' whatever that means exactly. But I do believe one can say some good things about that particular picture.

Figure 1.2. *The Bird Cage*. John Beaver, 2004.

How did I do it? It is such a long story, that I am not sure even where to begin. This is a good thing, because if I did know where to start, I might proceed. And if I proceeded to recount all that went in to that picture, you would be a very bored reader indeed. It is not that there is nothing interesting to say about how that picture came about. But *most* of the story is mundane, the result of what can only be described as a long slog through a lot of work.

This example is particularly illustrative because, in order to make that picture, I had to develop a new photographic technique that I call *cyanonegative photography*, and much of the power of this particular photograph is intimately connected to the technique used to make it.

To take advantage of inspiration when it arises, one needs tools. For a musician, the seemingly endless scale exercises are part of the necessary preparation for creative acts such as improvisation and composition. A photographer must understand principles of two-dimensional design in order to get more than the lucky shot, and this takes practice. Without having studied many topics very hard, an astronomer is just a person stumbling around in the dark. So part of the purpose of this book is simply to give you some useful tools. Make of them what you will!

1.5 Criticism and self esteem

Criticism is an absolutely necessary part of both science and art. To succeed in either, one must have a thick skin. Perhaps I should put it differently; it may sound

as though I am saying that one must develop an ability to tolerate abuse, and that is not at all what I mean. Rather, one needs to be able to distance oneself, and look at one's own work as if it were seen from the outside.

This ability to look at one's own work with some distance is necessary in order to hear the criticism of others without being paralyzed by it. Don't take it personally! Notice the exclamation mark. And to emphasize this, I say it again, but also with italics. *Don't take it personally!* Instead, criticism can be seen as information—sometimes useful, sometimes not—rather than a statement about one's ultimate worth as a human being.

The role of criticism in art is well known, but it is every bit as important in science, and a successful scientist internalizes the positive dynamics of criticism early on. In fact one tries very hard to anticipate critique, and to answer it before it occurs. This may seem as though it is simply deference to the authority (perhaps legitimate, perhaps not) of others. But for a successful scientist, it is not. Rather, it is an essential part of keeping oneself honest, and also for simply getting things done. Without looking critically at one's own work, how else can one tell whether or not it is even finished? One must actively look for flaws to find and address them.

As I write this, I am faced with a dilemma[2]. I have been working for *three years* on a photometric analysis of a single star cluster in the constellation Scutum. In collaboration with a colleague from UW Oshkosh, I have been working on a photometric analysis of this cluster. Although the cluster is an important and famous one, no one had previously studied this cluster in the particular way we have.

The problem is that our data are not fully cooperating, and it is *my* part of the analysis that is problematic. Much of it is as one would expect, but some of it makes no sense. Meanwhile, we discover (literally today, as I write this) that *someone else* will soon present similar data on this very same cluster. What to do?

Well, we make the best of it. I work very hard to get the most coherent, most interesting, and the most honest results I can out of our data. And I do it as fast as possible. Perhaps all of that work will come to something, perhaps not[3]. But in any event, I have learned a lot by doing it, and I now have many new tools at my disposal. And so when inspiration arises at some later date, it is more likely that I will be able to make the most of it.

Self-criticism is particularly important for photography because it is so easy to quickly accumulate a *lot* of photographs. And so a big part of making a good photograph is the ability to tell the bad from the good—in order to show the good ones to others, and throw away (or archive) the bad ones. This takes practice, because it is easy to become attached to one's own photographs, even when they are awful.

Once, lying in bed awake at 3 a.m., I had a great idea for a series of photographs. I won't go into the details. Suffice to say that it was a very clever idea, but pulling it off required a significant amount of work. I finished the first couple of photographs,

[2] I wrote this sentence in November of 2012.

[3] By December, 2013 all was well. The research project finally came together and we published a paper (Beaver *et al* 2013).

of which I was very proud, and excitedly began the detailed planning for the rest of them. But events required me to leave town for a while, and so I had to temporarily set this project aside.

When I returned, I pulled out my notes and the two photographs I had already taken, and I really looked at them. The pictures were pretty bad. And the more I stepped back from myself and tried to see, through the eyes of others, the finished product I had planned, the more I realized that my whole idea was rather lame.

So it goes. Afterwards, was it difficult to part with a project with so much time invested? No, not really; I'll make other pictures instead. Was the time I spent wasted? Possibly, but I doubt it. Much of the experience of thoughtfully but wrongheadedly making those bad pictures will likely find practical use someday, some way. I'm just thankful that I stopped myself before I hurt someone by hanging those dreadful things on the wall[4].

1.6 Looking at art

How can one look at something and know whether or not it is art? In the novel *Bluebeard* (Vonnegut 1987, p 148), the protagonist gives the following response (attributed to the painter Syd Solomon) when asked how to tell a good painting from a bad one: 'All you have to do, my dear, …is look at a million paintings, and then you can never be mistaken.'

To know how to look at art, there is no substitute for looking at a lot of it. Go to museums as often as possible, and look at not only photography, but also painting and sculpture. As a simple, practical matter to get much more out of this book, *look at every single picture in its companion*, *The Photography Book* (Cooke and Kinneberg 2014), *read the descriptions and consider the cross-references*.

References

Beaver J, Kaltcheva N, Briley M and Piehl D 2013 Strömgren H-β photometry of the rich open cluster NGC 6705 (M 11) *Publ. Astron. Soc. Pac.* **125** 934

Chomsky N 2000 *New Horizons in the Study of Language and Mind* (Cambridge: Cambridge University Press)

Cooke T and Kinneberg C (ed) 2014 *The Photography Book* 2nd edn (London: Phaidon Press Limited)

Rucker R v B 1977 *Geometry, Relativity and the Fourth Dimension* (New York: Dover)

Sokal A and Bricmont J 1998 *Fashionable Nonsense: Postmodern Intellectuals' Abuse of Science* (New York: Picador)

Vonnegut K 1987 *Bluebeard* (New York: Dell Publishing)

Weinberg S 1992 *Dreams of a Final Theory* (New York: Pantheon Books)

[4] One of the pictures, on further reflection (and input from others) turned out not to be so terrible, and so I did eventually put it in an exhibition. It was better standing on its own and would have been much diminished in the context of the overblown, multi-photograph project I had originally envisioned.

Part II

The nature of light

IOP Concise Physics

Chapter 2

What light is

The properties of light underlie much of photography; it is, after all, the *photo* in the word itself. Light has several different aspects, and we will look at each of these in turn. In particular, light has a *wave-like* nature, a *particle-like* nature, and a *geometric* nature.

When a photographer speaks generally of light, they often mean a combination of several different things (see figure 2.1). One of these is the intensity of the light on the subject, in relation to how the camera will need to be set for a proper exposure. But they may also be speaking of a subtle combination of the intensity, angle and color

Figure 2.1. *Algae No.2* John Beaver, 2005. A photographer could mean many different things when talking about 'light' for a photograph such as this.

doi:10.1088/2053-2571/aae1b6ch2

of the light on the subject. For example, the *magic hour* of one to two hours before sunset is famous for its 'light.' And this reputation comes from a complex combination of many aspects of light, only some of which are physical. For just as important are aesthetic concerns that have as much to do with the human mind as with physically definable aspects of light.

We will take up many of these issues throughout *The Physics and Art of Photography*, but in this chapter we concentrate on some strictly physical aspects of light that are independent of photographic concerns. We will see that a better understanding of these physical aspects of light will give us a stronger foundation for understanding some of the more subtle ways in which light relates to photography as an art.

There is no really good two-sentence definition of light, and so I will simply describe, as best I can, its properties. There *is* a unified, coherent physical description of light, and it is quite a long and complex story to go through in detail. But there are a few ways to approach the physical nature of light that, although simplified in various ways, still manage to capture the majority of its properties.

The biggest simplification is to describe light as a series of *rays* that travel in straight lines until altered by various devices such as lenses or mirrors. This is the subject of *geometrical optics*, and we take that up in chapter 6, as it has many practical applications for photography.

In this chapter we go a bit deeper, but even so there are two approaches, each of which works best for particular circumstances. One is to model light as a stream of *particles*, and the other is to view light as a *wave* phenomenon. It is the latter approach that turns out to be most useful for the purposes of photography, and so we will spend more effort on waves than particles.

So if we can describe light in these three seemingly different ways, then what is light really? Well, it is what it is, and its physical nature is a bit unlike anything in our direct intuitive experience. So in order to fully answer that question, we would have to go into a full description of its physical nature, and that means a study of *quantum physics* and much else that is beyond the scope of this book. But suffice to say that the full quantum description of light is a coherent theory that both incorporates and unifies (and thus explains) its wave-like, particle-like and ray-like natures.

So we will describe light separately as waves, particles and rays, and try to lay out under what circumstances each model is most appropriate. And although this approach is not entirely correct, and strictly speaking is not entirely coherent either, in practice it actually works most of the time.

2.1 The speed of light

Light carries energy from one place to another, and so by means of it events in one location can influence events at another location. How fast does this influence travel? Very fast; try to open the refrigerator door before the light comes on.

Many years ago, my friend Doug Fowler bought a new Chevrolet half-ton pickup truck. It was the 'Custom Deluxe' model, meaning it was so stripped down and featureless that it lacked even a radio. It came off the assembly line in Flint,

Michigan in 1976. Doug drove the car to college at the University of Montana, regularly driving it back and forth between the Rocky Mountains and his home town of McDonald, Ohio. Eventually, Doug moved back to Ohio, and that is where I met him, while a student in the physics and astronomy department at Youngstown State University.

I went to graduate school at Ohio State, and Doug went back out west, this time all the way to Bellingham, Washington. Eventually, he ended up back in Ohio again, although he drove the truck westward many times, including visiting me in Flagstaff, Arizona one summer while I was working at Lowell Observatory.

Finally, around 1991 or so, I ended up with his by then much-used truck. When I moved to Wichita, Kansas, in 1993, all of my possessions went west in that truck. It was already on its *second* engine, and even that one was unhappy about the trip. I had to change the oil twice on the way because the piston rings were so shot. About a month or two after I arrived in Wichita, it went to its final resting place in some salvage yard in Sedgwick County, Kansas.

What did the odometer read after all of that? Very nearly 186 000 miles, the *distance light travels in one second*:

> *The distance light travels in one second is comparable to the total distance traveled by an automobile that has been through a dozen or so years of moderately-heavy driving.*

This speed, the speed of light in a vacuum, is such a fundamental quantity, that it has its own special symbol, c. It has a value of $c = 2.998 \times 10^8\ \mathrm{m\ s^{-1}}$.[1] This is roughly 300 000 km $\mathrm{s^{-1}}$, or 186 000 miles $\mathrm{s^{-1}}$.

A speed is a distance per time, and with such a large speed as this, one can get a mental grip on either the distance or time, but not both. And so if we talk about an intuitive interval of time—the second, the distance light travels is outside our direct intuitive experience. 300 000 km is nearly seven times around the circumference of Earth.

We can make the light-travel distance manageable—30 cm for example, just slightly under one foot. And so we could express the speed of light as a little under one foot in a billionth of a second (1 ns), or 30 cm $\mathrm{ns^{-1}}$. But now it is the *time* that is far too *small* for our direct experience. In section 2.1.1 we try to meet both the huge distance and the small time somewhere in the middle.

2.1.1 The speed of light with a shortwave radio

I moved from Wichita in the summer of 1997, again hauling (in a different vehicle, obviously) my belongings halfway across the country—this time eastward, to Appleton, Wisconsin. Because I was starting a new job teaching physics and astronomy at a two-year campus of the University of Wisconsin, I spent much of that summer preparing for my upcoming courses.

[1] A brief overview of physical dimensions, units and scientific notation can be found in appendix D.1.

But one of the first things I did upon moving in to an apartment was to set the clocks. In my family[2], this had always been accomplished by using a shortwave radio to tune in WWV—a transmitter broadcasting from Fort Collins, CO, and operated by the National Institute of Standards and Technology (NIST). The NIST uses WWV to broadcast standard time signals on several shortwave frequencies.

And so on the evening of June 30, 1997 I tuned my radio to 10 MHz, the WWV frequency I most expected to deliver a good signal. Within less than 45 s, I heard something I had never heard before on WWV—a *female* voice. For WWV announces each new minute with a male voice. As soon as the female voice announced the minute, the familiar male voice repeated the same announcement.

I knew what was up. The conditions were just right, for the first time in my life, to receive not only WWV from Colorado, but also WWVH, its sister station in Hawaii. For just this situation, the two stations use different voices for their every 60 s announcements, and one finishes before the other begins.

I immediately recorded the broadcast, and my quick thinking has paid off; for two decades now, I have used this recording to torture physics and astronomy students. As the first lab exercise in many of my beginning physics and astronomy courses, the students take a digital sample of my tape recording, and use graphical analysis to measure the time of arrival of the signals (Beaver 2000).

Figure 2.2 shows the approximate paths traveled by the two radio waves as they arrived at my receiver in Wisconsin. Shortwave radio waves have the ability to bounce back and forth between the ground and layers of ionized gases (called the *ionosphere*) in Earth's upper atmosphere. And so they can, in effect, bounce around the curvature of the Earth. This ziggzaggy route means the radio waves really

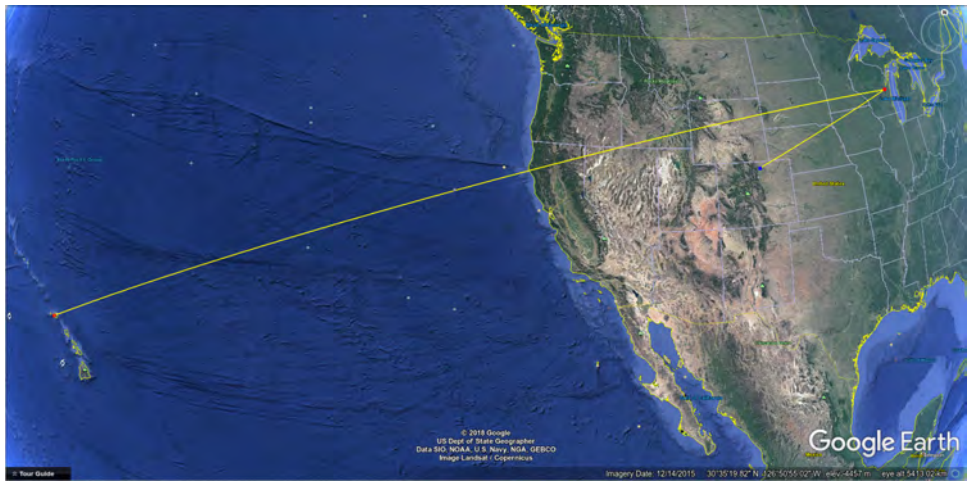

Figure 2.2. The paths of radio waves broadcasting simultaneous time signals from both Hawaii and Colorado, and arriving in Wisconsin, as mapped by Google Earth. Since the signal from Hawaii had further to travel, it arrived slightly later. Recording available at https://doi.org/10.1088/978-1-64327-332-7.

[2] My family included many ham radio operators.

traveled a little bit further than what is shown here; but the difference is less than 10%.

What does this have to do with the speed of light? First, light is an electromagnetic wave, and both visible light and radio waves are but two example of the many kinds of these waves. And so to measure the speed of radio waves is to measure the speed of light, because from a physical standpoint, radio waves *are* a kind of light, even though we can't see them with our eyes.

And so we have the following. Two signals were sent *simultaneously* from two different locations. Both signals traveled at the same speed—the speed of light. Both arrived at my receiver in Appleton, but one had to travel a greater distance to get there. Thus, the signal from Hawaii arrived later than the signal from Colorado. If one listens carefully to my recording, it is just possible to hear that each of the clicks is doubled, as if one tapped two fingers on the table, but with one finger slightly behind the other.

See figure 2.3 for a graph of two seconds of the data. Just a quick glance at the time axis on the graph (my students do this more precisely) shows that there is a difference of about 0.02 s between the arrival of the signals. To calculate the speed then, all one needs to do is divide the excess distance the Hawaii signal traveled by the excess time it took to arrive.

One can easily use an online tool such as Google Earth to see that the ground-level distance for the WWVH signal is about 6.9×10^6 m while for the WWV signal it is roughly 1.4×10^6 m. And so if we divide the difference in those travel distances by the excess 0.02 s for the arrival of the WWVH signal, we have:

$$\frac{6.9 \times 10^6 \text{ m} - 1.4 \times 10^6 \text{ m}}{0.02 \text{ s}} = 2.8 \times 10^8 \text{ ms}^{-1} \tag{2.1}$$

This is a little less than the speed of light, but remember that the radio waves really bounced between the ground and the ionosphere, and so the distances used in equation (2.1) are a little too small. Taking that small difference into account gives the correct answer (Beaver 2000).

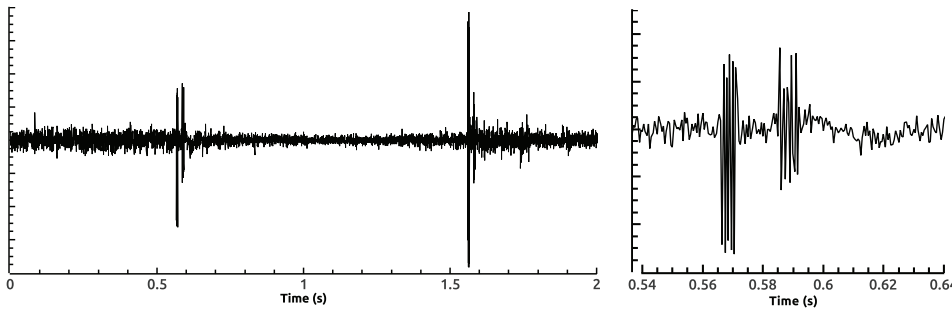

Figure 2.3. A graph of sound level versus time for 2 s of the signals received from WWV and WWVH. **Left:** The signals arrive as 'clicks' once every second. **Right:** a magnified detail of the time of arrival of the first click shows that it is actually two closely-spaced clicks. The first signal arrived from Colorado, and about 0.02 s later, the signal arrived from Hawaii.

2.1.2 Relativity and the speed of light

The speed of light is unlike other speeds; *all other speeds are relative*. The speedometer in a car measures the speed of the car relative to the road and, incidentally, to the tree alongside the road. The speed relative to another car following behind is quite different. And this difference is not just philosophical; it has real physical consequences, as would be immediately evident if your car came into contact with one or the other.

The speed of light, on the other hand, is absolute. *It is the same for all observers regardless of their relative motion*. This may seem impossible, but it is not, and the fact is well established by experiment. Light always travels (in a vacuum) at the same speed of 300 000 km s^{-1}, regardless of any relative motion between the source or observer. If I shine a flashlight at you while running towards you at 200 000 km s^{-1}, the light goes away from me at 300 000 km s^{-1} and it moves toward you at that same 300 000 km s^{-1}, even though the source of the light is moving toward you at 200 000 km s^{-1}.

This basic fact about light was first hinted at by Maxwell's complete theory of electricity and magnetism, formalized in the early 1860s. But it took four decades until sense was made of that idea; it is the foundation for Einstein's Special Relativity, first published in 1905.

When light interacts with matter—a piece of glass in a camera lens is a good example—then it may travel at a significantly slower speed. This has important consequences for photography; it is the reason we can bend light with lenses.

2.2 Geometry

Whether light behaves as a wave or as a stream of particles, it travels in straight lines until it interacts with matter in some way. Thus some of the most important aspects of light are more about the geometry of straight line *rays* of light, deflected by various objects.

And so we can often analyze light with geometry alone, only using particular details from our knowledge of electromagnetic waves when it is absolutely necessary. This approach is called *geometrical optics*, and we will use it often. And so when we talk about a ray of light, we are implicitly using this geometrical approach because it is convenient, even though we know light is really much more complicated than this. When we need to use the more-correct wave or particle models of light, we will. But otherwise, why make things unnecessarily complicated?

2.3 Waves

Wave phenomena allow the transfer of energy from one place to another without actual stuff having to make the trip (see figure 2.4). Move your hand up and down in the water at one end of the bathtub, and eventually the rubber duck at the other end bobs up and down too. Yell, 'Hurry up!' to your room mate in the bathroom and very quickly their eardrums vibrate, even though no air from your mouth traveled to their ear.

Figure 2.4. *Shoreline and White Caps.* John Beaver, 2006.

Light has this nature too, although it is not 'stuff' vibrating like most other waves. Rather, *a light wave alters the very electrical and magnetic properties of space itself*, even in a vacuum where there is nothing of substance to move. We call theses electric and magnetic properties *fields*, and so an *electromagnetic wave* is a changing pattern of the electric and magnetic fields. These changing fields can, in turn, affect matter, and that is essentially what happens when light has some physical influence. Since matter has its own electric and magnetic properties, it is affected by the changing electric and magnetic fields of a passing electromagnetic wave.

One might point out that if no stuff actually makes the trip when a wave moves from one place to another, what do we really mean then by the 'speed' of a wave? If the wave is very complex, then this question may have a complex answer. But for a simple wave, there is a simple answer—*a wave is a repeating pattern in space that moves as time passes. The speed of the wave, then, is the speed at which this pattern moves*.

This calls to attention the fact that a wave by its very nature is extended in both space and time. It is a pattern spread out in space that moves with time. And even at a given location, the wave changes as time passes. And so there is no real meaning to assigning a precise location and time to a wave. We can talk about what the wave does at a particular place and time, but we need to consider all of the other places and times in order to describe any particular wave. In this way, it is very unlike, say, a stone moving through space, which has a much-more precisely definable position at any specific time.

Like all waves, light in its purest form has four basic attributes: speed, amplitude, wavelength and frequency. We will consider each of these in turn.

2.3.1 Amplitude

As a light wave passes a point in space, the electric and magnetic fields change back and forth. The fields themselves actually point perpendicular to the direction of motion of the wave, and as the wave passes they alternately grow in strength, reach a maximum value, weaken to zero, then reverse direction and do the same. The maximum strength of the electric or magnetic field as the wave passes is called the *amplitude* of the wave.

For light, the amplitude is related to the brightness of the light. All else being equal, the larger the amplitude—the greater the maximum strength of the electric or magnetic field as the wave passes—the brighter the light. But if one doubles the amplitude, the brightness of the light does not also double; instead it quadruples. And so *the brightness of the light scales with the amplitude squared.*

This scaling by squaring has another consequence. The electric and magnetic fields alternately switch back and forth in direction, and we can describe this with positive and negative numbers. But the brightness of the light is proportional to the *square* of this, and the square of a number is always positive, even if the number being squared is negative. Thus, although the fields reverse directions, it is only the *magnitude* of the strength of the fields that represents the amplitude of the wave. This may seem to be a distinction without a difference, but it has profound and surprising consequences whenever two or more waves interact with each other, in a phenomenon called *interference*. We will revisit this idea in section 3.6.

2.3.2 Speed, wavelength and frequency

Let us consider a wave traveling down a string, in the way that a wave travels down a garden hose when one end is waved up and down. For now, let us imagine the string to be infinitely long, so we can ignore the interesting complication of waves reflecting off the ends of the string. If one were to take a flash picture of the wave, freezing it in time, a repeating pattern in space would be evident. One can represent this pattern in the form of a graph, as in figure 2.5.

The meaning of the amplitude of the wave is clear from the diagram. But there is another equally-important measure. The *distance* over which the wave completes one repetition is called the *wavelength* of the wave. It is a length, so our SI unit (see appendix D.1) of wavelength is the meter (m). By historical convention, we use the Greek letter λ (lambda) as our symbol for wavelength. The human eye is only sensitive to a very narrow range of wavelengths, between about 0.4 to 0.7 millionths of a meter, and so it is this range of wavelengths that defines what we call *visible light*.

These tiny lengths make the meter a bit cumbersome, and other smaller units are more commonly used when referring to the wavelength of light. The most common of these are described in table 2.1. Thus we can represent the range of visible wavelengths as '0.4 μm to 0.7 μm.' Or if we prefer, we could say the same thing as

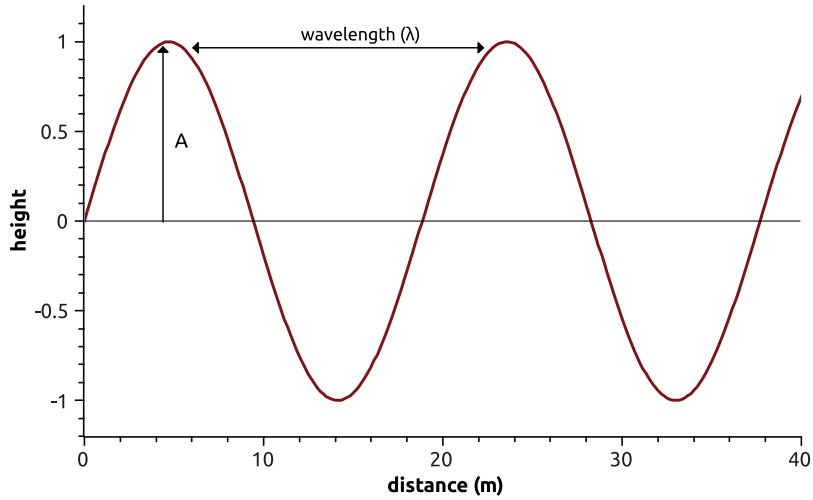

Figure 2.5. Wavelength (λ) and amplitude (A) of a wave. Notice that the horizontal axis of the graph is length. The wavelength is the *distance* over which the wave repeats itself, *at a particular point in time*.

Table 2.1. Useful small units for wavelength.

Unit	Symbol	Meters	Notes
Micron or micrometer	μ or μm	$\times 10^{-6}$	1000 μm = 1 millimeter
Nanometer	nm	$\times 10^{-9}$	1000nm = 1 μm
Angstrom	Å	$\times 10^{-10}$	10 Å = 1 nm

'400 nm to 700 nm.' As an alternative, if we can find the special symbol in our word processor, we could say with complete equivalence, '4000 Å to 7000 Å.'

We can also represent a wave in terms of changes in time, rather than in space. Figure 2.6 looks, at first glance, the same as figure 2.5, but look at the horizontal scale. Instead of showing the wave at different points in space (at a given instant of time), figure 2.6 shows the wave as time passes (but at some particular point in space). And so the *period*, T, of the wave is the *time* required for one repetition of the wave to pass a particular point in space, as the wave goes by. In SI units, the period is measured in seconds; for most electromagnetic waves of concern to photography, it is a tiny fraction of a second. More commonly we refer to the *frequency*, f, the reciprocal of the period:

$$f = \frac{1}{T} \tag{2.2}$$

The frequency then would be measured in inverse seconds, labeled s^{-1}, or Hertz (Hz). It is the number of wavelengths that go by per second.

While the period for electromagnetic waves is typically a very small number, the frequency is thus typically a very large number. There are two reasons for this. First,

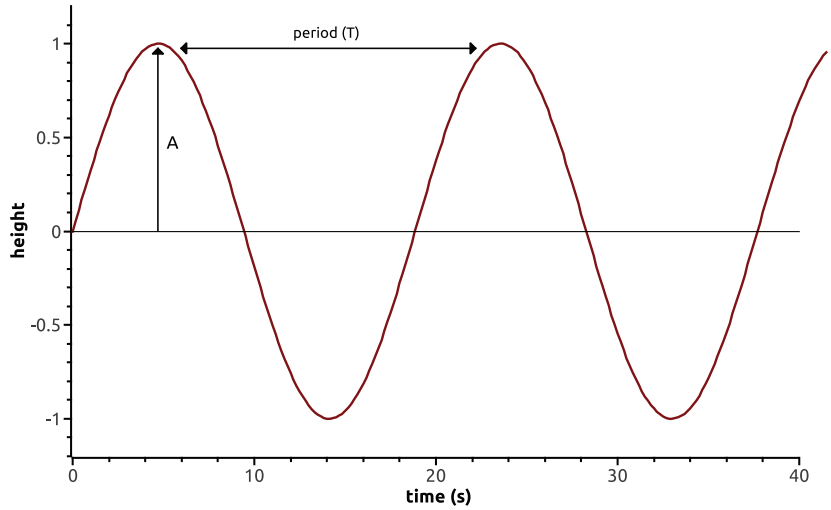

Figure 2.6. Period (T) and amplitude (A) of a wave. Notice that the horizontal axis is time. The period is the *time* over which the wave repeats itself *at a particular point in space*. The *frequency* (f) of the wave is simply one over the period.

the wavelengths of visible light are very tiny. Secondly, the speed of light is very large.

Clearly, the faster is the speed of the wave, the more wavelengths would go by in one second. But also, the shorter the wavelength, the more that would go by per second, for a given speed. Thus we have a relation between the frequency, f, the wavelength, λ, and the speed of light, c:

$$c = f\lambda \tag{2.3}$$

Since the speed of light, c, is a constant, then we can see that there is a relation between wavelength and frequency. Thus any given wavelength corresponds to a particular frequency, and vice versa. We can rearrange equation (2.3) as follows:

$$f = \frac{c}{\lambda} \tag{2.4}$$

$$\lambda = \frac{c}{f} \tag{2.5}$$

These are *reciprocal relations*; if frequency is larger, then wavelength is smaller, and vice versa. It also means that, for light, we can choose either frequency or wavelength for our description. If given one, the other can be easily calculated.

2.3.3 The electromagnetic spectrum

Different wavelengths (or frequencies) of electromagnetic waves interact with matter in different ways. Since both the absorption and emission of light are examples of such interactions, one would need different strategies to *produce* light of vastly

different wavelengths. Likewise, different methods are required to *detect* light of very different wavelengths.

It was not known that light is an electromagnetic wave until the late 1800s. And it wasn't until the 20th century that most forms of electromagnetic waves were finally identified. Some types had previously been detected, but it wasn't recognized until much later that they were just different wavelengths of electromagnetic waves. And so different ranges of wavelength of electromagnetic waves have different names, in part for historical reasons.

Table 2.2 shows the ranges of possible wavelengths, along with their customary names. Taken together, this is called the *electromagnetic spectrum*. Keep in mind that the ranges of wavelengths or frequencies are only approximate; the boundaries are fuzzy and overlap each other. The names really come from the different ways in which we produce or detect them, and that has changed over the years as technology has changed.

And so let us very briefly consider each of these basic parts in turn. I will start at the long-wavelength bottom of the list; this may seem strange, but remember that *long* wavelength is the same as *low* frequency.

- **Radio waves** are made by moving an electrical current back and forth in a wire, and radio waves induce currents to oscillate back and forth in wires they pass through.
- **Microwaves** can be thought of as very high-frequency radio waves. They can sometimes be made in the same fashion, but other processes (besides electronic circuits) are also used to produce them. Because of their shorter wavelength, they can be focused with special mirrors and more easily guided along paths.
- **Infrared light** is usually produced in ways similar to visible light, but the wavelengths are too long for the human eye to detect it.
- **Visible light** is the name for the narrow range of wavelengths (about 400–700 nm) to which they human eye is sensitive. Different wavelengths of visible light produce different color sensations; violet for short wavelengths and red for long wavelengths, with blue, green, yellow and orange in between.
- **Ultraviolet light** also is often produced in ways that are similar to visible light, but the wavelengths are too short to be detected by our eyes.

Table 2.2. The electromagnetic spectrum.

Name	Typical λ (m)	Typical size	f (Hz)
Gamma ray	$< 1 \times 10^{-11}$	Atomic nucleus	$> 3 \times 10^{19}$
x-ray	$1 \times 10^{-11} - 3 \times 10^{-8}$	Atom	$1 \times 10^{16} - 3 \times 10^{19}$
Ultraviolet	$1 \times 10^{-8} - 4 \times 10^{-7}$	Virus	$7.5 \times 10^{14} - 3 \times 10^{16}$
Visible light	$4 \times 10^{-7} - 7 \times 10^{-7}$	Bacteria	$4.3 \times 10^{14} - 7.5 \times 10^{14}$
Infrared	$7 \times 10^{-7} - 1 \times 10^{-3}$	Protozoa	$3 \times 10^{11} - 4.3 \times 10^{14}$
Microwaves	$1 \times 10^{-4} - 0.1$	Person	$3 \times 10^{9} - 3 \times 10^{12}$
Radio	> 0.1	Building	$< 3 \times 10^{9}$

- **x-rays** have wavelengths similar to the size of individual atoms, and so they usually interact with matter in ways that involve individual atoms on a one-on-one basis.
- **Gamma rays**, with their enormous penetrating power, interact directly with the tiny nuclei of atoms. And thus they are associated most strongly with nuclear reactions.

We will discuss different examples in more detail as we go along. But for most of *The Physics and Art of Photography* we will consider the specific properties of *visible* light, and wavelengths of infrared and ultraviolet that are close to visible light, since these wavelengths are of most practical interest for photography.

2.4 Particles

A wave is spread out in space, and it transfers more energy as more time passes; this fact will be very important when we consider the physics of exposure in photography. But there are situations where light doesn't act like this at all. Instead it acts like a stream of particles (called photons), each of which individually delivers its energy all at once and at a particular place. It was the development of quantum physics, in the early 20th century, that finally unified these two seemingly-opposed wave-like and particle-like natures of light.

A particle affects other particles by way of *collisions*, which cause sudden changes at particular places. A wave, on the other hand, gradually causes things to happen over spread-out regions of space. And so the two ideas—waves and particles—seem on the surface to be utterly contradictory. It is beyond the scope of this book to go into detail as to how quantum mechanics unifies these two seemingly incompatible concepts, but it does.

Suffice to say that sometimes light behaves more like a wave and sometimes more like a particle. The particle-like picture for light is most important regarding how light is actually detected by a light-sensitive material: film, say, or the CCD digital detector in a digital camera. By 'detection' we mean that light causes some physical change in the detector that can be recorded in some way, thus eventually producing an image for us. If one looks carefully enough, this physical detection occurs at particular instants of time, and at particular locations on the detector. It is in this way, then, that light behaves like a stream of particles.

This is not to say that when light behaves like a stream of particles that our discussion of light as a wave will be invalid. For even when light behaves like particles, its wave-like nature is still important. In fact there is always a fundamental connection between the two ways of looking at light.

We describe light as a wave in terms of its amplitude and wavelength (or frequency), whereas a particle would be represented by the energy and momentum it could transfer in a collision. One of the central results of quantum physics is that in a given situation, the wave-like and particle-like natures of light are explicitly connected to each other. If we denote the energy and momentum of a single photon

by, respectively, E and p, they are always related to the wavelength, λ and frequency, f, of the light by the following:

$$E = hf \tag{2.6}$$

$$p = \frac{h}{\lambda} \tag{2.7}$$

In these equations the letter h represents a very tiny number, one of the fundamental constants of nature, known as the *Planck constant*. The overall consequence is that light of short wavelength or high frequency (when it is acting like a wave) is made of individual photons (when it is acting like a stream of particles) each of which has a high energy and large momentum.

Thus one can take a source of light and allow it to interact with matter in a wave-like way. From that interaction, one can measure the wavelength. Then take the same source of light and allow it to interact in a particle-like way, and in the process measure the momentum of those particles (photons). If one does both, the measured momentum of the photons is related to the wavelength of the light by equation (2.7).

What about the brightness of the light? We have already seen that it is related to the square of the amplitude of the electromagnetic wave. But what is the 'brightness' of a *particle* of light—a photon? To answer this question for an individual photon is to delve into some of the subtle strangeness of quantum physics (we will consider these interesting issues somewhat more in Volume III of *The Physics and Art of Photography*). But there is a simple answer for a stream of many photons. For light of a particular wavelength, *a brighter light means that there are more photons per second.*

Reference

Beaver J 2000 The speed of light with a shortwave radio *Phys. Teach.* **38** 172–4

IOP Concise Physics

Chapter 3

What light does

When matter produces light, it is called *emission*; chapter 4 lays out some of the ways in which this can happen. But when light reaches matter, many different things can happen. We call this an *interaction*, and several different types of light–matter interactions are described in the next several sections.

Like matter, light carries both energy and momentum as it travels through space. But unlike matter, light can pass through other light unaffected; *light waves from multiple sources can coexist in the same place at the same time and still retain their separate identities*. This basic fact leads to the rather surprising phenomena of *interference* and *diffraction*, subjects we take up in sections 3.6 and 3.7.

3.1 Reflection, absorption and transmission

What happens when light arrives at some new material, a piece of glass for example? To help us make sense of this, it is useful to talk of the *plane of interface* between the two materials—the material the light came from and the different material the light has arrived at. The plane of interface then, is the *boundary surface* between those two materials. A good example is the boundary surface one side of a glass window makes with the air next to it. The other side of the glass window would then form a second plane of interface with the air next to it. The plane of interface may be literally flat as in this example, but it could also be curved. Think, for example, of the boundary between a crystal ball and the air around it.

When light arrives at a plane of interface with a new material three basic things can happen. Some of the incoming (incident) light can bounce off (reflection), some can pass through into the new material (transmission) and some can disappear altogether (absorption). More often than not, all three of these happen at once, to varying degrees.

But light carries energy, and energy must be conserved. So ultimately, the energy of the incident light must be equal to the sum of the energies of the reflected, transmitted and absorbed light. That is to say, the energy carried by the incident

doi:10.1088/2053-2571/aae1b6ch3

light cannot just disappear; it must be divided up somehow between these three different things that happen to it.

In the case of reflection and transmission, the light itself does not disappear, and so the energy from the incident light carries over to the energies of the reflected and transmitted light. Absorption is different; *when light is absorbed its energy is transferred to other forms of energy* (most commonly thermal energy). And so the energy *as carried by light* decreases, while other forms of energy (thermal energy for example) increase by the same amount.

A complication is that all of these three processes, reflection, transmission and absorption, depend on the wavelength of the light. Ordinary window glass is a good example. At a visible wavelength of 500 nm, window glass transmits a lot of light and reflects little. At wavelengths 20 times that, however, in what we call the thermal infrared part of the electromagnetic spectrum, window glass does the opposite; very little passes through and much more of it reflects. So keep this in mind as we go along. What is good for the goose wavelength may not be good for the gander wavelength.

3.2 Specular reflection

When light bounces off a smooth and shiny surface, it obeys a simple rule, known as the *law of reflection*. Figure 3.1 illustrates the geometry of this type of reflection. Here we have a reflective material on the right, and incident and reflected rays on the left, labeled with the Greek letter θ (theta: it's use is traditional for angles) with subscripts 'i' and 'r' to represent, respectively, 'incident' and 'reflected.' Specular reflections most commonly occur in two situations:

1. Light encounters a smooth, opaque but shiny (usually metallic) surface. This is the ordinary example of a mirror.
2. Light encounters a smooth interface between two *different* transparent materials.

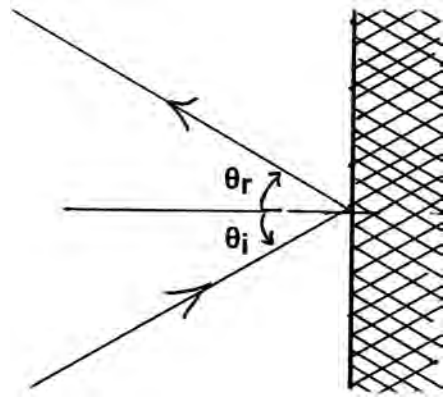

Figure 3.1. The law of reflection. The reflected ray of light makes the same angle with the surface as the incoming (incident) ray. But instead of measuring the angle the ray makes with the surface itself, it is more convenient to measure the angle made with a line *normal* (perpendicular in any direction) to the surface.

If instead of a single ray, a whole bundle of rays all parallel to each other arrives at that same shiny surface, they all reflect by the same angle, since they all arrive at the same angle. This organized type of reflection is called a *specular reflection*, and it is the sort that is produced by a mirror. In section 3.4 we consider equally-important *diffuse reflections* that are not so organized.

We define the direction of these two rays according to an angle made not to the reflective surface, but rather to a line *normal* to that surface. Here the word 'normal' means 'perpendicular to a plane in all directions parallel to that plane.' Place your pencil normal to your horizontal desktop, and you can draw a 90° angle between the pencil and the desktop in all horizontal directions. See figure 3.2. And so in this particular case the pencil points vertically. This is an unambiguous way to describe the relation between a line and a plane, and so we will use it whenever we describe the relation of a light ray to some plane (a new material the ray suddenly encounters, for example).

Given figure 3.1, the law of reflection can be expressed very simply—*the angle of reflection is equal to the angle of incidence*. That is to say:

$$\theta_r = \theta_i \tag{3.1}$$

But what if the reflective surface is curved? The same relation still holds, so long as we take our angles to mean they are made to the normal of the *flat plane locally tangent to the curved surface*. A plane tangent to a curved surface means it touches the curved surface at no other nearby points; see figure 3.3. And so we still have, at any given location, a *local* law of reflection. It is just that the curve of the surface makes the normal lines (and thus the reflected rays as well) point in different directions for different parts of the curve. A curved mirror can thus be used to converge or diverge parallel rays of light.

Figure 3.2. A pencil positioned *normal* to the plane of a horizontal wood floor. When describing rays of light compared to flat *planes*, we measure the angles compared to the normal, not the plane itself.

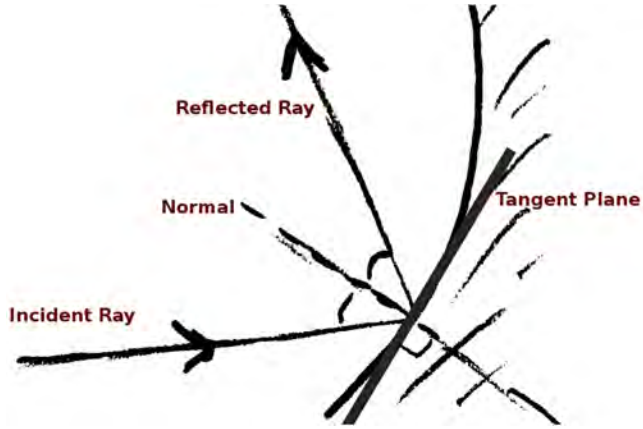

Figure 3.3. Specular reflection from a curved mirror. The ordinary law of reflection can be applied, but it will come out differently for different points on the mirror. And so we imagine a plane *tangent* to the surface at a particular point on the mirror, and measure the angle the ray makes with the normal to this imaginary tangent plane.

3.3 Refraction

While the speed of light in vacuum is a universal constant, the speed is altered when it travels through a transparent material such as water or glass. This is because light is a changing pattern of electricity and magnetism (an electromagnetic wave), and physical materials (glass for example) alter the electric and magnetic properties of space. In the cases of most interest to photography, materials such as transparent glass or plastic weaken electric fields, and this causes light waves to slow down.

It is helpful to define some terms. Let us say that we have two transparent materials, which we label 1 and 2. They could be water and glass, glass and air, air and acrylic, acrylic and water, or any other combination of two transparent materials. As light travels from material 1 to material 2, the plane of interface is the two-dimensional boundary—the surface—where the two materials meet.

See figure 3.4. According to tradition, I've labeled the angles with the Greek letter theta, θ, and I've put subscripts on them to denote which side we are talking about. The key finding is that when this happens, the light makes different angles in the two materials. This bending of light by a sudden change of material is called *refraction*.

The reason the two angles are different is because the speed of light suddenly changes when passing from material 1 to material 2. In the case I illustrate here, the angle with the normal is smaller in material 2 than in material 1. It turns out that this means that light must travel more slowly in material 2. The mathematical relation between the two angles is fairly simple, and it is called *Snell's law*, or *the law of refraction*:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{3.2}$$

n_1 and n_2 are pure numbers (without units or dimensions), and their values are properties of the two materials themselves. This *index of refraction* of a material is a

number greater than or equal to 1, and it is usually determined experimentally for a given transparent material.

The specific example shown in equation (3.2) illustrates one of the most important properties of the law of refraction: *the angle with the normal is smallest on the side where the index of refraction is largest*. As is the case for reflection, we apply Snell's law to a curved surface by measuring the angles with the normal to the plane tangent to the curved surface at each point. See figure 3.5 for an example of light refracted to the bottom of a lake by the wavy surface of the shallow water.

And so what does the law of refraction have to do with the speed at which light travels in the two materials? A purely geometrical analysis of light provides no answer except that 'it is what it is,' and this is how we observe light to behave when we perform experiments.

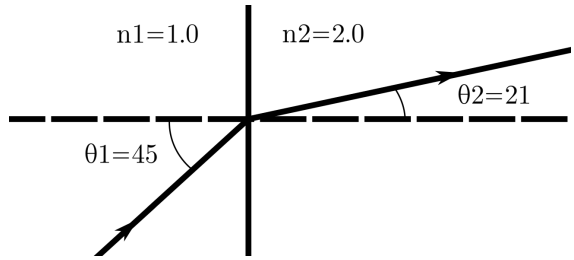

Figure 3.4. Refraction and the plane of interface between two materials. The *plane of interface* is the border between two different kinds of transparent materials. As light changes speed going from one material to the other, it bends at this plane of interface. It does this in such a way that it makes the *smallest* angle with the normal in the material through which light travels the *slowest*, and thus has the *largest* index of refraction, *n*.

Figure 3.5. The wavy surface of the water refracts sunlight into a complex pattern on the lake bottom.

For a physical *explanation* of Snell's law, we must consider the wavelike nature of light. To do this we will use one of the most important tools of wave optics: *Huygen's construction*, invented by C Huygens in 1678.

But first, let us define the concept of a *wave front*. If light moves coherently in a particular direction in a given region of space, we can imagine a line that connects all of the places where a particular wave is at a peak. In physical terms, for a light wave, this could be the points in space where, at a given instant of time, the electric field is at its maximum. In three dimensions this region would be a plane (not necessarily flat), but on a two-dimensional (2D) drawing this will appear as only a line (not necessarily straight). An important point is that *the direction of travel of a wave is always perpendicular to the wave front*. Since our 'rays' point in the direction of travel of the light wave, it follows that the wave fronts are perpendicular to the rays.

Huygen's construction says that if the location of a wave front is known, the next one can be found by imagining an infinite number of spherical wavelets emanating simultaneously from different parts of the first wave front, and *all propagating for the exact same amount of time*. The next wave front is then marked by a tangent that connects the edges of all of those spherical wavelets. For a 2D drawing, this means we can imagine circular waves propagating outward from every point on the first wave front. We stop them at a time of our choosing, connect them with a tangent line, and we have found the next wave front.

So let us look at an example such as figure 3.4, but this time drawn with a wave front indicated perpendicular to the incident ray; see figure 3.6. I have chosen to put the first wave front so it is starting right at the boundary between material 1 and material 2. I pick a point on the incident wave front, and I allow just enough time to pass for an imaginary spherical wave coming from that point to just make it to

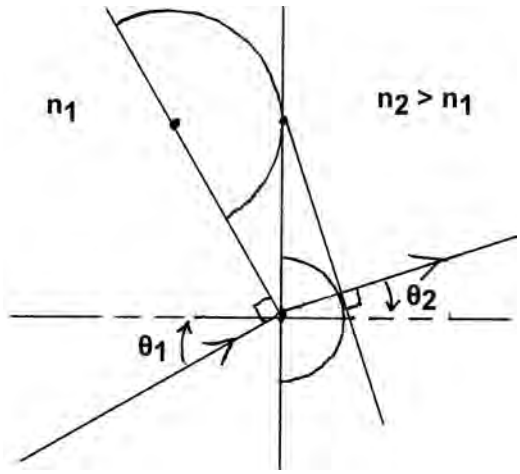

Figure 3.6. Wave fronts in refraction. Because light travels more slowly in the material on the right (higher index of refraction), a wave front in that material travels a smaller distance in the same amount of time. Connecting the wavefronts, according to Huygen's construction, shows that the light must bend in toward the normal as it enters the higher-index material.

material 2. Meanwhile, I allow another imaginary spherical wave to begin at the exact point the ray intersects with material 2.

But remember that light travels more slowly in material 2, so in the same amount of time, this imaginary circular wave doesn't travel as far. So where is the next wave front? Huygen's construction says that one simply needs to connect those circular waves with tangents. This second wave front is completely within material 2 while the first wave front was completely within material 1. If we recall that the rays are always perpendicular to the wave fronts, we have now found in what direction must be the second ray.

Clearly, the ray in material 2 makes a smaller angle with the axis than does the ray in material 1. In fact, if we analyze figure 3.4 with a little trigonometry, it is not hard to show that there is a fairly simple mathematical relation between θ_1 and θ_2:

$$\frac{c}{v_1}\sin\theta_1 = \frac{c}{v_2}\sin\theta_2 \tag{3.3}$$

where c is the speed of light in a vacuum, v_1 is the speed of light in material 1 and v_2 is the speed of light in material 2. It is easy to see that equation (3.3) is Snell's law—equation (3.2)—so long as we make the following identification for each of the two materials:

$$n = \frac{c}{v} \tag{3.4}$$

where n is the index of refraction in a particular material and v is the speed of light in that material.

In a purely geometrical optics, the index of refraction is simply a number that is measured by experiment for a given material—a piece of glass, for example. Wave optics and Huygen's construction reveal its physical meaning. The index of refraction tells us by what factor the material reduces the speed of light as compared

Table 3.1. Angles of refraction for rays of light entering glass (with $n = 1.5$) from air. The incoming (incident) ray in the air makes an angle θ_i with the normal. It is then refracted, and so makes a different angle, θ_r, with the normal while inside the glass.

θ_i	θ_r
0°	0°
10°	6.65°
20°	13.18°
40°	25.37°
60°	35.26°
80°	41.03°
85°	41.62°
88°	41.78°
89°	41.80°
89.5°	41.81°
89.9°	41.81°

to c, its speed in vacuum. And so, for example, if the index of refraction of a particular type of glass is $n = 1.523$, it means simply that light travels 1.523 times more slowly in that particular type of glass than it does in a vacuum.

The second column of table 3.2 gives approximate indexes of refraction for some different materials. Since in all but rather odd circumstances a light wave travels more slowly in a material than in a vacuum, indexes of refraction are greater than 1.0. An important point to remember is that the index of refraction of air is 1.0003, very nearly 1.0 exactly. A perfect vacuum really does have an index of refraction of exactly 1.0 (this should be obvious from its definition), and most gases have approximately that value as well.

Notice that Snell's law makes no reference to which side the light was coming from and which side it was traveling toward. Snell's law simply says that the angle is smallest where n is largest. And so we could reverse the arrows on the rays in figure 3.4 and it would still be a valid diagram.

The same could not be said, however, for figure 3.7, where we have taken into account the fact that when light gets to a sudden change in material, a certain percentage will *reflect*, instead of passing into the material and refracting. If we reverse the arrows on the top image, the result instead is something like what is shown in the bottom image, where the ray of light moves from plastic to air instead of from air to plastic. Notice the refracted rays do the same thing in both diagrams; the angle is larger in air and smaller in plastic. But not so for the reflected ray; it always reflects from whichever side the light is incident.

The images in figures 3.7 also illustrate a common misconception about light rays. A light ray is an abstract idea; the ray itself is not *visible* from the side; you can only see it if it points directly at your eye. A laser makes a very good approximation of a ray of light, and so why does the 'ray' of laser light show up in these pictures even though the lasers point not towards the camera, but rather crosswise to our vantage point?

It is because of a trick. I used a particular combination of laser and plastic to take advantage of *fluorescence*, a topic we consider in section 3.8. In this example, a small fraction of the laser light is absorbed by the plastic, and that energy is used by the

Table 3.2. Critical angles for light coming from various materials into air.

Material	Index of refraction	θ_{crit}
Water	1.33	48.8°
Plexiglass	1.49	42.2°
Salt	1.54	40.5°
Crown glass	1.52–1.62	38.1°–41.1°
Flint glass	1.57–1.75	34.8°–39.6°
Sapphire	1.77	34.4°
Lanthanum glass	1.82–1.98	30.3°–33.3°
Diamond	2.42	24.4°

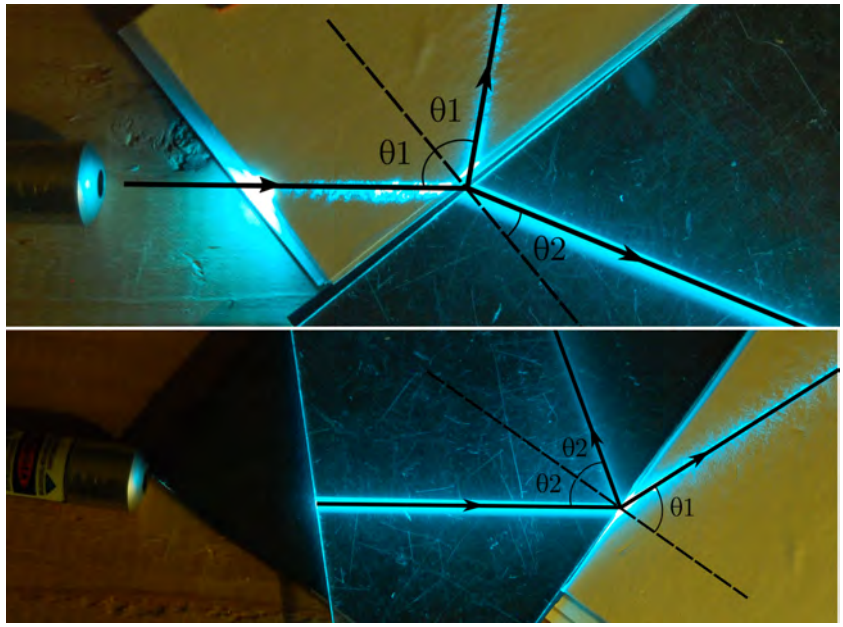

Figure 3.7. Top: as light arrives at a sudden change in index of refraction, the light will refract upon entering the new material. But a portion of the light will also reflect off the plane of interface. Here, light traveling through air (material #1) refracts as it passes into plastic (material #2). Notice that $\theta 1$ (air) is *larger* than $\theta 2$ (plastic), because the index of refraction of air is *smaller* than that of plastic. Bottom: the same arrangement as in the top image, but with the light instead coming from inside the plastic and refracting as it emerges into the air. Notice that, exactly as in the top image, $\theta 1$ (air) is *larger* than $\theta 2$ (plastic), because the index of refraction of air is *smaller* than that of plastic. But this time the reflected ray is inside of the plastic instead of in the air.

plastic to emit its own light. But it emits that light not in the direction the laser light was traveling, but rather in random directions. So the path of the laser light is marked by the fluorescent glow of the plastic sending light in all directions—including toward the camera. For the part where there is no plastic, I let the laser beam partially graze along the surface of a sheet of white paper. You see the little bit of the intercepted beam, scattered toward you by diffuse reflections from the paper, and thus tracing out its path. Notice that there is a portion of the laser light's path (on the left side) for which neither of these tricks is performed—and that part of the beam is invisible from the vantage point of the camera.

The amount of refraction depends not only on the angle of incidence, but also on the *ratio* of the indexes of refraction of the two sides. As an example, table 3.1 gives values for the angles of refraction, given various angles of incidence, for the special case of a ray of light passing from air ($n = 1.0$) into some type of glass with $n = 1.5$.

3.3.1 Total internal reflection

Notice in table 3.1 that for a ray of light passing from air to glass, the angle of refraction seems to approach a maximum value (41.81° in this case), as the angle of incidence approaches 90°. Clearly, the angle of incidence cannot be greater than 90°,

simply as a matter of logic. Figure 3.8 shows this limiting case, which can only be approached.

Remember that for refraction, one can reverse the arrows and the diagram will still be valid. Table 3.1 shows the angles of incidence and refraction for rays of light in air entering into a piece of glass with index of refraction $n = 1.5$. But we could instead imagine the reverse—rays of light coming from the glass and emerging into the air—and all we would need to do is reverse the columns of table 3.1.

But a moment's thought reveals something odd. The reversed table would only show angles of incidence of 41.81° or smaller, and those would have refracted angles of almost 90°. And so what would happen if one simply made the angle of incidence *greater than* 41.8°? The numbers in the table seem to imply that a refracted angle of *greater than* 90° would result. But what would that even mean? An angle with the normal of greater than 90° means the light has not even left the glass.

So what does Snell's law say happens in this case? You can find out by using a calculator to plug $n_1 = 1.0$, $n_2 = 1.5$ and $\theta_2 = 45°$ into Snell's law, and asking it to solve for θ_1. It will give you an error message. Snell's law is simply invalid for this case, and something else happens that is not predicted by Snell's Law alone.

Let us now consider the *reflected* ray in the bottom image of figure 3.7. If we were now to gradually increase the angle of incidence, four things would happen:

1. The angle of refraction would increase, according to Snell's law. Since it is in the side with the smaller index of refraction, the angle of refraction would increase faster than we increase the angle of incidence.
2. The reflected ray would increase according to the law of reflection. Which is say that the reflected ray would increase so it is always equal to the angle of incidence.
3. The refracted ray would get dimmer.
4. The reflected ray would get brighter.

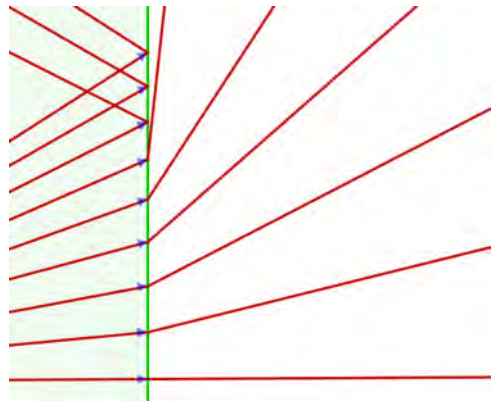

Figure 3.8. If light comes from a material with a *larger* index of refraction, there will be a particular angle of incidence at which the refracted ray would be 90°. For angles of incidence equal or greater than this, there is no refracted ray, and all of the light reflects internally. This is called *total internal reflection*, and it is often used in place of mirrors made with shiny metal coatings.

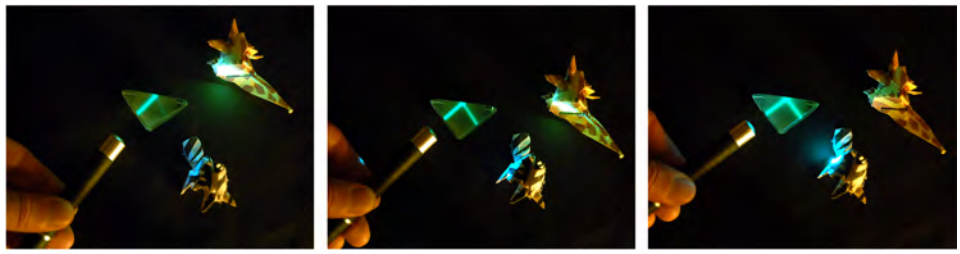

Figure 3.9. A 45° glass prism used to reflect light by 90°. Left: the light meets the backside of the prism at less than the critical angle. It mostly passes through, deflected by refraction. Notice, however, that a small portion of the light reflects off the surface internally, like a mirror. Center: the light meets the back face of the prism at very nearly the critical angle, and refracts almost parallel to the back face of the prism. Notice that the internally reflected portion is now much brighter. Right: the ray meets the back surface at greater than the critical angle, and all of the light reflects internally.

As one increases the angle of incidence, the refracted ray gets dimmer and dimmer, *disappearing completely as the angle of refraction approaches 90°*. Simultaneously, the reflected ray gets brighter, becoming as bright as the incident ray as the refracted ray disappears. If one then increases the angle of incidence even further, there is only a reflected ray, bouncing inside the glass and never getting out into the air. This is called *total internal reflection*, and it is very useful; it is, for example, a way to make a mirror from nothing but clear glass. See figure 3.9 for a demonstration.

From table 3.1 it is clear that, for the case of a ray of light trying to get out of a piece of glass with air around it, so long as the angle of incidence is greater than about 42°, the light will reflect inside the glass. This is called the critical angle, θ_{crit}, and we can calculate it as follows[1]:

$$\theta_{\text{crit}} = \sin^{-1}\left(\frac{n_1}{n_2}\right) \tag{3.5}$$

Table 3.2 shows indexes of refraction and critical angles for several different materials, assuming air ($n = 1.0$) is on the other side of the interface.

Total internal reflection has many practical uses. The right-angle prism of figure 3.9 is in some cases a better way to reflect light than the shiny metal surface of a good mirror. Fiber optics employ thin transparent fibers of glass or plastic. Once a ray of light enters, it inevitably meets the outside edge of the fiber at a very large angle to the normal, and so undergoes total internal reflection many times as it bounces from one edge to the other, even if the fiber is tied into knots. And even beads of water can act like little mirrors, if there is a layer of transparent air trapped underneath. See figure 3.10.

3.3.2 Dispersion

The index of refraction for a typical transparent material is not actually a simple constant; it depends somewhat on the wavelength of the light. Most often, n is

[1] In equation (3.5), $\sin^{-1}$ represents the *arcsine* or *inverse sine* of the number. This is the angle one would have to take the sine of in order to get that number. And so the $\sin(0°) = 1$ and $\sin^{-1}(1) = 0°$.

Figure 3.10. Left: tiny hairs on the surface of these leaves of American Lotus trap air underneath the water drops. When light rays passing through the water drop encounter this layer of air at greater than the critical angle of 48.8°, total internal reflection occurs. And so the transparent drops of water look like little mirrors reflecting the sky. Right: a bundle of optical fibers uses total internal reflection to guide light around tight curves.

slightly larger for *shorter* wavelengths. This means that the angle of refraction for a given incident ray of light depends on the wavelength of the light, with shorter wavelengths usually refracted by larger angles. This effect is called *dispersion*, and a graph of index of refraction versus wavelength, for a given material, is called the *dispersion curve* of the material.

The left side of figure 3.11 shows examples for some different types of glass. Notice that the graph also shows *overall* differences in index of refraction. And so for example at any given wavelength SF10 glass has a greater index of refraction than BK7 glass. But the blue curve for SF10 is its dispersion curve—it shows a different index of refraction for each wavelength. Also notice that some of the dispersion curves seem to zoom up to huge values at some very short wavelength in the ultraviolet part of the spectrum. The dispersion curve for F2 flint glass, for example, increases very rapidly at a wavelength of about 0.25 μm (250 nm). What the dispersion curve does not show is that the glass also becomes less and less transparent at these wavelengths. And if the glass does not transmit light, then the index of refraction is a moot point.

Most sources of light consist of many wavelengths simultaneously. And thus, if a ray of such a source of light enters a piece of glass, it is not all refracted at the same angle. Shorter wavelengths are refracted at a greater angle, and so the piece of glass *disperses* light of many wavelengths (but one direction) into many different directions (but each consisting of only one wavelength). See the right side of figure 3.11 for an illustration.

Dispersion can be either useful or a pain in the neck, depending on the situation. If one wants to measure the spectrum of a source of light, then a wedge-shaped piece of glass (a prism) can do the trick. Different wavelengths will refract at different angles, and so one can measure them separately, as the prism will deflect them by different angles. If the dispersion curve for the type of glass is known, one can mathematically convert this information into wavelength, and thus produce a graph of brightness versus wavelength—a spectrum.

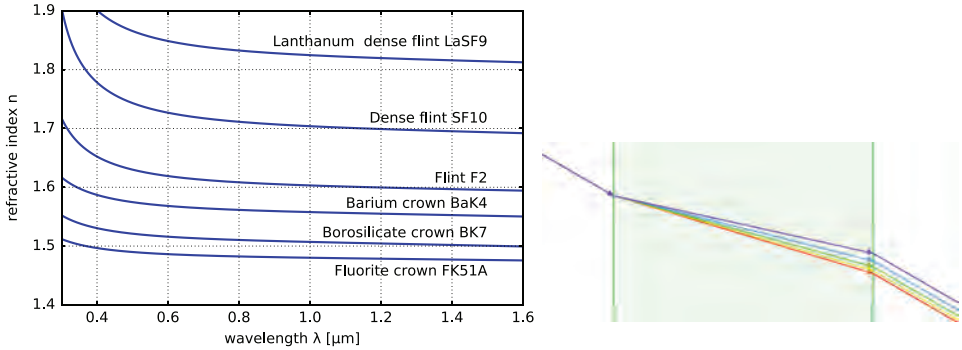

Figure 3.11. Left: dispersion curves for several different types of glass. The vertical axis represents index of refraction while the horizontal axis is wavelength. The shaded area marks the visible portion of the spectrum. We can use the dispersion curve to read the index of refraction for any particular wavelength. For most materials the index of refraction is higher at shorter wavelengths. This property of dispersion means that different wavelengths bend by different amounts when light of many wavelengths refracts in the material (graphic: Geek3 - Own work, CC BY 4.0). Right: an example of dispersion by a piece of glass.

But if the goal is instead to design a lens to produce sharp images, then dispersion is the enemy. It means that if one designs the shape of the lens so as to focus red light, then blue light will not be in focus, because the two wavelengths are refracted by different angles. This basic problem is called *chromatic aberration*, and it is discussed in more detail in chapter 6, section 9.6. For now, suffice to say that overcoming chromatic aberration usually requires both much cleverness and expense.

If we consider the phenomena of dispersion, refraction and total internal reflection, we can begin to see why different transparent materials 'look' different. A diamond has facets, and this means light refracts as it makes different angles with the different facets. But also, there is dispersion, so different colors refract differently, and one can see colors even with white light. Finally, many rays get trapped inside, undergoing total internal reflection multiple times before finally meeting a facet at less than the critical angle, and so escaping. Thus a diamond can look colorful and glittery.

Why does a piece of glass ground to exactly the same shape not look as impressive as a diamond? It is because of the considerably-smaller index of refraction of glass compared to diamond, which means less dispersion, smaller angles of refraction, and less internal reflection.

3.4 Diffuse reflections

It is often the case that when light reflects off an object, a single ray will reflect in many directions because of microscopic irregularities in the surface. This is called a *diffuse reflection*, in contrast to the specular reflection already discussed. If one could look closely enough, a diffuse reflection is really a myriad of specular reflections. Instead of one ray, there is really a bundle of many parallel rays very close together. If the surface is rough, then each individual ray in this thin bundle hits a seemingly randomly-oriented surface, and thus goes off in its own peculiar direction.

A mirror is a surface that makes only specular reflections and no diffuse reflections. To be a good one, the mirror surface must be smooth down to the level of the wavelength of the light itself. And so, for example, if one grinds glass with successively finer and finer abrasives, it will appear perfectly smooth only once the abrasives (and thus the scratches made in the glass) are smaller than the wavelength of light used to reflect off it. At that point is will produce a good specular reflection, and thus appear mirror-like and shiny. With larger scratches, from larger abrasives, the surface will appear dull from the diffuse reflection.

Diffuse reflection is the most common process by which we 'see' solid, opaque objects. A cat (see the left side of figure 3.12) does not create light of its own, at least not in the visible part of the spectrum. We see light coming from it because light from some other source has reflected off it, making diffuse reflections with every part of the cat. Thus rays of light go off *in every direction* from *each part of the cat*, just as if it were emitting its own light. Our eyes intercept whatever rays, coming from each part of the cat, happen to be going in our direction. The rest go elsewhere (toward a frightened mouse, for example).

We can see the difference between specular and diffuse reflections by looking at the right side of figure 3.12. The rays of light that came from the different objects in the picture, and arrived at the camera, were all due ultimately to reflection from the room lighting. Light coming from the eggs was due to diffuse reflection. In this case, some parts appear light while other parts appear dark because of differing amounts of diffuse reflection; whatever was not reflected was absorbed. Since those diffuse reflections go off in all directions, it is guaranteed that some of the rays from each part of the eggs make it to the camera.

The light coming from the glass surfaces, however, was due to specular reflections. In addition some light reflected diffusely off the eggs in the background has passed through the glass, its path altered by refraction on its way to the camera. In these cases, what appears light and what appears dark is much more complex; it

Figure 3.12. Left: *Boris and Quail Eggs*. John Beaver, 2010. A cat does not emit light of its own in the visible part of the spectrum. We see it because of the diffuse reflection of light from other sources. Right: *Bottle and Quail Eggs*. John Beaver, 2012. The light from the eggs is due to diffuse reflection (see section 3.4) while that from the surfaces of the metal and glass is due to specular reflections. In addition, some light from behind has passed through the glass and been altered by refraction (see section 3.3).

depends on the location of the camera, the location of the light source, and the angle and curvature of the surfaces themselves as they both reflect and refract the light.

3.5 Scattering

A light ray is *scattered* if, upon interacting with matter, it splits into many rays in random directions. A diffuse reflection from a solid, opaque object can be thought of as a form of scattering, but the term is usually used to describe other physical processes that deflect light randomly from within transparent or semi-transparent materials.

There are many kinds of scattering, but there are two main categories, wavelength-dependent and wavelength-independent scattering. Wavelength-dependent scattering occurs when different wavelengths are scattered by different amounts. Wavelength-independent scattering on the other hand affects all wavelengths equally.

3.5.1 Wavelength-dependent scattering

The most common experience of wavelength-dependent scattering is called Rayleigh scattering, and it occurs when light interacts with particles, such as individual atoms or molecules, that are much smaller than the wavelength of the light. Rayleigh scattering of sunlight by air molecules is responsible for the blue of the sky on a clear day. As individual photons encounter individual air molecules (N_2 or O_2), a certain percentage of the light will deflect off in a different direction. But *short wavelengths scatter by larger angles than long wavelengths.*

In this case, much of the deflection is at angles less than 90°, and so it is mostly scattering into the same overall direction, a process called *forward scattering*. A significant fraction, however, is also *back scattered* to angles between 90° and 180°.

Since it is the shorter-wavelength blue light that scatters the most, when one looks up at a cloudless daytime sky one sees blue light. This blue sky light comes from the Sun, ultimately, but it is light that would have missed you, had it not been deflected in your direction by Rayleigh scattering.

On the other hand, the setting Sun appears noticeably reddish. In this case you are seeing the light that has *not* been scattered out of your line of sight. Since it is the blue light that mostly scatters off in other directions, you see sunlight that has had blue light subtracted from it. Thus the Sun appears more reddish.

A kind of wavelength-dependent scattering very much like Rayleigh scattering can also occur with much larger particles, closer in size to the wavelength of the light itself. For visible light, that means particles that are just a bit too small to be seen with an ordinary microscope. But this size is still enormous compared to the individual atoms and molecules that cause Rayleigh scattering.

This type of scattering from larger particles is called the Tyndall effect, and probably the most commonly-seen example is the blue smoke from the tailpipe of an automobile engine that is burning oil and needs an overhaul. The smoke looks blue for the same reason the sky looks blue. You see it by light from other sources scattering off of it, and short wavelengths scatter more than long wavelengths. Not

all smoke does this, because in order for the Tyndall effect to be prominent, the smoke particles must be the right size.

The Tyndall effect can also be seen in certain liquids that have sub-microscopic globules of one material suspended in another. A tiny bit of milk mixed with water sometimes works, and looks bluish when illuminated from the side. Glass can be manufactured with impurities that emphasize the Tyndall effect, and these *opalescent glasses* are often used for their ornamental properties.

Figure 3.13 shows two photographs of an opalescent glass egg. When white light (of a mix of wavelengths) shines on it, you see only the light scattered toward you. And so it appears blue. When you look *through* it at a source of light behind it, on the other hand, you see the reddish light that has *not* been scattered.

3.5.2 Wavelength-independent scattering

Wavelength-independent scattering occurs when light interacts with particles of matter that are significantly larger than the wavelength of light. In that case, the physics is simpler, but the overall effect is similar; an incoming parallel bundle of rays is 'scattered' into many rays traveling in many directions. But the seemingly-random nature comes not from the probabilistic laws of quantum physics as individual photons interact with matter, but rather from the too-complex-to-describe microscopic details of the scattering material.

Light shining on microscopic particles of dust provides a good example. If one takes a high-power laser pointer and shines it up into the night sky, it seems as though one can see the beam of light. But the light from the laser is all going in the same direction, and this is not toward you (or at least it had better not be). What you are seeing is light scattered toward you from zillions of microscopic dust particles

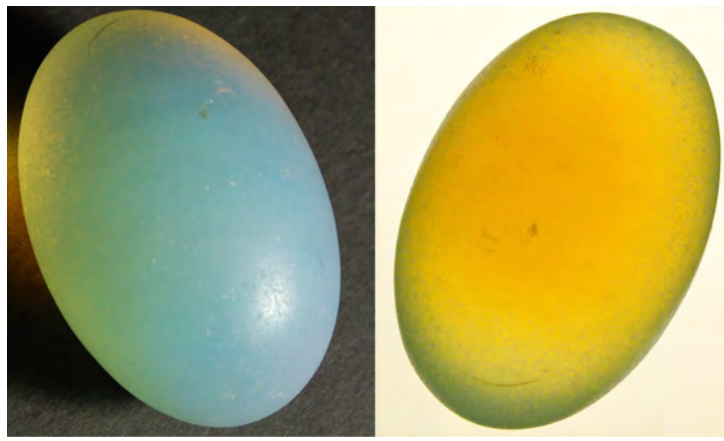

Figure 3.13. Tyndall-effect scattering in an egg-shaped piece of opalescent glass. The photo on the left was taken with the egg against a black background and illuminated from above. We see the mostly-blue light that is scattered back towards us. The photo on the right was taken with the egg illuminated from behind. In this case we see the light that was *not* scattered by the egg on its way to us. Since mostly blue light is scattered, what is left over appears reddish in color.

located all along the beam. Since this scattering occurs most strongly in the backward direction, the beam is more visible if one looks along it from behind.

Fog provides another example of scattering from relatively large particles. Microscopic droplets of water suspended in the air alter the directions of the rays of light by both refraction and total internal reflection. If the fog fills a given space, then a portion of the light that reaches you may have been scattered many times before it gets to you. But fog scatters all visible wavelengths equally, for the most part.

A puffy white cloud in the blue sky appears as it does due to scattering. It is, after all, nothing but countless microscopic, transparent droplets of water or crystals of ice. These tiny particles deflect light rays, each according to the basic laws of refraction and reflection. But the details are far too complex to follow individually, and so the light is scattered in random directions.

In dense fog, light is scattered many times before it gets to you, and it appears as though the air itself is emitting light. Light rays coming from more distant objects have more opportunities to scatter on their way to you, and so the light coming directly from objects is more diminished with greater distance. But while light that would have reached you (had there been no fog) is scattered out of your line of sight, other light—from random directions that would have missed you had there been no fog—is scattered in your direction. Thus the diminished light from distant objects is filled in by a random glow, and this reduces *contrast*.

Fog has little effect on your view of nearby objects, but it has a profound effect on distant objects. For photography, this means the presence of fog can add depth to a picture. Haze causes a mixture of both wavelength-dependent and wavelength-independent scattering, and it too can add depth to a landscape. It makes distant objects appear of lower contrast and slightly more bluish or purplish in color. Painters often use the term *perspective* to mean, in a generic sense, any type of depth cue. And so when a painter adds features that mimic the visual effects of haze scattering in order to show depth in a painting, it is called *atmospheric perspective* (see figure 3.14). But for a photographer, who must deal with whatever haze there happens to be when taking a picture, the lowering of contrast by haze scattering may or may not be desirable.

3.6 Interference

It is possible to put a thin coating of transparent material onto the surface of a glass lens that actually *increases* the amount of light that passes through the lens. It is surprising that one could put something on glass to make it more transparent; would not anything you put on it simply absorb more light and make matters worse? How could the coating cause *more* light to pass through?

When looking through a typical glass lens, most loss of light is due not to absorption in the glass, but rather because a significant percentage of the light reflects off the glass surface instead of passing through. An *anti-reflection coating* decreases the reflection, while increasing the transmission. It is easy to recognize such a lens from the faint greenish or bluish tint of light that glances off the lens surface.

Figure 3.14. *Rio from Sugar Loaf Mountain*. John Beaver, 2012. *Atmospheric perspective*, which can give a sense of physical depth to a picture, is here introduced by haze and fog.

It is possible to do this by taking advantage of the wave nature of light, in this case making use of a phenomenon called *thin-film interference*, just one example of the phenomenon of interference in general. All kinds of waves, not light alone, are able to interfere with each other; this is one of the defining properties of waves.

If two waves are present at the same place at the same time, they add together. A given wave, at a given point in space and time, causes some kind of disturbance, and this disturbance can be in either one direction or its opposite. We can (arbitrarily) call one direction positive and the other negative. For light it is the direction of the electric and magnetic fields. For sound in air, it is the direction of slight displacements of air molecules.

If one of the two waves is trying to make stuff zig, while the other is trying to make it zag, it is possible for both effects to cancel out, leaving nothing, an effect called *destructive interference. Constructive interference*, on the other hand, means that both waves are trying to cause the same disturbance at the same place and time—and so a wave of twice the amplitude results. There are also all of the possible circumstances in between, where the two waves either partially add or partially cancel each other. See figures 3.15 and 3.16.

There are many ways, in practice, to make this happen. For light, one can split a wave with some kind of obstacle, and then send both waves off on different paths, eventually bringing them back together. When they recombine, they will most likely no longer line up with each other, and so they will interfere. *Thin-film interference* is a good example. If one takes a very thin layer of transparent material, light falling on it will reflect partly off the top and partly off the bottom. See figure 3.17. These two reflected waves will then interfere, either constructively or destructively

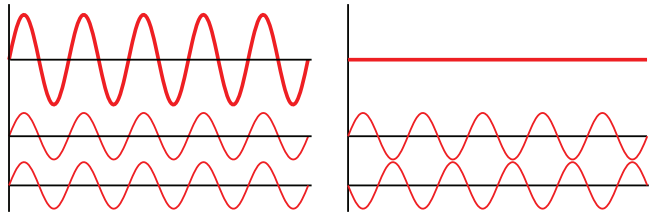

Figure 3.15. When two waves arrive in the same place at the same time, they may undergo *constructive interference* (left) or *destructive interference* (right), or something in between the two (graphic: Haade, Wjh31, Quibik - Own work, CC BY 3.0).

Figure 3.16. The waves made by these water striders show interference. At any given location, the height of the water is given by the sum of whatever waves happen to be at that place and time. In some places, waves can cancel out, while in others they add together.

depending on the details of the index of refraction of the film, its thickness, and what material is on either side.

And so an *anti-reflection coating* on a lens is cleverly chosen so that visible wavelengths of light interfere destructively for reflection, but constructively for transmission. Technically, for a single thin layer, this only happens completely for a rather narrow range of visible wavelengths. But if one uses many layers of coatings, all carefully chosen in thickness and index of refraction, then a broad range of reflecting wavelengths can be canceled, while transmitted wavelengths are reinforced.

Thus one can find three types of lenses: uncoated, single-coated and multi-coated. Uncoated lenses reflect, typically, about 10% of the light that falls on them, transmitting 90%. This might not seem so bad, but most optical systems (camera lenses included) require many lenses, each reflecting 10% of the light at every

boundary, both front and back. And that *is* bad, and not only because less light gets through, resulting in a dimmer image. Worse still, half of those internally-reflected rays *do* make it through eventually, but they are no longer where they should be, and so they add a diffuse glow to the image which reduces its contrast.

Single-coated lenses help with this greatly, but the best lenses are multi-coasted. Multi-coated lenses only reflect about 1% of the light (thus transmitting 99%). Since a real optical system contains many lenses, the best is 'fully multi-coated,' which means all lens elements are multi-coated at every surface.

Figure 3.18 shows examples of the reflections from the surfaces of both coated and uncoated lenses. Notice that each lens shows *multiple* reflections; what looks like

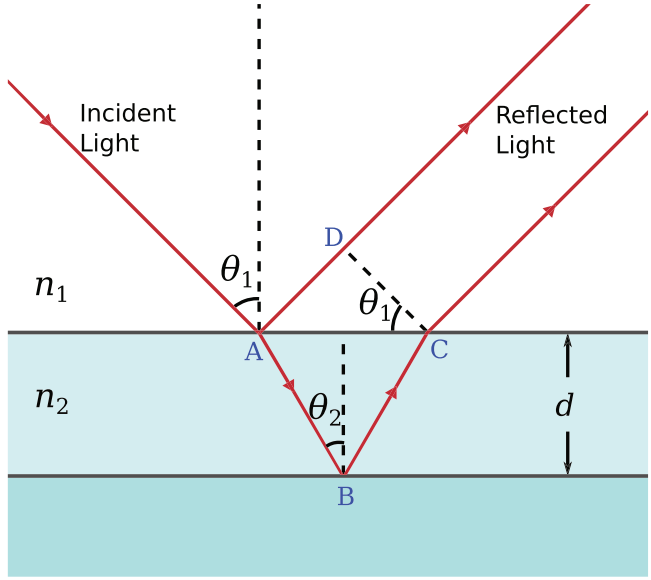

Figure 3.17. A wave can reflect off both the front and back surface of a thin transparent film. The two reflected waves undergo interference due to the fact that they have now traveled by different paths, and so are no longer in phase. Whether constructive or destructive interference occurs (or something in between) depends on the details of the wavelength of the light and the thickness and index of refraction of the thin film, as well as what materials are on either side of the film (graphic: Nicoguaro - Own work, CC BY 4.0).

Figure 3.18. The lens on the left (from the 1950s) is single coated, while the even-older lens on the right is uncoated. Note the slightly-purplish and much-fainter reflections from the coated lens.

a single lens is actually several lens elements one behind the other, and each surface causes a reflection. But note also that the reflections off of the coated lenses are much dimmer than the reflections off the surfaces of the uncoated lenses.

The purplish color of the reflections from the coated lenses arises because, while some wavelengths interfere destructively, others interfere constructively. The type and thickness of these lens coatings have been chosen to give destructive interference for reflected light that has wavelengths near the middle of the visible spectrum. But this means that reflected wavelengths near the violet edge of the visible spectrum are not canceled so well. And so the reflection appears purplish in color from these single-coated lenses. Multi-coated lenses show reflections that are even dimmer still, and in most versions show a slight greenish tinge.

3.7 Diffraction

Recall Huygen's construction, from section 3.3. It says that we can find the location of the next wave front by imagining a bunch of spherical wavelets propagating from the most recent wave front. We can use this to determine what happens when light encounters some kind of obstacle. For example, we can use Huygen's construction to determine what happens to light when it encounters a thin slit (see figure 3.19). The next wave front can be found by imagining spherical wavelets emitted simultaneously from every location along the slit.

But we know that when one adds up light waves, they interfere. So to figure out what would happen, one would need to add up those waves at every point, taking into account the fact that, by the time they all get to a particular point in space, *each had traveled a different distance, and so would be in a different part of its wave cycle*. The specific part of a wave cycle is called the *phase* of a wave, and it is the difference in phase between two waves that results in either constructive or destructive interference.

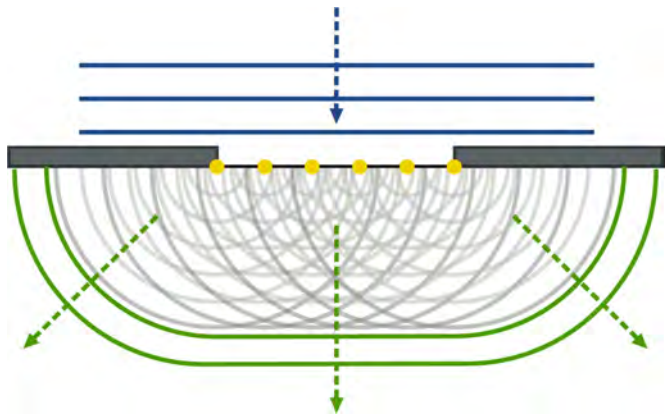

Figure 3.19. Huygen's wavelets propagating from a thin slit and interfering with each other. Huygen's construction allows us to model this by imagining many tiny wavelets emitted simultaneously from different parts of the slit. In some places these wavelengths will line up and cause constructive interference, while in other places they will cancel each other out (graphic: Arne Nordmann (norro) - Own illustration, CC BY-SA 3.0).

And so an infinite number of wavelets, sent out from all parts of the slit, arrive together at some distant point in space. But they arrive with different phases, and so they interfere with each other. This situation is called *diffraction*: the addition of an infinite number of waves, all of different phases (and sometimes also of different intensities). The mathematics of this requires calculus, which is intimately concerned with the business of adding up an infinite number of slightly different things. Here I will simply give you the answers for some important cases. Even without working through the mathematics in detail, we can still recognize some basic principles of diffraction theory:

1. When light encounters an obstacle that restricts it to a particular region of space, it will spread out into the region that one would expect there to be a shadow. Sending light through a slit is a good example. It will spread out beyond the confines of the slit.
2. The angle of spreading is *greater* for a *smaller* restriction (a narrower slit, for example). And so a more narrow slit causes the light to spread out more.
3. All else being equal, the spreading out of the light is greater for longer wavelengths and less for shorter wavelengths.
4. The spreading of the light is proportional to the ratio of the wavelength to the size of the obstacle. So, for a slit, it is the value of the wavelength divided by the width of the slit that decides how much the light spreads out.
5. This process of diffraction will produce an alternating pattern of constructive interference (bright light) and destructive interference (dim light). The size of this pattern when projected onto a screen is proportional to the ratio of the wavelength to the size of the obstacle. Such a pattern is called a *diffraction pattern*, and the bright regions of constructive interference are called *fringes*.

Consider the single slit as an example. The left side of figure 3.20 shows the diffraction pattern for a slit, both as an image and as a graph of intensity versus position. If the slit has a width, a, then θ is the angle by which the diffraction pattern spreads, as measured from its center. Most of the light ends up in one central region of constructive interference. The half-width of this bright central fringe is given by:

$$\sin\theta = \lambda/a \tag{3.6}$$

where λ is the wavelength of the light. Notice that it is the *ratio* of λ to a that counts—not either one by itself. If a is much greater than the wavelength of the light, then the angle between the fringes is very small. For example, if $a = 1$ cm, then for green light of $\lambda = 500$ nm, we have $\sin\theta = 5.0 \times 10^{-5}$, which is an imperceptibly small angle, and the diffraction pattern would not even be visible.

If one wants a noticeable diffraction pattern, the slit must be much smaller. For example, if one shines 500 nm light through a slit the width of a human hair, about 50 μ across, then:

$$\sin\theta = \frac{500 \times 10^{-9}\text{m}}{50 \times 10^{-6}\text{m}} \tag{3.7}$$

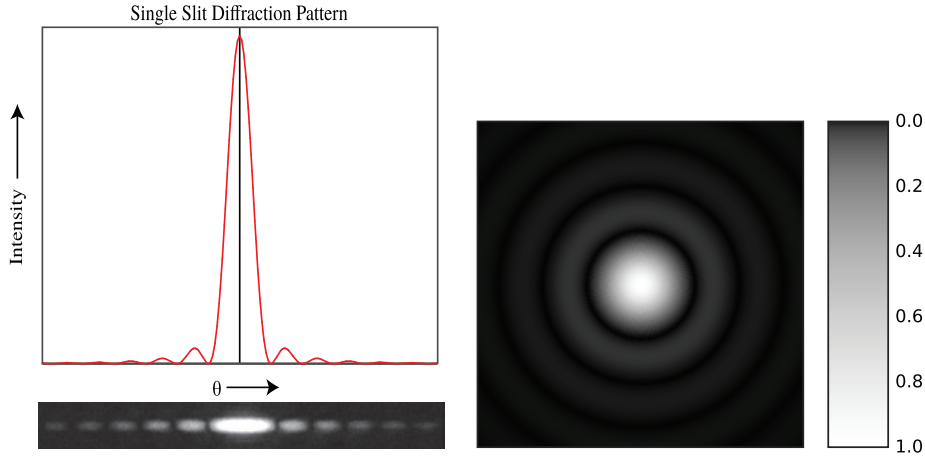

Figure 3.20. Left: the diffraction pattern from coherent light shining through a single slit. The pattern as projected onto a screen is at the bottom. At the top is a graph of the pattern's brightness versus position. A more narrow slit produces a *larger* diffraction pattern. For the pattern to be large enough to see easily with the naked eye, the width of the slit must be microscopic in size. Right: diffraction pattern for a circular hole. A circular hole produces a bullseye pattern. Light shining through a *smaller* hole will produce a *larger* pattern. For the pattern to be large enough to see easily with the naked eye, the hole must be microscopic in size.

$$= 0.01 \tag{3.8}$$

$$\Longrightarrow \theta = \sin^{-1}(0.01) \tag{3.9}$$

$$= 0.57^\circ \tag{3.10}$$

This is still a small angle, but if the screen is far enough from the slit, it would make a noticeable pattern.

What if it is not a slit, but rather some other shape, a circular hole for example? The same basic principles apply, but the detailed formula is different. In the case of a circular hole, the diffraction pattern is something like a bullseye; see the right side of figure 3.20. The half width of the central spot is given by a formula almost the same as equation (3.6); it differs by only a factor of 1.22.

A bunch of tiny circular dots, all the same size, will also give essentially the same pattern as a circular hole, although with less contrast. Figure 3.21 shows the diffraction patterns made by shining lasers through smears of red blood cells. Clearly, the angle of diffraction is relatively large, and so the blood cells must be small indeed.

Another general feature of diffraction, evident from figure 3.21, is that the size of the diffraction pattern depends not only on the size of the obstacle, but also on the wavelength of the light. Equation (3.6) implies that a longer wavelength of light, all else being equal, will produce a larger diffraction pattern. Equation (3.6) is for the special case of a single slit, but the same rule holds true for any diffraction pattern.

Figure 3.21. Left: lasers of three different wavelengths shine through smears of human red blood cells. The diffraction pattern is large because the cells are so tiny. The central portion of the bulls-eye pattern is smallest for the violet (short wavelength) light and largest for the red (long wavelength) light. Right: two lasers of the same wavelength shine through different smears of red blood cells. The diffraction pattern on the left appears slightly larger because the smear was made with red blood cells from a dog, and they are slightly *smaller* than the human red blood cells used on the right.

3.8 Fluorescence

Some materials can absorb the energy from light of a particular wavelength, and then use that energy to re-emit light of a completely different wavelength, in a process called *fluorescence*. Most commonly, light of violet and near-ultraviolet wavelengths are absorbed, and then light of wavelengths closer to the middle of the visible spectrum is emitted.

Sunlight contains a fair amount of violet and near-ultraviolet (UV) light, and so fluorescent dyes can make something appear *brighter than white*. Unbleached paper looks duller than new white copy paper because the latter has a fluorescent dye that absorbs some of the energy of the non-visible UV, and uses this energy to re-emit visible light. This mostly-bluish added light makes the color whiter overall, but it also makes it look brighter. The non-fluorescent unbleached paper can only reflect as much visible light as falls on it. The fluorescent paper does this too, but then adds its own visible light, the extra energy coming from absorbed UV.

3.9 Polarization

There is an additional aspect to electromagnetic waves, apart from the four properties—amplitude, wavelength, frequency and speed—discussed in chapter 2, section 2.3. In an electromagnetic wave, the electric and magnetic fields point perpendicular to each other, and the wave travels in a direction perpendicular to both. But this leaves open an infinite number of possibilities. For example, if an electromagnetic wave is traveling directly upward, then the electric field must point perpendicular to that—i.e., horizontally. But does the electric field point horizontally north, south, east, west or one of the infinite possibilities in between? Any one particular choice of orientation of the electric and magnetic fields is called a *polarization* of the wave.

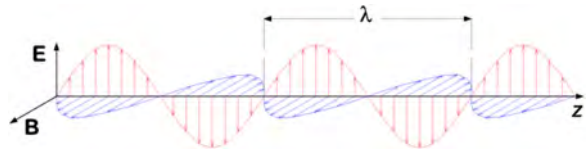

Figure 3.22. A polarized electromagnetic wave traveling from left to right. The electric and magnetic fields (marked **E** and **B**) in a polarized wave maintain a particular orientation. In an unpolarized wave the fields have random orientations perpendicular to the direction of travel (and to each other) (graphic: P.wormer - Own work, CC BY 3.0).

For most sources of light, countless individual atoms and molecules are each producing light in their own fashion, with little coordination between them. Thus many waves are emitted each with its own essentially random polarization. The overall effect is that *all* polarizations are present at the same time. This is called *unpolarized* light, and it is the light one gets from, say, an ordinary light bulb. If, on the other hand, a source of light contains mostly only one polarization, we say the light is *polarized*, as in figure 3.22. In practice, this is a more-or-less thing, rather than either-or, and so we find ourselves often describing light as *weakly polarized* or *strongly polarized*. This just means that, while all possible polarizations are present, one of them is more strongly represented than the rest.

There are many ways in which light interacts with matter, for which polarization is a crucial factor. A good example is the Rayleigh scattering discussed in section 3.5. Not only does the amount of scattered light depend on the wavelength, it also depends on the polarization. And so, for a given scattered angle, some polarizations scatter more than others. For Rayleigh scattering in the Earth's atmosphere, it turns out that light scattered by an angle of 90° is strongly polarized. Why? Because only one particular polarization scatters well at this angle. On a clear, cloudless day of low humidity, the light of the sky comes mostly from Rayleigh scattering. And so the part of the sky, making a circle around the Sun, that is everywhere 90° from the Sun is thus strongly polarized.

Reflections provide another example. Recall from section 3.1 that a certain percentage of a light wave will reflect when it encounters a sudden change in transparent material, such as when it encounters a piece of glass or the surface of water. This reflected wave may be strongly polarized, depending on the angle of the reflection. At a particular angle, called the Brewster angle, the reflected light is completely polarized.

You don't directly see the effects of polarization because, to the human eye, one polarization looks just the same as any other. But there are clever and important ways to distinguish between them with special filters, and we will take up this topic again in volume 2 of *The Physics and Art of Photography*.

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Chapter 4

Sources of light

4.1 Light and its spectrum

There are many ways in which light can be created by matter. Some of these processes are best understood from within the particle model of light, while others are more easily explained in terms of waves. Nonetheless, any particular method for producing light is essentially an interaction between light and matter, and the wavelength (or frequency) of the light is of crucial importance.

For any method of producing light will inevitably not create all wavelengths equally. If by 'light' we mean the entire electromagnetic spectrum, then this fact is true in spades. For example, to make electromagnetic waves of a frequency of one million Hertz, simply use an electronic oscillator circuit to make a current go back and forth in a wire at that frequency. A frequency of one million Hertz is in the radio part of the electromagnetic spectrum, and so this is essentially the basis of a radio transmitter.

But for visible light with frequencies several powers of ten higher, this strategy simply will not work. Instead, however, one can just heat up a tungsten wire to a high temperature, and the individual tungsten atoms will vibrate at high frequency and visible light (and infrared too) will be emitted.

The point is that any particular method of producing light will make some wavelengths well and other wavelengths poorly, or not at all. And so to really describe a particular source of light, we need to describe how much of each wavelength has been produced. Such a description is called a *spectrum* (plural, spectra) of the light source, and the most productive way to represent it is with a graph.

Figure 4.1 shows the spectra of the two stars that make up the binary stars system Albireo, also called β (beta) Cygni. The horizontal axis of the graph is wavelength, with short wavelength on the left and long wavelength on the right. The axis goes from about 4200 Å to 6000 Å (420–600 nm), and this is within the range of wavelengths sensitive to the human eye. The spectrum of one of the stars is marked in red while the other in blue. Clearly, β Cygni A emits more long-wavelength (red)

doi:10.1088/2053-2571/aae1b6ch4

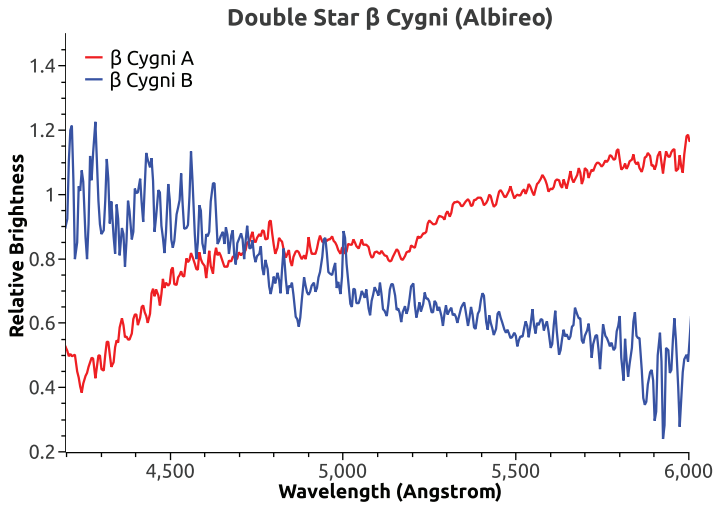

Figure 4.1. The spectra of two stars making up the binary star system β Cygni. Each colored line represents the spectrum of one of the stars. The fine-scale squiggles result from noise (random measurement uncertainty), but the overall trends in the spectra are accurate. The spectra span the range from 4200 Å to 6000 Å (420 nm to 600 nm), thus covering the range from blue-violet on the left to orange-red on the right. β Cygni A is yellow in color, and so emits more long-wavelength light than short-wavelength light. β Cygni B on the other hand, is blue, and emits more short-wavelength light than long (data from Beaver 2012).

light than short-wavelength light (blue), while the opposite is true for β Cygni B. Thus one can see from this graph that they would not appear the same color overall; β Cygni A appears yellowish to the eye, while β Cygni B appears more bluish. We will use this idea of the spectrum of a light source many more times throughout this book; it is one of the most important tools for understanding light. In the case of β Cygni and other stars, for example, astrophysicists can determine many things about their physical natures by analyzing the spectra of the light they emit.

4.2 Thermal radiation

The term *thermal radiation* refers to electromagnetic waves emitted due to the normal motions of atoms and molecules in some material. Whether it be liquid, solid or gas, the atoms and molecules that make it up are constantly in motion. Although the individual particles are moving with a wide range of energies (called kinetic energy), there is in many circumstances a meaningful average. This average of the kinetic energies of the individual particles in a material is directly related to what we call *temperature*. Thus the atoms and molecules making up a hot potato are moving, on average, with greater energy than are those of a frozen potato.

These motions of the individual atoms and molecules generate electromagnetic waves, and the waves produced have a wide range of frequencies, just as the atoms and molecules have a wide range of energies. Thus a *spectrum* of frequencies (or wavelengths) is produced. If all of the individual atoms and molecules are in perfect balance with each other—and with the light they produce—then a particular kind of spectrum is produced called a *blackbody spectrum*. That perfect balance of matter

and light, called *thermodynamic equilibrium*, ensures that *the details of the blackbody spectrum depend only on temperature*.

Figure 4.2 shows three different blackbody spectra, with temperatures of 3000 K, 4000 K and 5000 K, where the 'K' stands for kelvin, our SI unit of absolute temperature. The kelvin scale starts at absolute zero, and room temperature is roughly 300 K. The surface temperature of the Sun is 5770 K, so the temperatures represented in figure 4.2 are very high indeed. Nearly all of the chemical elements are vaporized at 5000 K.

The graph is silent about what *type* of material made these spectra for a very simple reason—it doesn't matter. A 5000 K blackbody is a 5000 K blackbody, whether it is produced from calcium, carbon, cobalt or vaporized chocolate-chip cookies.

The overall shape of a blackbody spectrum is that of a broad hump, with a peak at some particular wavelength. There is a particular mathematical formula that describes the exact shape of a blackbody spectrum, but there are two basic features that are obvious from the graph alone. First, the hump is bigger—which means more energy is emitted—if the temperature is higher. This is a huge effect, as the brightness is related to the total *area* under the graph. Notice how much larger is

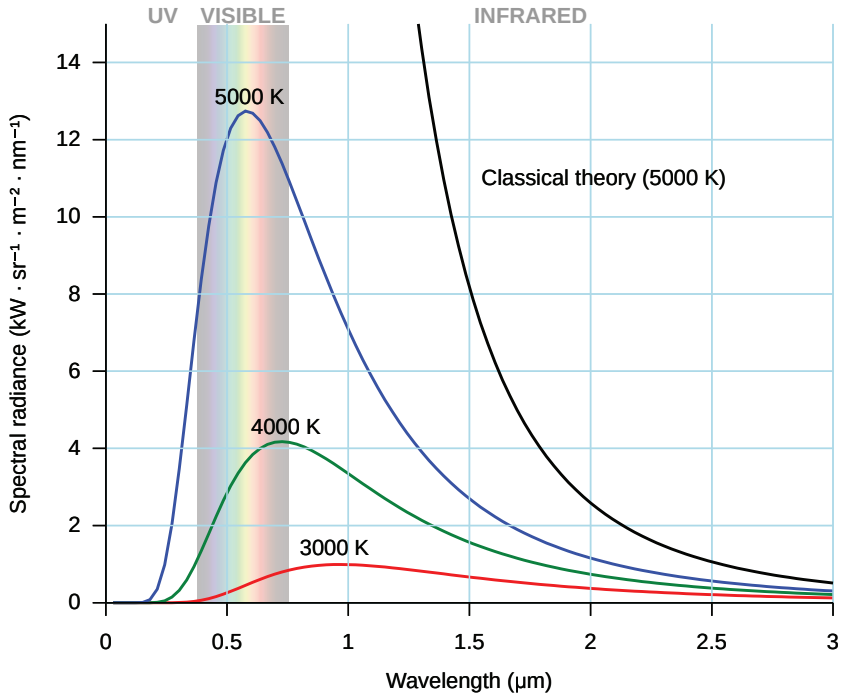

Figure 4.2. The blackbody spectra for several different temperatures. Blackbody radiation is the most ideal form of thermal radiation. The range of wavelengths corresponding to the visible spectrum is highlighted in rainbow colors. The curve marked 5000 K is similar to the spectrum of the Sun (5770 K would be closer). Note that higher-temperature objects are brighter and radiate at predominately shorter wavelengths. The curve marked 'classical theory' shows how badly theory fit the data before the development of quantum physics.

Table 4.1. Peak wavelengths for different temperature blackbody spectra.

Temperature (K)	Peak wavelength	Part of E–M spectrum	Example
2.7 K	1100 μm	Microwave	Cosmic microwave background
300 K	10 μm	Thermal infrared	Room temperature objects
3000 K	1 μm	Near infrared	Light bulb filament
6000 K	0.5 μm	Visible light	Visible surface of Sun
10 000 K	0.3 μm	Near ultraviolet	Visible surface of Vega

the area under the 5000 K graph as compared to the 3000 K graph. It turns out that the brightness of a blackbody scales with the fourth power of the temperature. And so a doubling of the temperature increases the brightness by $2^4 = 16$ times.

It is also clear from figure 4.2 that a *higher* temperature produces a blackbody peak at a *shorter* wavelength. This is described by a simple inverse proportionality. As an example, we can see from the graph that a 3000 K blackbody 'peaks' at about 1 μm (1000 nm). And so a room-temperature 300 K blackbody, with one tenth the temperature, peaks at ten times that wavelength, or 10 μm, in a region of the spectrum known as the *thermal infrared*. Table 4.1 shows the peak wavelength for different temperatures.

A blackbody the temperature of the Sun, a little less than 6000 K, peaks right in the middle of the visible spectrum. The part of the Sun's atmosphere that makes the visible surface of the Sun is not in perfect equilibrium (if it were, the light could not escape). And so the visible spectrum of the Sun is not quite the same as a blackbody—but it is a rough approximation of one, with the same overall shape. Notice that the light from an ordinary light bulb is on average of much longer wavelengths than that of the Sun. This has important implications for photography that we will discuss in more detail in volume 2 of *The Physics and Art of Photography*.

Figure 4.2 also shows the spectrum of a 5000 K blackbody as predicted by the electromagnetic theory of the late 19th century, and it is a catastrophic failure of agreement between theory and experiment. At the dawn of the 20th century Max Planck demonstrated that the observed shape of the blackbody spectrum could be explained only if one assumed that energy came in discrete clumps, or *quanta*. This *Planck quantum hypothesis* was the first step in the development of quantum physics, and the idea was extended by Einstein and others to form the modern concept of the photon—a particle of light. And so although we observe the spectrum emitted by a blackbody as a wave phenomenon, the fact that light can act as individual particles is essential to its creation.

4.3 Non-thermal radiation

Thermal motion of atoms and molecules is not the only way to produce light, and so not all sources of light emit with a spectrum approximately like that of a blackbody. A given laser pointer, for example, emits only one wavelength of light, not a broad

range like a thermal source. And there is no simple connection between temperature and the wavelength of the laser. An ordinary fluorescent light is another good example. Mercury vapor is excited by a high voltage, and the individual mercury atoms in the gas emit light of only very specific wavelengths, characteristic of mercury gas. This produces an *emission-line spectrum*. Most wavelengths are devoid of light altogether, while at certain very specific wavelengths, a lot of light is emitted.

Some of the brightest *emission lines* in the spectrum of mercury are in the ultraviolet part of the spectrum, which wouldn't do us any good if we wanted to use this as a source of light to see by. But in a fluorescent lamp, this light is reprocessed. Before it gets out of the lamp, it strikes a powder of small crystals, called phosphors, that line the inside of the glass lamp. These crystals have the ability to absorb ultraviolet light and use its energy to emit its own visible light. Since the phosphors emit a range of many wavelengths after excited by ultraviolet light, the overall effect is that the invisible ultraviolet light is converted to visible light. If phosphors with the right properties are chosen, the overall effect is a whitish light. But it is not produced by high temperatures, and the spectrum is very different from that of sunlight.

Figure 4.3 shows the spectrum of a typical fluorescent light. One can see the Mercury emission spectrum poking through, but superimposed on this is the broad spectrum of the glowing phosphors. Although the overall perception by the eye of fluorescent light is roughly similar to that of white light, it is a very different spectrum. Thus we will need to take this into account when taking photographs with fluorescent light.

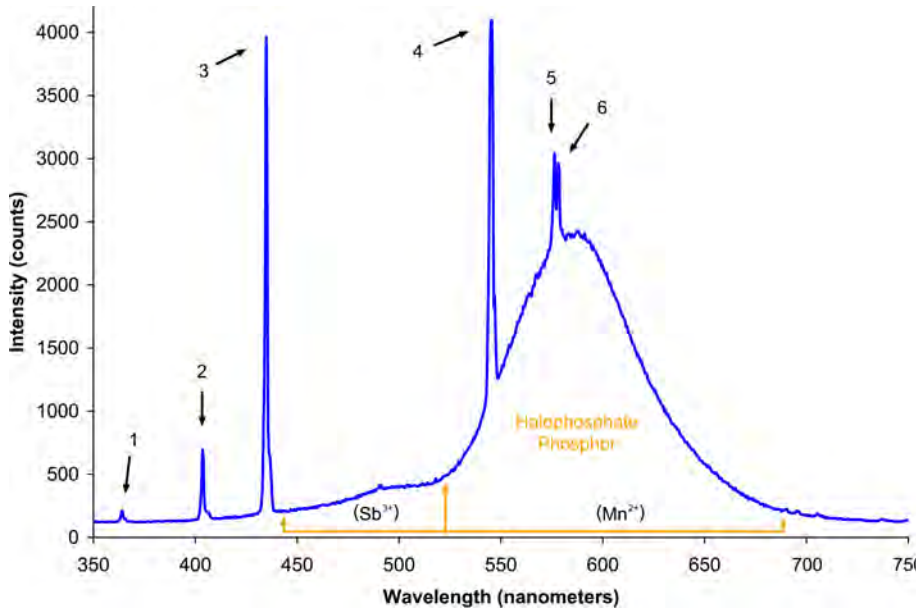

Figure 4.3. The spectrum of a fluorescent tube light. The numbered peaks are emission lines from the spectrum of mercury. The smooth parts of the spectrum is due to the glowing phosphors excited by the ultraviolet emission lines in the Mercury spectrum.

Non-thermal sources of light such as lasers and emission-line sources produce light at the atomic level. That is, each atom produces its own light, emitting individual photons. And so the particle-like nature of light must be considered to understand these sources. They are one-on-one interactions between light (in the form of individual photons) and matter (in the form of individual atoms). As such, the spectrum of the light produced *depends critically on the type of atoms emitting the light*. This is the opposite of the case of thermal radiation. *For purely thermal radiation, only the temperature matters*; the type of atom is irrelevant. Many real sources of light, when considered in more detail, involve a combination of both thermal and non-thermal emission.

Reference

Beaver J and Conger C 2012 Extremely low-cost point-source spectrophotometry (ELCPSS) *Society for Astronomical Sciences 31st Annual Symp. on Telescope Science* 113–20

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Chapter 5

Wavelength reconsidered

We have described many different ways in which matter can both create light and also alter light that interacts with it: absorption, reflection (diffuse and specular), refraction, scattering and emission (thermal and non-thermal). We have seen how emission processes produce a spectrum of light—different amounts of light emitted at different wavelengths. But *all* of these processes depend critically on the wavelength of the light. And this means that for one particular region of the electromagnetic spectrum (visible light, for example), one or more of these processes may be important while others are negligible. But at some other region of the electromagnetic spectrum (infrared, for example) the situation could be entirely different.

Let us reconsider, for example, Boris the cat in figure 3.12. In the visible part of the spectrum, we see Boris by diffuse reflection of light that has arrived at him, but was created by other means. In this particular case, Boris was illuminated by an incandescent light bulb—a white-hot tungsten filament that *emitted* light by means of (mostly) thermal radiation.

The light emitted by the light bulb and arriving at Boris had a spectrum similar to the red curve in figure 4.2; it was mostly visible light (400–700 nm, or 0.4 μ–0.7 μ) and the infrared part of the spectrum right next to visible light, what is sometimes called the *near infrared* (700 nm to about 5000 nm, or 0.7 μ to 5 μ). In this case, the film that took the picture ignored whatever near-infrared light reflected from Boris, only recording the little bit that was in the visible part of the spectrum. And so the picture shows visible light, emitted by a light bulb and reflecting diffusely off of Boris.

Every part of Boris also *absorbed* a portion of the visible light that arrived there. But his stripes absorbed a little more, and so they reflected a little less, and thus appear slightly darker in the picture.

Figure 3.12 shows only what happened to *visible light*; it says nothing about other wavelengths. And so what about that near-infrared light that also illuminated Boris, but was not detected by the film and so is not represented in the picture? What would have happened if we had instead taken the picture with a camera that is sensitive not

doi:10.1088/2053-2571/aae1b6ch5

to visible light, but instead only to near-infrared light? Would the picture have been different?

See figure 5.1 where I have photographed a different cat[1], Tobias, with two different cameras, one sensitive to visible light (left) and the other sensitive to near-infrared light (right). Just because the striped part of the fur absorbs more (and so reflects less) visible light, it does not follow that it must also do so for near-infrared light.

For both visible light and near-infrared light, we see what is reflected by Tobias; the source of the light is elsewhere. But for *thermal infrared* light, Tobias *is* the source. In figure 5.2 I photographed him with an ordinary visible-light digital camera on the left. But the right-hand image was made with a *thermal imaging camera*, sensitive mostly between the wavelength range 8–14 μm.

Figure 5.1. Tobias the cat illuminated by sunlight, and photographed with, left: visible light (0.4–0.7 μm). Right: near-infrared light (0.7–1 μm). The stripes of Tobias absorb more (and so reflect less) visible light than the surrounding fur, but not so for near-infrared light.

Figure 5.2. Tobias illuminated by daylight, and photographed with, left: visible light. Right: thermal-infrared light. The visible-light image depicts how well Tobias reflects or absorbs visible light (0.4–0.7 μm). But the thermal-infrared image shows mostly how well Tobias (and his surroundings) *emit* light in the wavelength range 8–14 μm. The window glass is transparent to visible light, but it blocks thermal infrared light from passing through, while reflecting some off its surface like a mirror.

[1] Rest in peace, Boris, under the peonies!

Thermal infrared (TIR) is called such because it is the range of wavelengths *emitted by* objects at roughly room temperature. But the amount emitted depends strongly on *temperature*, and since Tobias is warmer than his surroundings, he emits more thermal infrared than anything in the picture.

The window glass is transparent to visible light, but it is opaque to TIR light; it does, however, reflect a significant fraction. Thus none of the outdoors for which Tobias pines is visible in the TIR view, and it looks as though he is sitting in front of a mirror admiring his own handsome reflection.

The TIR image looks almost as though it were a negative, but it is not; the brightest parts of the picture are those that emitted the most TIR light. The gaps and crevices in Tobias's fur reveal areas that are shadowed from reflected visible daylight, and so they appear dark in the visible-light picture. But they also reveal the skin of Tobias, which is much warmer than the outer parts of his fur. And so those are some of the brightest areas in the TIR image.

Part III

Geometry and two-dimensional design

IOP Concise Physics

Chapter 6

Geometry and the picture plane

The history of geometrical optics is tied intimately to the history of the visual art that pre-dates photography. Look at paintings from the middle ages and they are unlike photographs in many ways. For example, the sizes of forms in the paintings were not necessarily connected in a representational manner, as they are in a typical photograph, to the sizes and distances of objects in the real world. Instead, the size and positioning of forms in a medieval painting often had more to do with what was more or less important (see for example Duby, 1992, chapter 1). It was not until the Renaissance that painters really worked out the details of *linear perspective*, allowing one to duplicate a scene on canvas akin to what would later appear in photographs (Fichner-Rathnus, 1992, pp 333–4).

Photography clearly has much to do with the connection between the three-dimensional (3D) world and two-dimensional (2D) art, for at its most basic level a camera takes rays of light from the 3D world and redirects them to a 2D surface. But the 'fourth dimension' of time is also important. It is now common to see photography as an act of freezing a moment of time, and photographers often take advantage of this. See, for example, the disturbing use of this power in *Execution of Vietcong Prisoner* by Eddie Adams.

A case can be made that the idea that an instantaneous moment of life can be captured with paint on canvas came into its own with the Impressionists, who were very much influenced by then-new photography. In volume 3 of *The Physics and Art of Photography* we revisit this question of the relation between painting and photography.

6.1 From 3D to 2D

Part of the essence of photography is that elements of the inherently 3D world are directed onto a 2D surface. At first glance, one might think that, as with a camera, this is exactly what the human eye does. Rays of light from the world are focused by

doi:10.1088/2053-2571/aae1b6ch6

the eye's lens and cornea onto specific locations on the two-dimensional surface of the retina. But it is much more complex than that.

Part of the reason that 2D art works at all is that our brains seem to be specifically structured such that we can easily attach worldly, 3D meaning to patterns on a flat surface. Presumably, this is because the brain must make sense of the essentially 2D information coming from the eye. How the eye/brain does this is extraordinarily complex and only partially understood, but much of the detail has little direct counterpart in the operation of a camera.

For one thing, the image from a camera (or eye) lens literally represents only directions in space, not distances. An *image* is only a 2D representation, and yet we construct a 3D universe out of that image.

This fact is particularly (and painfully) obvious to an astronomer. Take a picture of the night sky, and one sees a pattern of bright dots—stars. But there are an infinite number of possible 3D arrangements of those stars consistent with that same picture. To determine that third dimension of distance requires much additional information that is not available in the photograph alone, and this task is a central preoccupation of astronomers.

But the human brain, in its second-by-second workings, doesn't really operate in this considered, mathematical way. The brain adds its own interpretation to the mix of data coming from the eye; it makes stuff up, in a sense. And so it is possible to take advantage of this fact and trick the brain; the phenomena of *optical illusions* provides evidence for this, and clues to the actual mechanisms at work.

6.2 The human brain's construction of three-dimensional reality

Experiments have shown that the image on the retina of the human eye is very different from what we actually perceive as sight. The human eye in its construction has much in common with a camera; a lens focuses an image onto a surface (the retina) that is sensitive to light. But the similarity ends there. Much is sometimes made of the fact that the image on the retina is, like in a camera, upside down; but that is not such a big deal. Most of 'seeing' is in the brain, not in the eye. Flipping the image right-side up is the least of it.

Some of the research carried out in the 1950s and 1960s by the Russian psychologist Alfred Yarbus provides a good example (Yarbus 1967). Yarbus recorded the eye movements of observers as they performed certain tasks. He found that for even something so simple as looking at a motionless face, the human eye darts around all over the place; see figure 6.1 for an example. It seems that we tend to look at the eyes and mouth a lot, as can be seen from the eye movement trajectories he recorded. If one pointed a video camera at the same face, and moved it according to such a trajectory, the resulting video would likely be unintelligible.

Yet somehow the brain makes sense of this seemingly confusing information coming from the eye. Clearly the brain *constructs* an image from this information, rather than simply recording it. And that image is of a 3D world with objects in it. And what we 'see' not only ignores much of the irrelevant information coming from the retina, it also includes elements not even present in that stimulus; the brain

Figure 6.1. An illustration tracking the eye movements of a subject looking at a photograph of a face, work pioneered by the Russian psychologist Alfred Yarbus. Even when looking at something so simple, the human eye is constantly darting about. Yet we perceive a motionless image (graphic: SpoonSpa, Simon Viktória, CC BY 2.0 Generic).

makes stuff up. And this means that *we can attach meaning to even the simplest marks on a flat surface*. For painters and photographers, who make images on a flat surface, this is a lucky break! A photograph is just a flat piece of paper with marks on it. But we can see the world when we look at a good one.

6.3 Linear perspective and the *Camera Obscura*

Which came first, the camera or the light sensitive material that goes in the camera to make the photograph? It is a surprise to most people that the correct answer is the camera, which preceded the invention of photo-sensitive materials by a couple of hundred years. But when one realizes that the word *camera* is simply the Latin word for an enclosure, then maybe it is not so strange. Long before cameras were used to record images, they were used to view them. The *camera obscura* is simply a 'dark box.'

It has been known for centuries that one can use geometry to produce images. Place a small hole in one side of a dark box, and each ray of light coming from objects in the world will be restricted by the hole to only one spot on the opposite side of the box. See figure 8.1. Thus, an image of the outside world is automatically reconstructed on the inside of the box. This basic idea underlies both the ancient camera obscura, and the modern pinhole camera we discuss in detail in chapter 8.

A glass lens allows one to (among other things) brighten this otherwise dim image, and this too has been known for centuries, at least since the 1500s. When light-sensitive materials were invented in the 1800s, the already-existing camera obscura developed into the photographic camera (Marien, 2002, pp 3–7).

One simple way to view the image in a camera obscura is to make the dark box very large, and put yourself inside. The photographer Abelardo Morell has been doing this with great success. He uses paint to black out the windows of a room with a view, and then scratches a small hole in the paint. An upside down image of the outside world appears on the opposite wall. This is, in effect, a giant pinhole camera with the photographer inside. Very little light enters through the small hole, and so the image is very dim—barely visible to the naked eye. He photographs this image with a separate camera, set up inside the room, using a long time exposure to make up for the dim light.

A not-very-impressive example of this same technique is shown in figure 6.2. The darkened room of my physics classroom was photographed with a long time exposure. The image on the wall is from a small hole in the blinds opposite. The photograph has been inverted so the camera obscura image, of the grounds outside the lab, is upright.

The camera obscura and related optical and sighting devices were instrumental in the discovery of mathematical rules for laying out a realistic perspective in paintings, what art historians call *linear* or *mathematical perspective*. Some have speculated (the thesis is still controversial) that some of the Flemmish painters of the 17th century (Vermeer in particular) used a camera obscura to properly locate the details in their paintings (Steadman 2001). In 15th century Italy, Filippo Brunelleschi performed experiments with a sighting device much like a camera obscura in order to work out the first formal methods for incorporating precise linear perspective into paintings and drawings.

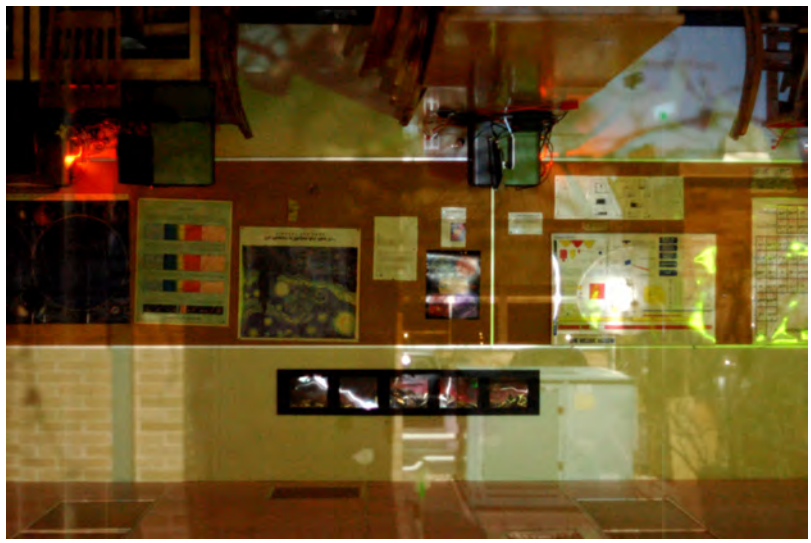

Figure 6.2. My physics classroom photographed from the inside. A small hole in the blinds produced the image of what is outside the lab. The picture has been turned upside down, and so the room appears inverted but the camera obscura image is upright.

6.4 The picture plane

Photographs are often *representational*; forms in the photograph are meant to directly represent things in the real world. But as is the case for all 2D art, there is more to it than that. An important element is the *picture plane–the flat plane representing the 2D surface of the photograph itself*. While this concept of the picture plane *as a physical surface* is of crucial importance for painting, it is sometimes downplayed or ignored in photography, which often strives to be purely representational.

This can be seen in the way many photographs are displayed. A mat is used to overlap the image area of the print. And so the opening in the mat appears as a window, with the picture behind. The intent seems to be to hide the fact that one is looking at a piece of paper; only the representation of the image matters.

Many photographers however, aligning themselves somewhat more with the history of painting and the early history of photography, go out of their way to draw attention to the picture plane. One method is to use alternative printing processes that introduce their own elements in the form of random detail and textures (Rexer 2002). We will explore some of these printing processes in more detail in volume 3 of *The Physics and Art of Photography*. Often the goal here is similar to that of traditional art printmaking. The image is the same, but each print is slightly different, and thus unique, due to the handmade process. And the more-interesting and less-uniform surface of the paper itself is often evident. And so it is not just an image; it is an image on a flat piece of interesting paper, and the paper is the picture plane.

But the negative process of film photography allows for another interpretation of the picture plane. One does not normally view a photographic negative directly; rather one looks at a positive print made from it. *But the negative also represents a flat plane, and it is what was actually inside the camera at the moment the photograph was taken*.

Thus in negative-based film photography, we actually have a possibility for two picture planes: the surface of the print itself, and the surface of the negative that was used to make the print. Details in each can draw attention to these two flat surfaces, forming an additional element in the formal construction of the picture. This is an especially important point in the modern age, where many pictures never exist as a physical print. Thus the picture may be an image on a large computer screen or on a tiny smart phone, depending on who is looking at the picture. Where then is the picture plane? Is the concept still meaningful when no *physical* image surface exists?

Figure 6.3 shows an image made from an old type of Polaroid large-format negative film. This type of film made its own instant print and a high-quality instant negative, that could then be printed in the darkroom. No darkroom was needed to produce the negative itself; one just needed to clear off the developer goo, in daylight in the field, with a simple sodium sulfite solution (often carried around by the photographer in a bucket).

In this case, I allowed the goo to remain on the negative, and in fact rubbed it in the dirt to make it even more yucky. I then allowed the negative to dry without

Figure 6.3. *Proof of Life on Earth*, John Beaver 2008. This photograph was scanned from an intentionally-damaged negative, thus making evident the flat surface of the negative itself. This surface detail introduces a picture plane *into the image itself*, independently of how it is displayed.

washing it. The result was a bad negative that would have been difficult, if not impossible, to print in a traditional darkroom. But it was possible to scan it with a high-quality scanner, and then use digital techniques to reverse it to a printable positive image.

The result is that the surface of the negative is introduced as a picture plane, independent of the manner in which the image is printed or displayed. In this particular case the flat plane of the negative plays off of the seemingly-curved plane of the table introduced by the distortion and vignetting of the particular lens used.

In the chapters that follow we explore this 2D nature of photography, and how it relates to our 3D world. We will see that even a virtual image that never exists as a physical object still has abstract relations between the parts of its 2D 'surface.' In volume 3 of *The Physics and Art of Photography* we revisit some of these ideas in consideration of the physical processes by which light from the world interacts with, and changes, a flat surface in order to form an image. And we explore some methods by which nature can have its own say in the making of photographic art, in ways that are only partially under the control of the artist.

References

Duby G 1992 *Medieval Art: Europe of the Cathedrals 1140-1280* (Paris, France: Bookking International)

Fichner-Rathnus L 1992 *Understanding Art* 3rd edn (Englewood Cliffs, NJ: Prentice Hall)

Marien M W 2002 *Photography: A Cultural History* (Englewood Cliffs, NJ: Prentice Hall)
Rexer L 2002 *Photography as Antiquarian Avant-Garde: The New Wave in Old Processes* (New York: Harry N. Abrams)
Steadman P 2001 *Vermeer as Camera: Uncovering the Truth Behind the Masterpieces* (Oxford: Oxford University Press)
Yarbus A L 1967 *Eye Movements and Vision* (New York: Plenum)

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Chapter 7

Light and shadow: photograms

A *photogram* is a camera-less photograph. An object is placed in direct contact with some light-sensitive material and then exposed to light, directly recording the object's shadow. Many of the earliest photographs were photograms, for the simple reason that the earliest light sensitive materials were of such low sensitivity that their use in a camera was impractical. Some of the first photographs using light-sensitive silver salts on paper, by William Henry Fox Talbot, were photograms of botanicals; he called them 'photogenic drawings.' And the first book to be illustrated with photographs was *Photographs of British Algae: Cyanotype Impressions* by the botanist and photographer Anna Atkins. She used John Herschel's newly-invented *cyanotype* process to illustrate her 1843 book with direct photograms of the specimens.

Since a photogram is essentially a shadow, it is important to understand some of the basics of how shadows are formed:

1. A shadows is larger than the object casting it if the source of light is very near. The object and its shadow are roughly the same size if the source of light is distant.
2. A large light source produces a shadow with fuzzy edges. A small light source produces a shadow with sharp edges.
3. The shadow of an object closer to a screen is sharper-edged than the shadow of the same object placed more distant from the screen.
4. Multiple light sources produce multiple shadows. But depending on the geometry, light from one source can partially fill in the shadow from another light source.
5. The Sun in a clear blue sky produces shadows that are partially filled in by the blue light of the sky, contrasting with the slightly yellowish (in comparison) non-shadow area from the direct sunlight. Thus we have the well-known rule that on a clear day, shadows are bluish in color.

doi:10.1088/2053-2571/aae1b6ch7

6. The shape of a shadow is distorted by the geometry of the projection angle onto the surface the shadow is cast upon.

With these points in mind, carefully examine the cyanotype photogram *Paloverde* in figure 7.1, where I have digitally reversed the original blue-tone negative to a positive. Some of the branches appear very sharp, while others are blurry. This gives the photogram a look of depth, as if it were a photograph taken with a camera using a shallow depth of focus. But this was simply a piece of light-sensitive paper held next to a tree branch while it was exposed by the Sun; no camera was involved. The branches touching the paper made sharp-edged shadows, while the branches in front of and further from the paper made fuzzier shadows.

Early photographic detectors were far less sensitive to light than those of the modern era, and so they lent themselves naturally to the making of photograms. But many contemporary photographers have found old photogram techniques to be useful tools for making new art, and so there is now a strong revival of these old processes (Rexer 2002). We explore these issues in some detail in volume 3 of *The Physics and Art of Photography*, but in appendix A of this volume, I describe how to make your own photograms, with both the traditional cyanotype process and my own new/old 'ephemeral process' photography (Beaver 2017). Both techniques are inexpensive, accessible, and a lot of fun.

Figure 7.1. *Paloverde* (John Beaver, 2005). In this photogram, the shadows of the branches that were in direct contact with the light-sensitive paper appear sharp, while those that were some distance away produced blurry-edged shadows. The effect looks similar to the selective focus of a camera lens; but here no camera was used.

7.1 Shadows and the source of light

A photogram is, essentially, a capture of a shadow cast directly onto a light-sensitive material. And so to experiment with photograms is to explore the nature of shadows. For the right-hand image in figure 7.2, I used ephemeral process photography to directly record the shadow cast by the object shown in the left-hand image. When the shadow is seen in context, it is clear that it is illuminated by a single source of light to the upper right of the cordial glass. But the photogram itself seems distorted; the circular base looks exactly as it does on the object, but the rest looks very different.

The seeming distortion arises from the fact that the photogram is a *projection* of light rays from the source through the object. And so the shadow of the top of the glass is to the side of the shadow of the bottom of the glass. It is also evident that the shadow of the top of the glass seems a bit disproportionately large, compared to the proportions of the different parts of the object that cast the shadow. Since the bottom of the glass was in direct contact with the light sensitive paper, it appears the same shape and size, and in the same position, as the object itself. Not so, for the top of the glass, some inches above the surface of the paper.

Careful inspection of the photogram in figure 7.2 also reveals that the blurriest portion is from the top of the glass, the most distant from the paper, while the sharpest portion is from the base that was in direct contact. We have already seen that objects more distant from the shadow surface cast blurrier shadows.

Figure 7.3 shows four photograms, all made with the object in the same arrangement as in figure 7.2. But for each I have changed the light source. For the first and third photograms, the source of light was positioned relatively close to the object, while for the second and fourth the light was positioned far away. But for both the first and second examples I used a very tiny source of light (shining through a small hole), while for the third and fourth I used a light that had a large emitting surface area.

Figure 7.2. Left: a cordial glass illuminated by a single source of light located to the upper right. Right: a photogram made with this same illumination. The shape of the photogram is distorted by the angle of projection.

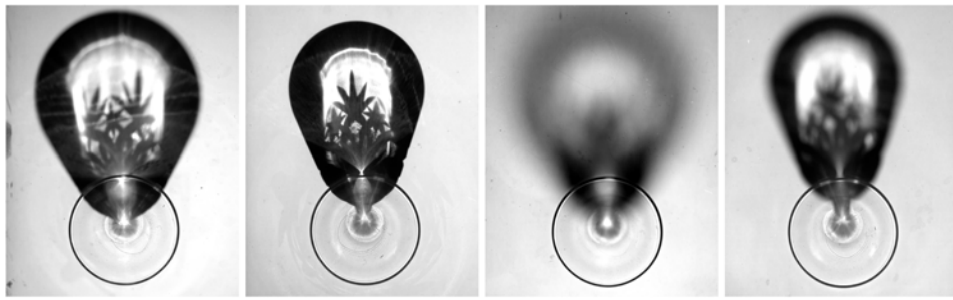

Figure 7.3. Four photograms of the same crystal cordial glass. From left to right: small light close up; small light far away; large light close up; large light far away. Notice that for all four examples the bottom of the glass (in direct contact with the light-sensitive paper) is identical. But the size of the top of the glass is more exaggerated for the images made with the close-up lights. Also notice that the shadows are sharper when the light source is either smaller or farther away (or both). And for all four photograms, the sharpest shadows are from the parts of the glass that were closest to the light-sensitive paper.

Let us first think about the relative distortions of size, and put the issue of sharpness or blurriness of the shadows aside for the moment. Comparing the first image to the second, we see that when the source of light is close by (as in the first image), objects more distant from the detector (and thus closer to the light source) appear larger than life. The same is clearly true when comparing the third and fourth images. As compared to the third image, the light source was moved farther away for the fourth image, and so the shadow of the top of the glass appears closer to its real proportion compared to the bottom.

For the first and third pictures, which show considerable distortion in the size of the top of the glass, the light source was only a few times further away than the distance between the top of the glass and the light-sensitive paper. And thus the light rays were spreading out significantly when they reached the nearby top of the glass. Following these same rays to the light-sensitive paper, they continue to spread, and so make an enlarged shadow of the top of the glass.

If the source of light is very far away, on the other hand, the light rays are nearly parallel, and so this effect is much less noticeable. Our most common source of light for outdoor photography, the Sun, is very far away indeed. And so for shadows cast by the Sun, the distortions in size illustrated by the first and third image in figure 7.3 do not occur.

The first and third images were both made with the light positioned at roughly the same distance. Yet one is very blurry while the other is very sharp. The same is true for the pair of the second and fourth images. Here the difference results from the *physical size* of the light source. For both pairs, the smaller light source produced the sharper shadow. But the *distance* of the light source also affects the sharpness of the shadow. The first and second shadows were both made by the same small light source; yet the second, made with the light source farther away, is clearly sharper. The same relation holds true for the third and fourth images. The same source of light moved farther away produces a sharper shadow.

A more distant light source produces a sharper shadow, while a physically larger light source produces a blurrier shadow. And so what if the light source is *both* larger *and* more distant? Will the shadow be blurrier or sharper? The answer is that it depends; we must look at the details for the particular case in question.

Consider again the four examples in figure 7.3. If we rank them from sharpest to blurriest, it is clear they would be in the order 2–1–4–3. Table 7.1 gives the relevant data for the four examples—the size of the light source and its distance from the light-sensitive paper—in order of increasing blurriness. By looking at each of the two variables individually, it is far from obvious why they would be in this particular order. But the fifth column of the table provides the answer. Here I have divided the distance to the light source by its size. It is clear that the smaller this ratio, the blurrier the shadow.

If we invert this ratio, instead dividing the size of the light source by its distance, we get numbers (fractions in this example) that are larger with increased blurriness, and smaller with increased sharpness (sixth column). This measure, the size of a light source divided by its distance, is closely related to the important concept of *angular size or angular diameter*. The angular size of an object is how large it *appears* as seen from a particular vantage point. The Sun is nearly a million miles across, but it has roughly the same angular size as a dime held at arms length. With this concept of angular size, we can now state the following simple rule for shadows:

> As seen from the point of view of the object casting the shadow, a source of light with a relatively small angular size will produce relatively sharp shadows, while a source of light with a relatively large angular size will produce comparatively blurry shadows.

Photographers often refer to sharp shadows as *hard* (especially if they are also of high contrast), while referring to blurry shadows as *soft* (especially if they are also of low contrast). Hard shadows may give a sense of drama to a photograph, while soft shadows tend to have a more relaxed feel. Portrait photographers often work in the controlled environment of a studio in part to exert detailed control over the hardness or softness of shadows cast by one part of a subject's face onto other parts. To

Table 7.1. In order of increasing blurriness from top to bottom, column one lists the four examples from figure 7.3. Columns 3 and 4 list the distance and size of the light source that illuminated the cordial glass to make these four photograms. From either of those columns alone, it is unclear why the rows of the table should be ordered in this way. But by dividing distance by size (or size by distance), one can see that neither the size nor the distance alone is the crucial issue, but rather *how they compare to each other*.

Example	Description	Distance (cm)	Size (cm)	Distance/Size	Size/Distance
2	Sharpest	132	0.64	206	0.0048
1	Less sharp	41	0.64	64	0.016
4	Blurrier	132	7.6	17	0.058
3	Blurriest	41	7.6	5.4	0.185

greatly soften shadows, a compact source of light may be enclosed in a large *soft box* that spreads the source of light over a large area.

7.2 Laser photograms

To make a photogram with perfectly sharp shadows, either the object must be thin and in direct contact with the light sensitive material, or the light source must appear, from the location of the photogram, as a only a tiny point. Since the Sun has a significant angular size, and so does not appear as only a tiny point, photograms of three-dimensional objects have fuzzy edges if sunlight is used. One could, in theory, make a mask that blocks all of the Sun except what comes through a tiny hole. But then the light would be too dim.

One solution I have discovered is to expose the photogram by 'scribbling' over it with a laser mounted from a fixed point. An example can be seen in figure 7.4. To do this I placed objects on light-sensitive paper, and patiently scribbled over it with the laser, exposing only one part at a time. The entire process took about 5 min.

But to produce such a sharp image in this way, the laser beam must come from exactly the same location the entire time. And so I mounted the laser on a tripod, but in such a way that it could pivot somewhat about the center of its light-emitting part. I could point the beam in different directions (within limits), and so expose any portion of the light-sensitive paper. But from the point of view of the light-sensitive paper, the beam always originated from exactly the same location. And thus the

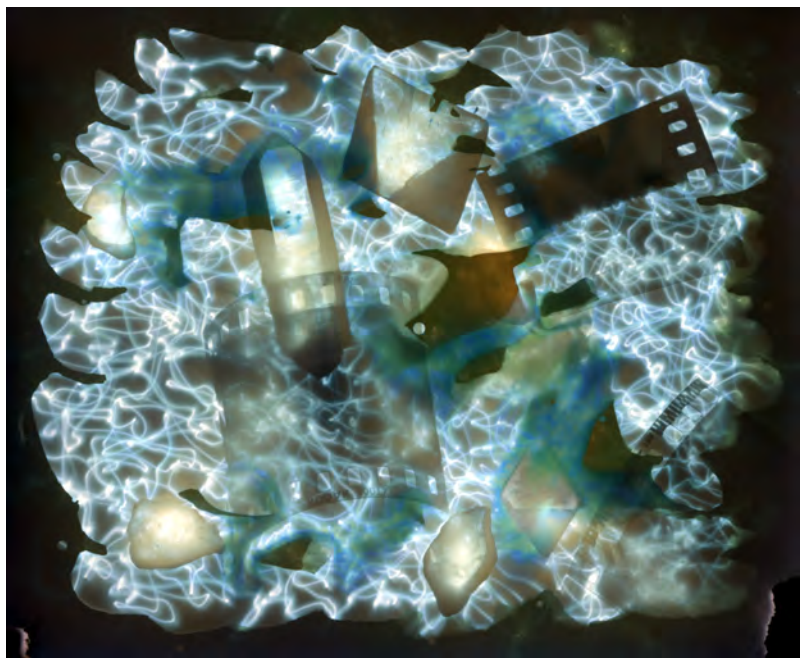

Figure 7.4. *The Six Cornered Snowflake*, John Beaver 2016. A laser swiveling about a fixed point was used to expose a photogram of three-dimensional crystals. Since the beam always came from the same fixed location, it was as if the source of light were a point source, and so the shadows are perfectly sharp.

effect, regarding the sharpness of the shadows, was the same as if the photogram had been illuminated with a point source of light.

Figure 7.4 was made using ephemeral process photography combined with an inexpensive deep-violet 405 nm laser pointer. The same laser can be used to make a photogram with cyanotype, although one must move the laser much more slowly (and so the process takes much longer). See appendix A for details on how to make your own photograms with cyanotype and ephemeral-process photography.

References

Beaver J 2017 Using new-antiquarian photographic processes to integrate art and science *AGU Fall Meeting Abstracts*

Rexer L 2002 *Photography's Antiquarian Avant-Garde: The New Wave in Old Processes* (New York: Harry N. Abrams)

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Chapter 8

Ray optics 1: pinhole photography

A lens uses refraction to redirect many light rays to the same place, and this allows for the creation of both bright and detailed images. We will consider lenses in more detail in chapter 9, but there is much about the formation of an image that has nothing to do with the lens *per se*. Rather it is about simple geometry. And so before we analyze lenses, it is helpful to consider the geometrical aspects of the simplest of all ways to record an image—the pinhole camera.

Let us consider that light undergoes diffuse reflection off of some object in the world—a cat, for example, such as in figure 3.12. The diffuse reflection means that every part of the cat reflects light in all directions at once. From the point of view of someone looking at the cat, or taking its picture, it is as if every point on the cat is *emitting* light in all directions. We know, of course, that the cat is only reflecting visible light, but if it *were* emitting its own light, a ray diagram of that process would be very similar.

We will adopt this approach as we go along in the next several sections. We will describe an *object* as if it is emitting, from each point on it, rays of light pointing in all directions. Of course, we can't literally draw all of these rays, as there would be an infinite number of them. Instead we choose to depict only those rays that, strategically, allow us to figure out what we want.

Consider for example rays of light coming from an object and then focused to an image by a lens. When analyzing this situation we gain nothing by depicting the plethora of light rays coming from the object that miss the lens all together. But in many circumstances, by being only a little clever, we will be able to disregard most of the rest of the rays as well, choosing only those that are necessary in order to determine whatever it is that we want to know. In fact, we will often be able to accomplish our task by considering *only two* out of the infinite number of rays coming from the object.

doi:10.1088/2053-2571/aae1b6ch8

A *pinhole camera* is essentially a dark box with a hole in it. A small hole in one side of a box is used to restrict light rays so as to form an image on the opposite side of the box. The image is then recorded with some light sensitive material.

One way to accomplish this is to use black and white enlarging paper, such as would be used for making prints in a darkroom, as the 'film.' While in the dark, place the paper on the inside of the box and put a piece of black tape over the pinhole. Then set the camera up outside on a tripod or stand of some kind, and uncover the hole. After enough time has passed (typically a minute or so), the print paper will be exposed. Cover the hole, carry the entire camera into the darkroom, remove the paper and develop it. A positive can be made from the negative print that results by placing it in contact with a second sheet of print paper, and shining diffuse light through the paper[1].

Figure 8.1 shows the geometry of a pinhole camera and why an image is formed. Since light can only enter through a small hole, only one ray from a given part of the object makes it into the camera. And then geometry does the rest; that ray can only go to one specific place on the light detector. And so each part of the object exposes its own place on the detector, and an image is formed. The image is, of course, upside down.

The left side of figure 8.2 shows the same diagram as figure 8.1, but with a larger hole. The big disadvantage of a pinhole camera is that the light enters only through a pinhole. This means the image is very dim, and requires either very bright light or very long exposures (or both). So one might very much want a larger hole for a brighter image. But clearly, a large hole means that rays from one point on the object go to a range of positions on the image. In other words, a larger pinhole should mean a blurrier image. So there is an inevitable trade-off for a pinhole camera; a brighter image means a blurrier image.

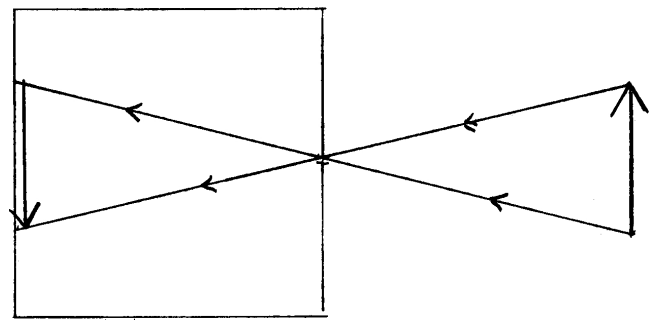

Figure 8.1. Formation of an image in a pinhole camera. Light rays reflect from the tip of the arrow in all directions, but only one can make it through the small hole in the pinhole camera. Thus rays from each part of the object are guided each to their own separate place to form an upside-down *image* of the object.

[1] There are many how-to guides available in print and online to fill in practical details for this process and its many variations, if you want to try it for yourself.

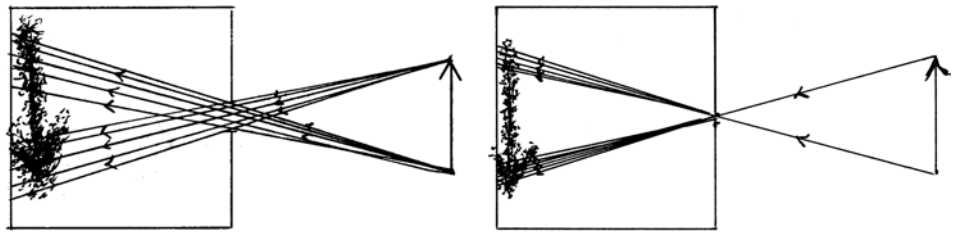

Figure 8.2. Left: a larger hole in a pinhole camera would let in more light and thus produce a brighter image. But the image would be blurry since many different rays from any given point on the object are allowed to strike many different locations on the detector at once. Right: if the hole is too small, the image in a pinhole camera will also be blurry; diffraction at the pinhole will spread the rays out on their way to the detector.

We can compensate for a dimmer image by taking a longer exposure. And so if the subject is motionless and the camera is on a stable mount, then perhaps one could make the image as sharp as one likes simply by using the tiniest possible hole, even if one had to wait all week for a proper exposure to build on the film. But this will only work up to a point because of the phenomenon of *diffraction*.

Recall that because of the wave nature of light, a ray of light passing through a tiny hole spreads out to form a bullseye pattern. *And the tinier the hole, the more the light spreads out.* This means that the light passing through a tiny pinhole will not simply travel in a straight-line ray; it will spread out, and dramatically so if the pinhole is only a little bigger than the wavelength of the light. See the right side of figure 8.2 for an illustration.

Imagine taking picture after picture with a pinhole camera, each time using a slightly smaller hole. For large holes, diffraction has little effect. And so at first, because of simple geometry, smaller holes would yield sharper images. But as one tried smaller and smaller holes, eventually the images would stop getting sharper and sharper. Instead, because of diffraction, smaller holes would make the image *blurrier*.

And so there must be a particular size hole that gives the sharpest image. Images from pinholes both larger and smaller than this optimum size would be blurrier—large holes give a blurry image because of geometry, while tiny holes give a blurry image because of diffraction. This is the best one can do with a pinhole camera; we are up against the very nature of light itself.

With a little bit of calculus and some algebra, one can derive an equation for that optimum pinhole size. The details of the calculation and its motivation are in appendix C, but the result is very simple, and given by equation (8.1)[2].

$$D \approx \sqrt{2.44F\lambda} \tag{8.1}$$

[2] There are slightly different ways to define what is the 'sharpest' image, and so there are other versions of equation (8.1) that are slightly different. When applied to specific numerical examples, however, they give similar results.

where D is the best diameter of the pinhole and λ is the wavelength of the light. If we choose $\lambda = 550$ nm (the middle of the visible part of the spectrum), and convert the units so D comes out in millimeters, we have:

$$D\ (\text{mm}) \approx \frac{\sqrt{F\ (\text{mm})}}{27} \tag{8.2}$$

I also define the *focal length*, F, of the pinhole camera—the *distance between the pinhole and the light detector*. And so a pinhole camera with a focal length of 8 inches (about 200 mm) should have a pinhole about half a millimeter across in order to produce the sharpest possible images.

8.1 Focal length and angle of view

The distance between the pinhole and the light detector is the focal length, but what about the other dimensions of a pinhole camera? And what is the difference between a large focal length on the one hand and a small focal length on the other? To answer these questions, we must distinguish between four different concepts: focal length, detector format, image size and angle of view. There is another quantity, focal ratio, that relates to focal length in a different way; we consider focal ratio in section 8.4. All of these concepts apply not only to pinhole cameras, but to all cameras in general.

Focal length: The focal length, F, is the distance between the pinhole and the light detector.

Image size: the size of the image on the detector, directly measured in some unit of length.

Detector format: the physical size of the light detector in the camera. Is the detector 6×9 cm or is it 8×10 inch?

Angle of view: the range of directions that can be imaged at once, as seen from the pinhole looking toward the subject.

We consider each of these in turn, and the relations between them.

8.1.1 Image size

The size of the image, s_i, on the detector of a pinhole camera depends on three factors: the size, s_o, and distance, d, of the subject, and the focal length, F, of the camera. The relation between these four quantities is simple, and follows directly from the law of similar triangles, as applied to figure 8.1:

$$s_i = s_o \frac{F}{d} \tag{8.3}$$

And so for a given subject at a given distance, a camera with a longer focal length will produce a larger image of that subject. We shall see in chapter 9 that equation (8.3) is strictly-speaking not true for the case of an image formed by a lens, although it is often at least approximately true even then.

8.1.2 Detector format

Most cameras are built around the particular light detector they employ. And so some cameras are designed to use a roll of light sensitive film, each frame of which is a rectangle that measures 36 × 24 mm. And so the detector format of such a camera is 36 × 24 mm. Such a camera has long been called '35 mm' as a shorthand.

There have been literally dozens of different detector formats in widespread use over the history of photography. But for any given camera, it is usually the one thing that is fixed and unchangeable by the photographer. A 35 mm camera uses a format of 36 × 24 mm and only that format, by way of its fundamental design.

The typical pinhole camera is the exception. It can be made by taping a piece of darkroom enlarging paper to the inside of a box. The light-sensitive paper can be purchased in many sizes, and it can be easily cut smaller. And so the photographer can put whatever size piece of paper they want in their pinhole camera, so long as it fits in the box. And if it does not, they can simply find a different box.

Equation (8.3) for the *image* size holds whatever the detector format. It may be that a particular desired image does not fit on the detector, because the format is too small. But even then, the part of the image that *does* fit will follow equation (8.3) for the part of the subject that is imaged.

Often we are concerned in the end with neither the image size nor the detector format *per se*. Rather, we may care instead about how the two compare to each other. And so consider two cameras of the same focal length situated right next to each other, both pointing at the same subject. The image size of the subject, as measured in inches, would be the same in both cameras. But if one camera has a larger detector format than the other, then that image will be a smaller percentage of the full picture. Does the picture only include your friend's head, or does it include their whole body as well?

8.1.3 Angle of view

Figure 8.3 illustrates the concept of angle of view, θ, for a pinhole camera. The angle of view depends on two factors: the focal length and the detector format. It is clear from these diagrams that for a given focal length, a larger detector will produce a larger angle of view. But for a given detector size, a longer focal length will produce a *smaller* angle of view.

Since many cameras have a detector format that is rectangular, a camera's angle of view may have different values for the horizontal and vertical parts of the picture. As such, the angle of view is sometimes described by that given by the diagonal of the format.

If one bisects the angle of view in any of the diagrams in figure 8.3, two right triangles are formed inside the camera. It is easy to show with a little trigonometry that the relations between the detector size, S, the focal length, F, and the angle of view, θ, are given by:

$$F = \frac{S}{2\tan(\theta/2)} \tag{8.4}$$

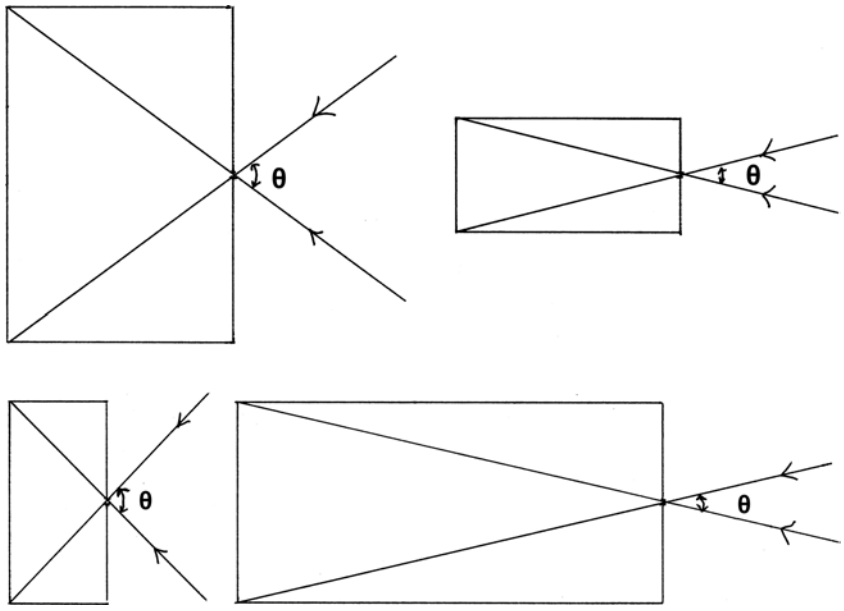

Figure 8.3. Top: two different angles of view, both with the same focal length. The larger detector on the left produces a larger angle of view. Bottom: two different angles of view, both with the same size detector. The shorter focal length on the left gives a larger angle of view.

$$S = 2F \tan(\theta/2) \tag{8.5}$$

$$\theta = 2 \tan^{-1}\left(\frac{S}{2F}\right) \tag{8.6}$$

A large angle of view is often called *wide-angle* while a small angle of view is sometimes called *telephoto*. This terminology is more clear if one thinks of it in terms of most professional cameras; they have a fixed detector size, but use interchangeable lenses of different focal lengths.

Imagine that I build two pinhole cameras, both to hold an eight-inch square piece of darkroom paper, but one of them has a longer focal length than the other. The one with the *longer* focal length would then have the *smaller* angle of view, as shown in the bottom of figure 8.3. But also for that camera, the longer focal length means that the image of a particular object (a squirrel, say) would be *larger*. On the other hand, the camera with the shorter focal length, and thus the larger angle of view, would show more than just a single squirrel. It would also show the squirrel's friends gathered around it in a large crowd. With my telephoto pinhole camera, I can see only the one squirrel, albeit in much greater detail. And so, unless I am able to read the sign it is carrying, with my smaller angle of view I would not know that it is only one of a mighty crowd of squirrels in open rebellion, angry and dangerous.

Pictures taken with an angle of view of about 35–40° seem approximately 'normal' to most people, and so that is often the dividing line between what is called wide angle (or short focus) and what is called telephoto (or long focus). From equations (8.4), one can see that an angle of view of 40° means the camera's focal length is about 1.4 times greater than the size of its detector.

8.2 Distortion and angle of view

Figure 8.4 illustrates two different ways to produce an image of a given object to be a specific size on the detector. One can either use a long focal length from far away or a short focal length from close up. In this example, the sizes of both the image and object are exactly the same, but the angle of view and the distance between the subject and the camera are different. If the object is simply an arrow on a flat wall, then both situations would produce identical images.

But a picture taken in the real world usually includes a variety of subjects of different sizes and distances, and three dimensional shapes that have depth. And this means very different results can arise when using different angles of view, even if they produce the exact same image size for the main subject.

Let us consider as an example two ways of making an image of the head of Tobias, the cat. Figure 8.5 shows the geometry of Tobias as imaged by a short focal length from close up. Although his nose is only about 1/3 the width of his head overall, in this image his nose would appear almost as big as the rest of his head.

An image of Tobias of roughly the same size could be made instead by using a long focal length from far away, as in the top example in figure 8.4. But in that case the image of the nose, as compared to the head overall, would be much closer to their real proportions.

From far away, the ears are only a tiny percentage farther away than the nose. But from close-up, the back of the poor kitty's head is twice as far from the camera as is his nose. This means that the image that results shows a much larger nose as compared to the size of his head. See figure 8.6 for the comparison of the two cases.

And so using a long focal length from far away will form an image such that the different parts of the image are in proportion to the actual sizes of the objects. Using

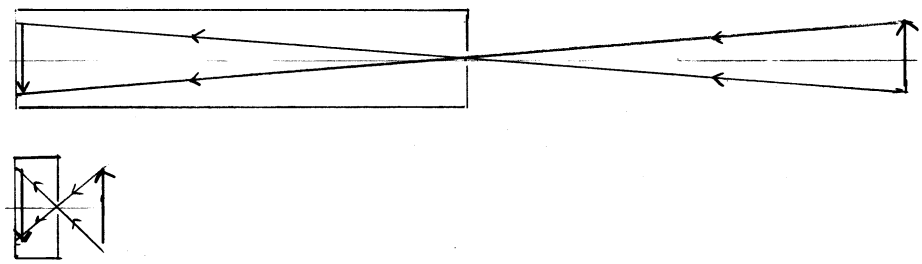

Figure 8.4. Two different ways to achieve the same size image of an object. On the top a long focal length is used from far away. On the bottom a short focal length is used from close up. Although the object portrayed in the diagrams will appear the same in both cases, objects at both nearer and farther distances will be portrayed differently.

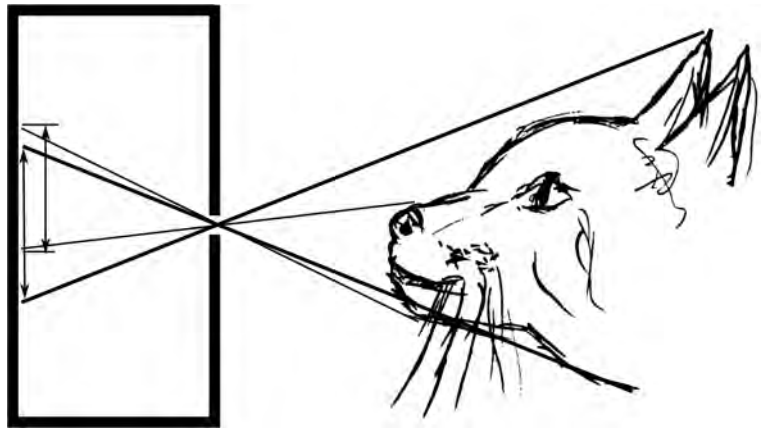

Figure 8.5. If the head of a cat is imaged at a large distance with a long focal length (telephoto), the sizes of the images of the nose and the back of the head are nearly proportional to their real sizes. But if instead an image of roughly the same size is made by moving close up with a short focal length (wide angle), then the image of the nose is disproportionately large compared to the image of the more-distant back of the cat's head.

Figure 8.6. The results of taking pictures of Tobias with two very-different angles of view. Left: a short focal length lens was used from close up, as in figure 8.5. Right: a long focal length lens was used from far away. In the wide-angle, close-up view, the nose was significantly closer to the camera than the ears, and so looks disproportionately large.

a short focal length from close up, on the other hand, has the effect of making the nearer objects look larger in comparison to more distant objects.

Using a long focal length (and narrow angle of view) tends to make things look flat, with little depth; distances seem compressed together. Using a short focal length (and wide angle of view), on the other hand, exaggerates depth and makes nearby objects look disproportionately large.

As humans, we are used to looking at things from a 'moderately close' distance, and so pictures taken from far away with a long telephoto seem unnaturally flat. Likewise, pictures taken with a very short focal length from very close up seem unnatural as well, but in the opposite sense. In practice, photographers have devised rules of thumb, guidelines, for what is the most useful angle of view for common

situations. An example is the rule that for a 'natural-looking' head shot, the angle of view should be roughly 20° (this corresponds to a 100 mm lens with 35 mm film). But as always, it is your picture. It is up to you to decide what is the most appropriate angle of view for the subject matter and the effect that you desire.

8.3 Vignetting

For a given value of the focal length, F, how large a light detector could one use in a pinhole camera? Would a pinhole camera with a focal length of only two inches, for example, produce an image that would cover a detector that is 8 × 10 inches?

Light coming straight in along the axis of the pinhole passes through a circle of a particular diameter. The brightness of the light that enters along that path is proportional not to the diameter of the pinhole, but rather to its area, πr^2 for a circular hole of radius, r.

But light rays passing through the hole at an angle see not a circle, but an ellipse, as in figure 8.7. This ellipse has the diameter of the pinhole as its long axis, but it is smaller than that along its short axis. Clearly, this off-axis ellipse has a smaller area than the pinhole itself, and so less light passes through at an angle.

This means that the image will be brightest in the center, exactly opposite the pinhole, and progressively dimmer farther from the center. This effect is called *vignetting*, and figure 8.8 shows a graph of brightness versus angle due to the differing appearance of the pinhole shown in figure 8.7. It turns out to follow a simple mathematical rule; the brightness is proportional to the cosine of the angle from the axis. One can see from the graph that at angles greater than 60°, only half the light enters the camera compared to rays passing straight through the hole. Clearly, an angle of view of 120° would produce significant vignetting.

In practice, vignetting is much more severe than one would expect from figure 8.8, since there are other factors besides the elliptical projection of the pinhole. See figure 8.9. For one thing, the thickness of the pinhole material enters into the calculation, and this means that there may be angles for which no light passes through at all. Furthermore, the edges of the detector are farther from the pinhole than is the center, and this causes additional vignetting.

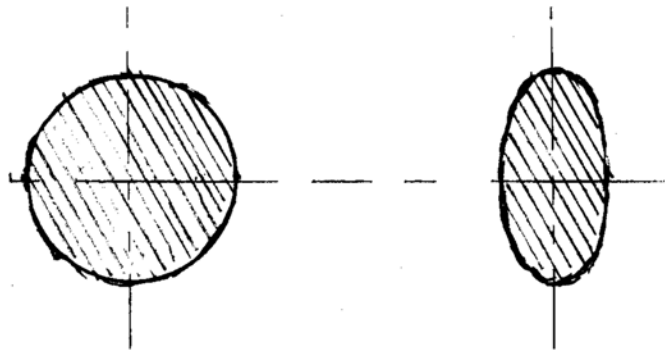

Figure 8.7. A circular hole appears elliptical, with smaller area (and thus admitting less light), when seen from an angle. Thus, the center is the brightest part of an image formed by a pinhole camera.

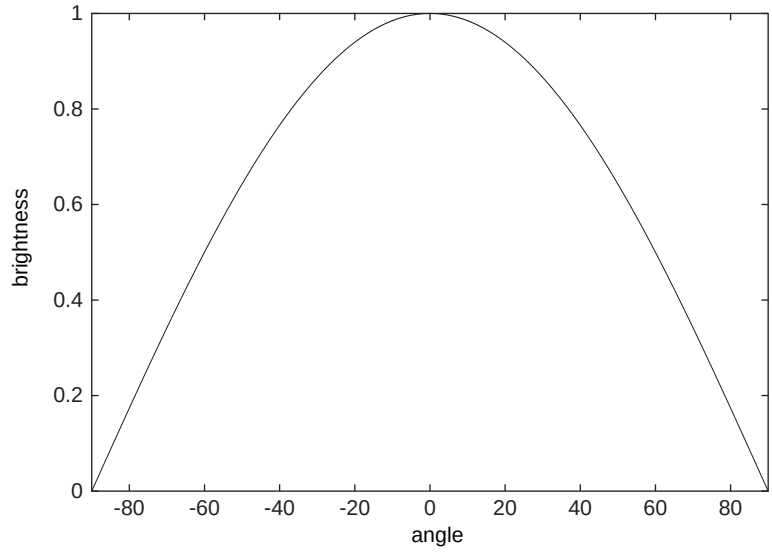

Figure 8.8. A pinhole camera naturally produces an image with vignetting, since light rays entering from an angle 'see' a hole of smaller area. This graph shows the percentage of light that makes it through the pinhole, at different angles.

Figure 8.9. *Polar(oid) Bear*, Teresa Patrick 2005. This wide-angle image from a pinhole camera shows vignetting at the corners.

8.4 Focal ratio

The *ratio* of the focal length, *F*. of the camera to the diameter, *D*, of the pinhole is called the *focal ratio*, *f*, of the camera:

$$f = \frac{F}{D} \tag{8.7}$$

The focal ratio of a pinhole camera is an important measure for a single reason—it is one of the factors that determines the exposure when the image is recorded by the detector. This topic is considered in detail in Volume 2 of *The Physics and Art of Photography*, but from the definition alone one can see that a *smaller* focal ratio means a larger hole, and so results in a *brighter* image.

For a camera with a lens instead of a pinhole, there is a second reason why the focal ratio is important; we consider that topic in chapter 9.

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Chapter 9

Ray optics 2: lenses

For a pinhole camera there is a necessary trade-off between image sharpness and image brightness, and the image sharpness has a natural limit set by the wave nature of light. We could, however use the property of refraction to redirect light rays from the edges of a large hole so that they end up at the same place on the image as do light rays passing through the center of the hole.

A wedge-shaped piece of glass, a prism, will do the trick. Near the edge of the hole, the rays must be deflected at a greater angle (see the left side of figure 9.1), and so a more-angled wedge is required. Nearer the center of the hole, the light rays need not be deflected as much, and so a flatter piece of glass would be needed. All of these little prisms can be put together into one smooth shape, called a lens (see the right side of figure 9.1). You may have noticed that a lentil has basically the same shape, and that the words 'lens' and 'lentil' have much in common (see figure 9.2).

And so a lens allows one to get around the inherent geometric restrictions of using a pinhole to make an image. We can now, in principle, use as large a hole as we want, so long as we can design a lens to deflect all of the light rays to their correct places on the image. In practice, there are of course limitations, and we will talk some more about those later. But for now, let us put a lens on our camera and see what happens.

9.1 Focus

One consequence of using a lens to form an image instead of a pinhole is that, for a given distance to the object, there will be only *one* lens–film distance for which the lens will bring the light rays from one point on the object to a single point on the image. This should be clear from figure 9.1. If one were to move the detector closer to or farther from the lens, the light rays would not all meet at one spot.

This is not the case for a pinhole camera. It is true that, because of the effects of diffraction and ray geometry, there is an optimum focal length for a given pinhole size. But this is a subtle effect, and it does not depend at all on the distance to the

doi:10.1088/2053-2571/aae1b6ch9

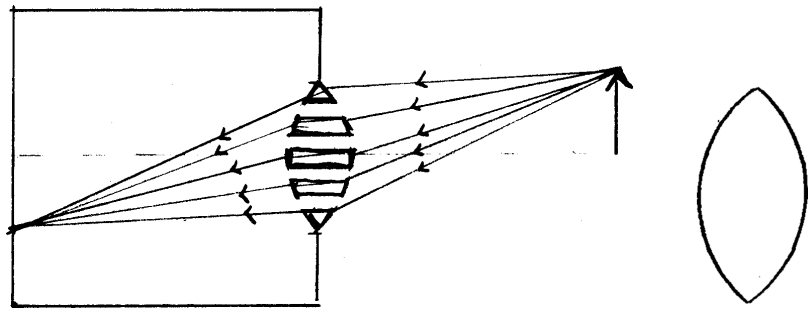

Figure 9.1. Light from a large opening could, in principal, be redirected to a common focus by appropriately-placed prisms, as illustrated on the left. On the right is a cross section of a double convex lens. Each part of the lens acts as a prism to refract light by different angles, such that they all reach a common focus.

Figure 9.2. Lenses and lentils.

subject. So for a pinhole camera, simply use the right size pinhole for your focal length, and everything at all distances will be equally in focus.

Introduce a lens, however, and for a given object distance there will be only one image distance for a perfect focus. This means that one cannot, in principle, get objects of different distances to all be in focus at once. It also means there must be some mechanism for adjusting the lens–film distance to achieve perfect focus for a given object. This is called, of course, *focusing* the camera, and there are many strategies for accomplishing this, all of them greatly complicating the construction of a camera. One of the appeals of pinhole photography is that no focus mechanism is needed, thus making the construction of a pinhole camera simpler.

We can use the most basic simple lens as an example to illustrate some important points about the focus of any lens. Figure 9.3 shows that light rays coming from a point on a nearby object diverge at a steep angle as they arrive at both sides of the

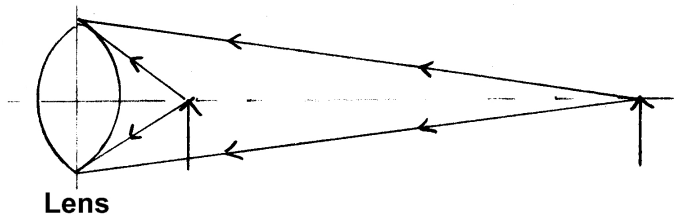

Figure 9.3. As they reach the edges of a given lens, rays of light from nearby objects diverge at a greater angle than those from objects far away.

camera lens. Light rays coming from a distant object, on the other hand, are more nearly parallel.

It should not be surprising that a given lens, which must bend the light rays inward to get them to converge, will bend those coming from the distant object at a steeper angle than those coming from the nearby object. For the nearby object with greatly diverging light rays, some of the refractive power of the lens must be used simply to straighten the rays before it can bend them inward. Thus *the image distance for a nearby object is greater than for a more distant object*.

What if we slightly change the distance of the two objects in figure 9.3? For the nearby object, a small difference in distance will make a big difference in how much the light rays are diverging when they arrive at the lens, and so will cause a significant difference in *image* distance. And so for nearby objects, small difference in distance lead to large differences in focus. For the distant object on the other hand, a slight change in distance will make much less difference in the angle of the light rays as they arrive at the lens. And so the focus changes little between two objects that are at different, but both very-distant, locations.

If the object were an infinite distance away, the light rays would be exactly parallel. But even if the subject is only *very* far away, the light rays arrive *practically* parallel to each other. For most lenses, put the subject one mile away or two miles away and the light rays are still so close to parallel that one can't tell the difference between them. And so every lens has an 'infinity' setting, but don't take the meaning literally.

9.2 Focal length

For a pinhole camera, we referred to the *focal length*, *F*, as simply the distance between the pinhole and the detector, the *image distance*. But for a camera with a lens, *the focal is the image distance for an object at infinity*. Said another way, the focal length is the distance required for a lens to bring parallel light rays to a focus. Since light rays coming from infinity are parallel, these two statements say the same thing. For a pinhole camera, the image distance and the focal length are the same thing. For a camera with a lens, this is only true if the object is very distant. For nearby objects, *the image distance is greater than the focal length*.

We can use the ray diagram in figure 9.4 to investigate the geometry of a lens in focus. The lens is represented by the vertical line in the center. The horizontal dashed line is the *axis* of the lens, an imaginary line everywhere perpendicular to the plane of

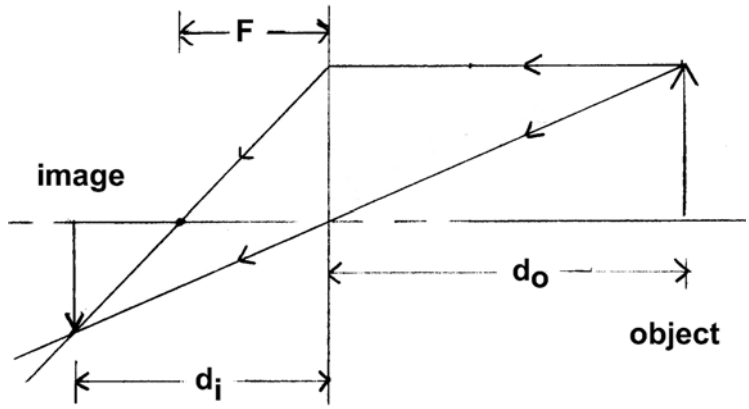

Figure 9.4. A ray diagram for a converging lens. Only two *principal rays* are needed to locate the image. The arrow on the right represents the object, and the lens (vertical straight line in the center) forms an upside-down image of it on the left. The distance at which rays parallel to the axis of the lens (the horizontal line in the center) come to a focus is called the *focal length*, *F*, of the lens. The distance between the lens and the image is the *image distance*, d_i, while the distance between the lens and the object is the *object distance*, d_o.

the lens. The arrow on the right represents an object to be imaged, and the arrow on the left side is the image of that subject.

In this diagram we only consider light rays from the tip of the arrow. There are an infinite number of such rays scattering off in all directions, but we only need to consider two of them, called *principal rays*. The *first principal ray* passes through the exact center of the lens. Since the center of the lens is flat, this ray passes undeflected. The *second principal ray* passes parallel to the axis of the lens. Since all light rays arriving at the lens parallel to the axis are—by the definition of *F*—focused at a distance from the lens equal to the focal length, then this ray must also cross the axis at that point.

If we then project these two principal rays further we see that they cross each other. If the lens is shaped correctly, any other ray emitted from the tip of the arrow and intercepted by the lens would also cross at that same point. Thus, the two principal rays show the location of the *image* of the tip of the arrow.

It is easy to confirm that if we do the same thing for a point half way down the arrow, the rays converge halfway between the lens axis and the image of the arrow's tip. Thus an image of the arrow is reconstructed as shown in the diagram. And so we have a lens, an object and an image of that object. We call the distance between the object and the lens the *object distance*, d_o. And we call the distance between the lens and the image the *image distance*, d_i. The other distance that is relevant is the focal length, *F*. This is a property of the lens itself; ultimately it depends on the shape of the lens and the type of glass from which it is made.

The two principal rays and the three distances, d_o, d_i and *F*, form a set of triangles. And so it only takes a little bit of geometry and trigonometry to see that there must be some mathematical relationship between those three quantities. I won't work through the details, but here is the answer:

$$\frac{1}{F} = \frac{1}{d_o} + \frac{1}{d_i} \tag{9.1}$$

This is called the *thin lens equation*, and it says in words that the reciprocals of the image and object distances add to give the reciprocal of the focal length. As we will discuss later, real lenses are almost always more complicated than this, but even for those cases, the thin lens equation is still a good starting point, as it is often approximately true even then.

So let us consider some implications of this equation. To help with this, it is useful to solve equation (9.1) for each of the three quantities separately:

$$F = 1/\left(\frac{1}{d_o} + \frac{1}{d_i}\right) = \frac{d_o d_i}{d_o + d_i} \tag{9.2}$$

$$d_i = 1/\left(\frac{1}{F} - \frac{1}{d_o}\right) = \frac{F d_o}{d_o - F} \tag{9.3}$$

$$d_o = 1/\left(\frac{1}{F} - \frac{1}{d_i}\right) = \frac{F d_i}{d_i - F} \tag{9.4}$$

These three relations imply the following:

1. For a pinhole camera, the focal length is simply the image distance. But for a camera with a lens, the focal length is a property of the *lens*, whatever the image distance happens to be.
2. Equation (9.2) shows that one can determine the focal length of a lens if one knows even one combination of image distance and object distance. Measure the distance between lens and subject. Then find the best focus (by, for example, moving a screen back and forth near the lens) and measure the image distance. Finally, use equation (9.2) to calculate the focal length.
3. For an object at infinity—or so far away that it might as well be at infinity—the image distance is equal to the focal length. And thus another way to determine the focal length of a lens is to simply measure the image distance for an object that is very, very far away (the Moon, for example).
4. Pick a focal length, and then plug different values for the object distance into equation (9.3), and it soon becomes clear that if the object distance is *smaller*, then the image distance is *larger*. Move the object closer to the lens, and the focus moves farther from the lens.
5. Equation (9.3) shows that if the object distance is *less* than the focal length of the lens, a *negative* image distance results. This means that there is no image at all of the kind we have been talking about. The light rays diverge so much that the lens is unable to even bring them parallel, let alone bring them together to a focus. Thus there is no *real image*—an image that can be focused onto a screen, piece of film or digital light detector. And so *a camera lens cannot focus on a subject that is closer to the lens than its own focal length.*

9.3 Depth of focus and focal ratio

In practice, careful focusing is sometimes very critical and other times not all that important. Consider the difference between the two examples in figure 9.5. Both have the same object distance, focal length and image distance. But the lens on the right has a larger *diameter*, D. Clearly, if one moves the detector in the right-hand example only slightly, there is a large error in how well the rays come together. But for the left-hand example, with a smaller diameter lens, the rays are less angled. And so the detector would have to be moved much farther to cause an equal error in focusing.

We can ask a different but related question. If the detector is located so as to form a perfectly focused image of a particular subject, by how much could I move that subject closer or farther away, and still achieve a not-perfect, but still-acceptable focus? This *range* of subject distances that still produce an acceptable focus is called the *depth of focus* or *depth of field*.

For the same reason that careful focus adjustment is more critical for the right-hand example in figure 9.5, the larger lens also has a smaller (more shallow) depth of focus. For a given focus setting, there is a smaller range of subject distances that are in acceptable focus for the large lens than for the small lens.

Since depth of focus is about the *angle* the rays make at the edge of the lens, it cannot be about lens diameter alone. For what if I made a lens with twice the diameter, but also twice the focal length? Clearly, everything would scale together and so the angles would be the same (so long as the object were also twice as far away). And thus depth of focus is primarily about not lens diameter, D, but rather how the lens diameter *compares* to the focal length, F. We have already defined a quantity that makes this comparison—the focal ratio, $f = F/D$. A smaller focal ratio means a larger diameter lens, as compared to its focal length. And so we have the following rule:

> *A smaller focal ratio means a more shallow depth of focus. A larger focal ratio means a larger depth of focus.*

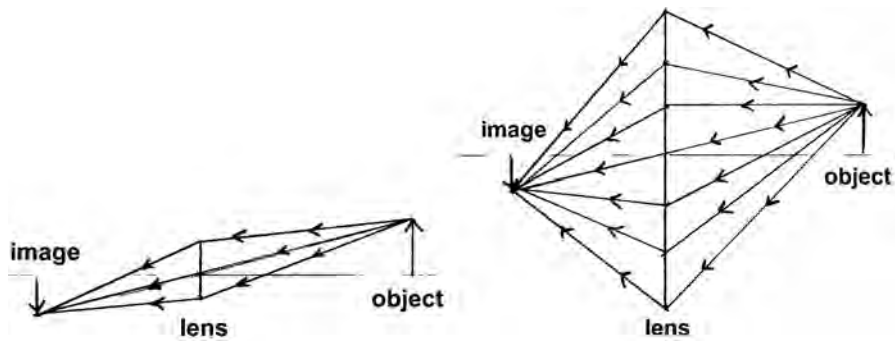

Figure 9.5. Left: ray diagram for a small diameter lens. Right: ray diagram for a large diameter lens. Since the rays converge to focus at a greater angle for the larger-diameter lens, the plane of focus must be located more precisely. And this means one must focus the larger lens more carefully than the smaller lens.

These effects of depth of focus are more pronounced, for all lenses, at small image distances, simply because the rays coming from the subjects are less parallel for nearer subjects. And so one must change the focus a lot when focusing on objects six inches and then twelve inches away. But for practically any lens one might ever use, the difference in focusing at ten miles compared to focusing at twenty miles is tiny indeed.

Most real camera lenses include a set of thin metal leaves that can slide against each other to make a hole of adjustable size in order to vary the *aperture*—the size of the hole where the light gets in. And so one can vary the aperture, the effective size of the lens, even though the glass lens itself is fixed. This is called an *iris* or *aperture stop*, and the adjustments are labeled with the focal ratio that results. For this reason focal ratio is often referred to as '*f*-stop.' Sometimes, to emphasize that *f* is a *ratio*, it is instead written as '*f/*.'

For most lenses, the *f* settings come in discrete increments, called *steps*. Recall that there is another reason to adjust *f* apart from depth of focus: the focal ratio is one of the factors that determines exposure when a picture is taken. And so the focal ratio steps are chosen to be equal steps in exposure, not depth of focus. This topic is considered in detail in volume 2 of *The Physics and Art of Photography*, but I note a key result here: equal steps in exposure are related not to focal ratio directly, but rather to its square.

This means that doubling the focal ratio changes the exposure by a factor of four rather than a factor of two. Since photographers choose exposure steps such that they result in successive doubling (or halving) of exposure, then our steps in *f* must be *square roots* of successive factors of two. And so successive steps in focal ratio are, for example, $\sqrt{1}$, $\sqrt{2}$, $\sqrt{4}$, $\sqrt{8}$, $\sqrt{16}$, $\sqrt{32}$, $\sqrt{64}$, etc. Approximating these square roots, the *f* settings on a typical camera lens is some subset of 1.4, 2, 2.8, 4, 5.6, 8, 11, 16, 22, 32, 45, 64, 90.

9.4 Zone focusing

The *infinity setting* on a camera lens is set for the exact focus at—literally—infinity. But because any given lens has a certain depth of focus, it means that things even 'beyond infinity,' if there could be such a thing, would be in focus as well. The landscape photographer often wants a large depth of focus, with not only the most distant objects but also subjects as close as possible to be in acceptable focus. We have already seen that a large focal ratio is the best choice for this situation. But one can do even better by putting the best focus somewhat closer than infinity—such that infinity is still in acceptable focus. But now the range of acceptable focus has been moved even closer. Such a setting is called *hyperfocal infinity*, and many manual-focus camera lenses have marks to allow one to set them that way.

With an understanding of depth of focus, one can predict ahead of time what range of distances will be in acceptable focus for a particular combination of focus and *f*. This leads naturally to the idea that one can *pre-focus the camera lens for the situation*. Whenever only a restricted range of distances is the center of interest, one can pre-set the lens focus to the middle of that range of distances. If the focal ratio is

then picked so that the depth-of-focus extends for the full range of distances of interest, then one can pre-focus in this way. This process of carefully choosing a pre-set focus *and* focal ratio to give the required depth of focus, is called *zone focusing*.

Zone focusing was common before the arrival of cameras with electronic autofocus, and many old cameras included markings to make zone focusing easier. But even now, zone focusing is sometimes useful. The autofocus process is not instantaneous, and so there is an inevitable slight delay for the camera to focus. Zone focusing, *when used appropriately*, is *faster* than autofocus. Zone focusing may also be a good choice when the scene includes many subjects moving rapidly in different directions; an autofocus mechanism may be unable to quickly and correctly identify a proper focus subject.

9.5 Ray tracing

For the design of a real camera lens, the thin-lens approximation is far too crude. For the simplest type of lens, with two surfaces that are simply sections of spheres, rather than some more complex shape, there is an equation (called the lensmaker equation) that allows one to find the focal length, F, given the curvature of each side and the dimensions of the lens. But even this is too crude as anything but a starting point for the complex situations faced by lens designers

Instead, optical systems are designed with the aid of *ray tracing*. Rays coming from the object are followed geometrically until they reach an interface—a sudden change of material, such as from air to the first surface of a glass lens, or from one type of glass to another. At that point of contact with the new surface, geometry is used to find the angle of incidence and then Snell's law is used to calculate the angle of refraction. The ray is then followed geometrically until it reaches the next interface. And on and on, until the image point is finally reached. See figure 9.6 for an example of rays traced through two very simple, but decidedly non-thin lenses.

It should be obvious that the tedious calculations of ray tracing are much easier now than before the advent of the computer.

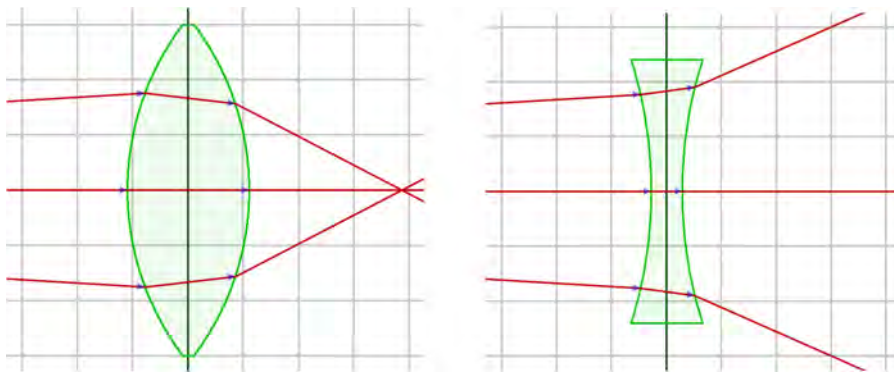

Figure 9.6. The thin lens approximation is too crude for precise optical design. Instead, individual rays are 'traced' from the source to the image, with Snell's law applied at each interface.

9.6 Aberrations and distortion

A camera lens has a lot of work to do. It must bring light rays of many wavelengths, coming from many directions, all to a proper focus at the right part of the detector. In practice, no single lens can do all of this, and so a real camera 'lens' incorporates many different *lens elements* to correct each other's mistakes. See figure 9.7 for an example of a photograph taken with only a simple double-convex lens. In this section we consider the typical distortions and the three most important *aberrations*—errors in bringing the image to a proper focus—that are inevitable with a simple lens.

9.6.1 Spherical aberration

A spherical lens surface is one that is shaped like a slice off the edge of a sphere. This is the easiest type of surface to make, by a long shot, for the simple reason that all parts have the same curvature. Lenses that have non-spherical surfaces are called *aspherical*, and they are much more expensive to manufacture. The problem is that spherical lenses do not form perfect images because different parts of the lens have different focal lengths; it is not really the correct shape needed to bring all of the rays to a common focus.

Figure 9.8 shows parallel rays traced through two lenses. They look almost identical to the eye, but they are not. The left-hand lens is spherical, while the right-hand lens has a more-complex hyperboloidal curvature. *Spherical aberration* means that rays of light passing through the edge of lens focus more closely than rays passing through the center of the lens. The (difficult to manufacture) hyperboloidal curvature eliminates spherical aberration in this particular example, but other

Figure 9.7. This photograph was taken with a simple single-element magnifying glass attached to a 35 mm camera. It resulted in large amounts of distortion and aberrations.

aberrations arise when the light rays are not parallel or arrive at the lens at an angle to the lens axis.

9.6.2 Coma

Even if a simple lens has the correct aspherical shape to bring on-axis light rays to a proper focus, it will still mis-focus light rays coming in from an angle to the lens axis. The result can be seen in figure 9.9. Severe coma spreads out light from point sources into fan shapes that point away from the center of the image. The effects of coma are greater farther from the image center, and with a wide-angle lens.

9.6.3 Chromatic aberration

The focal length of a thin lens is related to the index of refraction of the glass it is made from. Since a larger index of refraction means the light rays coming from air bend at a larger angle, a larger index of refraction results in (all else being equal) a

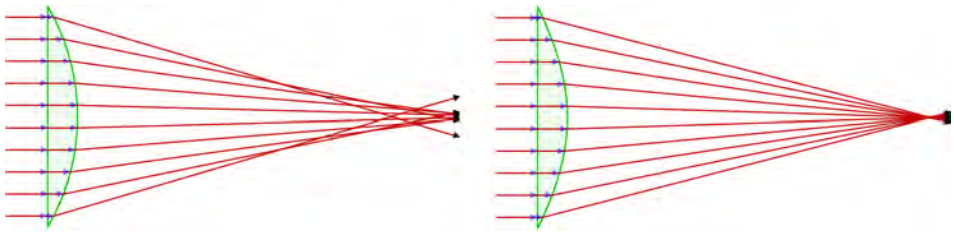

Figure 9.8. Although these two lenses appear almost identical to the casual eye, the mathematical shapes of their curvature are different. The example on the left is a spherical lens, while the lens on the right has a more-complex hyperboloidal curvature. The spherical aberration of the left-hand example is obvious from the closer focus of rays from the edge of the lens than from the center. Chromatic aberration has been ignored in both of these examples.

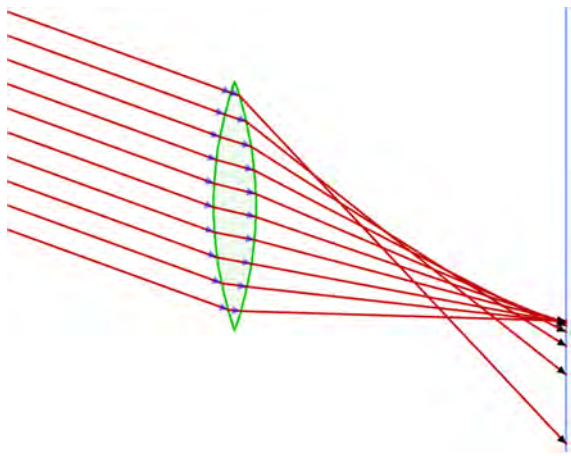

Figure 9.9. Parallel rays entering a simple lens at an angle focus in different places, making a point spread out into a comet-shaped blob.

shorter focal length. But the index of refraction of glass is not a simple number; it depends on wavelength. This is the *dispersion* discussed in chapter 3, section 3.3.2.

The examples of dispersion curves in figure 3.11 show that *shorter* wavelengths have a *larger* index of refraction. And thus, combining these two relations, one would expect an ordinary lens to focus shorter wavelengths to a shorter focus, and longer wavelengths to a longer focus.

See figure 9.10 for an example of rays traced through a simple lens, with this dispersion accounted for. Notice that the short-wavelength violet rays focus more closely, while the long-wavelength red rays focus at a larger distance. This error in focus is known as *chromatic aberration*.

Even though it is about wavelength, and thus color, chromatic aberration is important for black and white photography as well as color photography; even black and white detectors are sensitive to many wavelengths at once. And so those different wavelengths coming from the same subject would all focus differently, resulting in a blurry image.

One way to reduce chromatic aberration is to use a filter to restrict the light entering the camera to only a narrow range of wavelengths. See figure 9.11 for an example. But this is usually not a very practical solution, for two reasons. First, a more narrow range of wavelengths also means less light enters the camera, and so a proportionally longer exposure is needed. But also, if only a narrow range of wavelengths is admitted, then color photography is impossible.

The best way to reduce chromatic aberration is to abandon the use of a single, simple lens. We combine a strong—but low dispersion—converging lens with a weaker—but higher dispersion—*diverging* lens. The two lenses are made of different types of glass so they have different amounts of dispersion. The trick is to choose the types of glass and the shapes of the lenses such that the combination still acts as a converging lens; the diverging lens is weaker than the converging lens regarding the overall bending of light rays. But the *dispersion* of the diverging lens *does* cancel the dispersion of the converging lens.

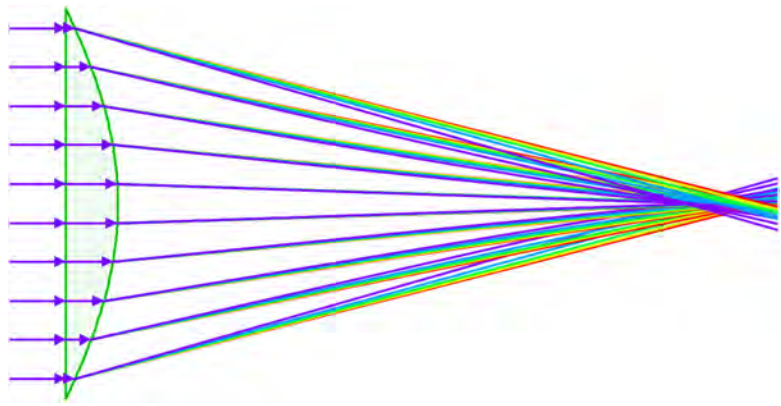

Figure 9.10. Even if spherical aberration is corrected, a simple lens will focus different wavelengths of light differently because of the dispersion of the glass.

See figure 9.12, where I have traced rays of many wavelengths through such a combination—called an *achromatic lens*, or simply an *achromat*. Chromatic aberration can be reduced even further by combining more than two lenses. A lens combination that effectively eliminates nearly all detectable traces of chromatic aberration over the range of visible wavelengths is often called an *apochromatic* lens.

9.6.4 Aperture and aberrations

One way to reduce aberrations in general is to use a lens with a very large focal ratio. Regarding chromatic aberration, for example, one can see from figure 9.11 that the rays entering the edges of the lens are refracted by a greater angle—but also those rays show greater dispersion. The rays entering through near the center of the lens, on the other hand, are bent very little and show much less dispersion.

For any given lens, all aberrations are generally worse when more of the lens is used. And so a common way to get a sharp image with a not-so-great lens is to 'stop

Figure 9.11. Both of these images were taken with the same simple double-convex lens, and they show pronounced chromatic aberration. But the detail is sharper in the example on the right because a filter was used to greatly restricted the range of wavelengths of light. The image is still blurry, especially at the edges, because spherical aberration and coma still remain.

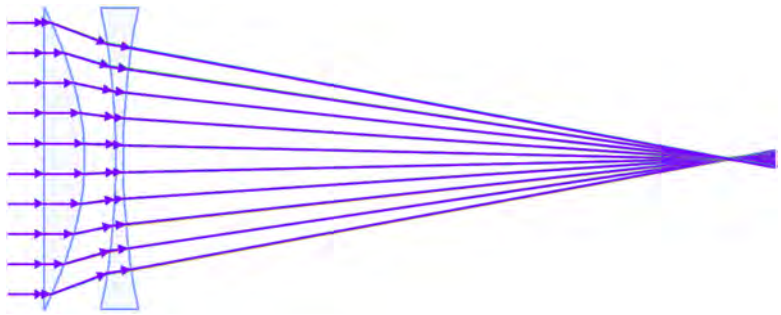

Figure 9.12. Chromatic aberration can be partially corrected if at least two lenses, of different dispersions and indexes of refraction, are strategically combined. In this example, the positive lens on the left is stronger than the negative lens on the right, so the overall effect is that of a positive converging lens. But the negative lens has a higher dispersion, and cancels out the dispersion of the positive lens.

it down' to a very small aperture. This is especially true for lower-quality lenses that are not very well corrected for aberrations (see figure 9.13).

If we take this idea to the extreme, we are back to pinhole photography, where the image is formed simply by geometry. Even then, however, the image would still be blurred by diffraction; no amount of lens-design cleverness can sidestep the fact that light is a wave. But also, one often wants to use a large aperture because more light is needed, or a shallow depth of focus is desired.

9.6.5 Distortion

Even if a lens does a good job at getting all of the rays in good focus, they may not end up on the detector in the desired locations. For example, it might be that a particular straight line in the subject might focus to a perfect line on the detector, but that it appears curved instead of straight. This is called *distortion*, and as with aberrations it is difficult to design a lens that produces a distortion-free image.

For a wide-angle lens especially, there is an important sense in which it is impossible. The detector is a flat surface, and it is recording the image of light from a three-dimensional world. And so it is often a choice of what type of distortion is the least-undesirable. A familiar example is the *barrel distortion* produced by a *fisheye lens*. See the left-hand image in figure 9.14 for an example. In this type of distortion straight lines passing through the center of the image appear straight, but straight lines off-center seem to curve around the image center.

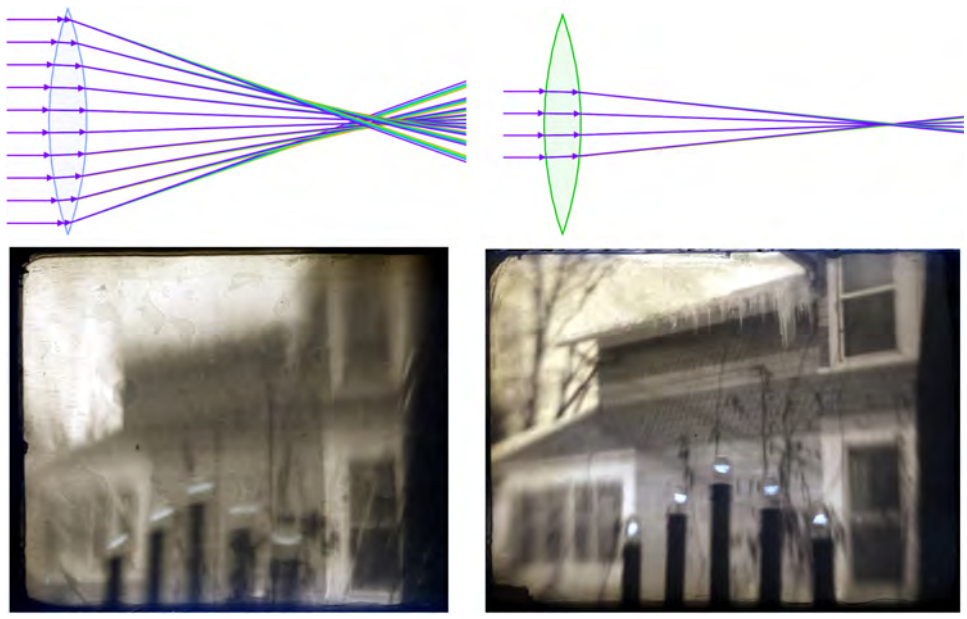

Figure 9.13. If a larger focal ratio (smaller aperture) is used, aberrations are greatly reduced. Both of these images were taken with the same simple double-convex lens, and they show pronounced coma and both chromatic and spherical aberration. But the detail is sharper in the example on the right because an aperture stop was used to only allow rays passing through near the center of the lens to enter the camera.

Figure 9.14. Left: *The World is Flat*, John Beaver 2010. This image, taken with a fisheye lens, shows considerable barrel distortion, which curves the horizon line. Right: *What Happened*, John Beaver 2012. This image, taken with a wide-angle rectilinear lens, preserves straight lines. But sizes of objects are exaggerated at the edges of the image.

The right side of figure 9.14 shows an image from a lens of nearly the same focal length as the image on the left. But it was made with a (considerably more expensive) *rectilinear lens*, rather than a fisheye. For this image, all straight lines in the subject appear as straight lines on the image, no matter where on the image they are. But the result is that *areas* are distorted; objects near the edges of the image appear larger than they really are.

9.7 Resolution

Let us imagine we want to take a photograph of the stars at night. This example brings up many technical details regarding exposure, but let us put those aside for now. A star, as seen from Earth, is essentially a point of light[1]. But because of diffraction, the light from the star will not form a perfectly point-like image, even if the lens is perfect, with no aberrations. Instead, the image will be the bullseye diffraction pattern of a circular opening (see chapter 3, section 3.7). Thus the point-like object will be imaged as something larger than a point. And if one tried to image two stars much closer together than the size of that diffraction pattern, then they would be indistinguishable as two stars; they would instead blur together as one.

This means the size of the diffraction pattern determines a maximum theoretical *resolution*—the maximum amount of detail one can discern in the image. To achieve a *higher* resolution, a point must form an image that is more like a point. That is, one wants the diffraction pattern to be *smaller*. For a particular wavelength of light (say the 500 nm in the middle of the visible spectrum), this means one needs a *larger* lens.

There is no way around this basic fact; it arises from the very wave nature of light itself. The size of the lens limits the detail in the picture, and for a given wavelength of light, nothing can increase it but to use a larger lens. This is not the same thing as

[1] As an example, the Sun and the nearest major star, α Centauri, are typical stars of similar size to each other, and the distance between them is typical of the distances between stars in our part of the Galaxy. Yet the scale of the distance, compared to the sizes of the stars, is like ping pong balls spaced hundreds of miles apart.

magnification. Magnify the image all one wants, but once the maximum detail is visible, there is nothing more to see. A large blur is still only a blur.

For most everyday photography, one does not often come up against this theoretical limit to resolution. Because of the aberrations discussed in section 9.6, the image usually has much less detail than this anyway. And so for most lenses, the resolution is improved by using a *smaller* lens aperture setting, simply because less of the imperfect lens is used. This is especially true for low-quality lenses.

But for some very high-quality lenses used for technical applications, there is a particular aperture, somewhere between the minimum and maximum settings, that gives the highest resolution. Larger apertures give a higher theoretical resolution because of diffraction, but smaller apertures have less-severe aberrations, and so there is some happy medium in between.

9.8 Lens design

What photographers call a 'lens' is actually a combination of three to as many as a dozen or more individual lenses of different shapes and spacing, and even different types of glass. Because of this we usually refer to the combination of many lenses that work as a unit, and that one attaches to a camera, as a 'lens', while each individual glass lens inside it is called a *lens element*. And so a given camera *lens* may contain many *lens elements*.

There are really only two ways to design a camera lens with minimal aberrations and distortions, even at small focal ratios. One can use a combination of many, many lens elements, each correcting for the aberrations of the others. Or one can use fewer lens elements, but with complex aspherical surfaces. Both choices are expensive. But using a large number of lens elements has the added disadvantage that every lens surface, both front and back, reflects a little bit of light. One problem with this is that the reflected light decreases the brightness of the image, since if it is reflected then it doesn't pass through. But even worse is the fact that about half of these reflected light rays *do* eventually make it through the lens and onto the image—but they end up completely in the wrong place.

Most of this unfocused light is spread diffusely over the image, and this adds an overall diffuse light to the image and *thus reduces contrast*. For color photography, not only is the contrast reduced, but the colors are less brilliant too. And if there is a very small and bright source of light (the Sun for example) within view of the lens (even if it is out of the picture), then a *flare* can be produced—an oddly-shaped streak of light on the image produced solely from these errant reflections.

Long ago, it was nearly impossible to make precision aspherical lens elements, and so there was an inevitable trade-off between making a sharper image with fewer aberrations and distortion and using fewer lens elements. Thus one could buy a lens that had a large aperture and produced sharp, low-distortion images, but it would inevitably produce pictures of low contrast due to the large number of lens elements. A high-contrast lens, with only a few lens elements, would have noticeable aberrations at small focal ratios. Many of these old and simple high-contrast lens

Figure 9.15. Left: a modern camera lens contains many lens elements and precision moving parts. Right: *St. Mary's Church*. John Beaver 2003. This picture was taken with only a simple double-convex lens.

designs were completely unusable at focal ratios less than about *f*/4, which is very modest by modern standards.

Modern lenses are much improved for three principal reasons. First, it is no longer impossible to make precision aspherical lens elements. It is still expensive, but not absurdly so. Secondly, a lens can now be fully multi-coated (see chapter 3, section 3.6) with antireflection coatings that use wave interference to cancel out most of the reflections. Thus, many more lens elements can be used to cancel aberrations, without reducing the contrast to unacceptably low levels. Third, computers can be used to optimize the design of the lens over a wide range of conditions. Thousands of rays can be now be easily traced through a particular complex lens design, so it can be tested before it is even manufactured. A good modern lens is still expensive, but an expensive lens today makes a much more 'perfect' image than did its equivalent-dollar counterpart of 50 years ago.

See the left-hand image in figure 9.15 for an example of a modern camera lens disassembled to show the individual lens elements. Such a lens also has a complex *mechanical* construction; the spacings between many of the lens elements must change in a precise way as the lens is focused to different distances. This particular example also had an adjustable focal length, so both wide-angle and telephoto pictures could be taken with the same lens (called a *zoom lens*). The zoom feature greatly adds to both the optical and mechanical complexity.

And so, are old camera lenses useless today? Not to me. A perfect image does not necessarily make a perfect picture. The right-hand image in figure 9.15 was taken with an intentionally-awful lens—just a single-element double-convex lens. For many years now, the Soho Photo Gallery in Tribeca, NY has hosted an annual Krappy Kamera Competition. The only rule is that the picture must be taken with an awful lens. In 2005 I won second place with this picture, and so I guess I can be almost as Krappy as the best.

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Chapter 10

Symmetry

10.1 Transformations and invariance

Everyone knows that the image in a mirror is reversed left-to-right; it seems too obvious to mention that the image in a mirror would be just what we call a mirror-image. But with this obvious fact comes an obvious question. Why does a mirror show an image that is reversed left-to-right, but not top-to-bottom? A mirror is simply a flat, reflective plane—how does it know the difference between left-and-right on the one hand and up-and-down on the other? The answer is surprising to many: a mirror reverses neither left–right nor top–bottom; we are the ones who do the reversing.

To make the picture on the left side in figure 10.1, I used a permanent marker to write the letters L, R, T and B (for left, right, top and bottom) on a sheet of two-tone origami paper. When I hold the paper in front of the mirror, the image of the letters has the same orientation as my direct view of them. The object on the left (the letter L) is on the left, the object on the right is on the right, and the same is true for top and bottom. If an object is to the left of center, its image is also to the left of center. And the analogous rule holds true for objects that are above, below or to the right of center. So why then *do* the words and letters appear reversed when one views them in a mirror?

The trick is obvious—the mirror images of the letters in the left-hand image in figure 10.1 are of the marker bleed-through, as if the letters were objects floating in space rather than simply written onto a surface. What one usually means by the phrase 'image in the mirror' is not what is represented by the left-hand image in figure 10.1, but rather that of the image on the right, where I have rotated the paper 180° to face the mirror.

The point is that when we look into a mirror, we want to see the image as if we were looking not from our vantage point in front of the mirror, but rather from the vantage point of the virtual image behind the mirror. And so we imagine ourselves (or the sheet of paper with writing on it) undergoing a *transformation*. When I stand

doi:10.1088/2053-2571/aae1b6ch10

Figure 10.1. Left: the images of object reflected in a mirror appear in the same left/right/top/bottom orientation as they do in real life. Right: by 'mirror image,' we usually mean the image in the mirror *once we have rotated it* (about a vertical axis) to face the mirror. It is our act of rotation, not the mirror, that reverses left and right.

in front of the mirror, objects behind me but to my left also appear to the left in the mirror. But for me to see those objects I would have to turn around—and then they would be to my right.

So back to our original question: why does this work for left/right but not for top/bottom? The answer is that it in fact *does* work just as well for a top/bottom reversal. It is all about how I transform the paper in order to position it according to the mirror's perspective, rather than my own. In figure 10.1, I have rotated the paper 180° about a *vertical* axis passing through the center of the paper. But I can also rotate it 180° about a *horizontal* axis, as in figure 10.2. In this case the letters instead read correctly left-to-right, but reversed top-to-bottom. Note that the vertical stem of the L is still on the left as it should be, but the horizontal bar is now on the top rather than the bottom.

And so text usually appears reversed left-to-right (but not top-to-bottom) when seen in a mirror simply because we most often rotate the text horizontally, about a vertical axis, so that it faces the mirror.

This concept of a transformation—a precisely defined change—is one half of the important concept of *symmetry*. Figure 10.3 shows the images in a mirror of the words 'WOW - OXIDE,' both rotated and unrotated about a horizontal axis (on the top) and a vertical axis (on the bottom). For the rotation about the horizontal axis, the word 'WOW' has reversed top-to-bottom, while the word 'OXIDE' seems to have not. But for the rotation about the vertical axis, the word OXIDE has reversed left-to-right, while the word WOW has not.

The reason of course is that for each kind of rotation, one of the two words has a particular kind of symmetry, while the other does not. The concept of symmetry is highly intuitive, but it is important for us nonetheless to define it precisely. An important mathematical way to define symmetry is as follows: *a symmetry is an invariance under a transformation*.

A transformation is a precisely-defined change, while an invariance is a precisely-defined quantity that remains the same. And so a symmetry is when there is something that stays the same while some other thing is changed in a particular way. And so for our example of the word 'OXIDE,' the two-dimensional shape of the word remains the same when it is rotated 180° through a horizontal axis passing

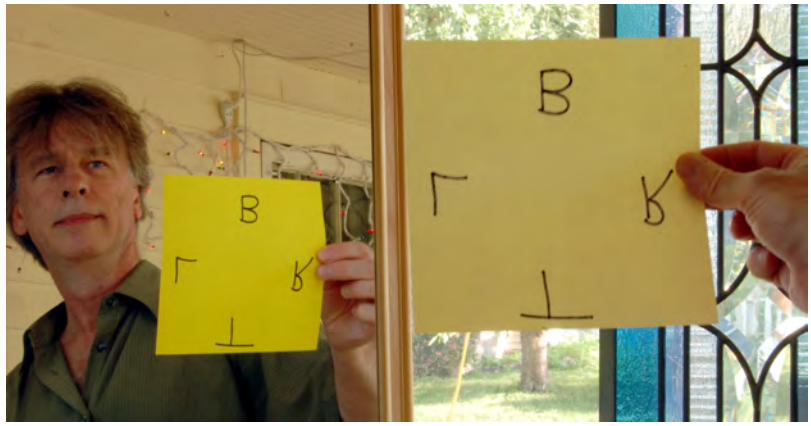

Figure 10.2. We can instead rotate the object about a horizontal axis to face the mirror. In this case, left and right are *not* reversed, while top and bottom are.

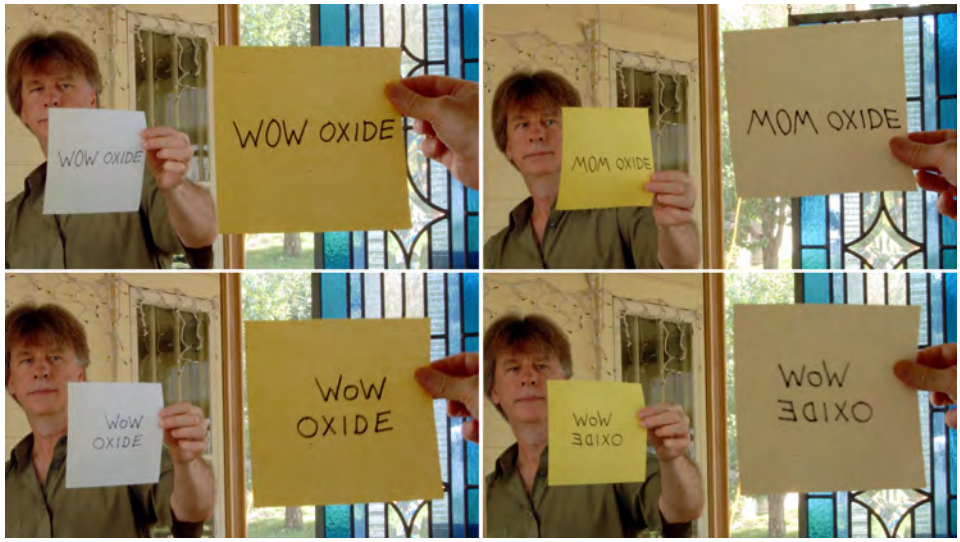

Figure 10.3. Top: the sheet of paper undergoes a rotation about a horizontal axis through the paper. The word OXIDE is invariant under this transformation, but the word WOW is not. Bottom: the sheet of paper undergoes a rotation about a vertical axis through the paper. The word WOW is invariant under this transformation, but the word OXIDE is not.

through its center. Thus, 'the two-dimensional shape of the word' is the invariance, while 'rotate the word 180° through a horizontal axis passing through its center' is the transformation. The word 'WOW,' of course, has a similar kind of symmetry. But the appropriate transformation is instead 'rotate the word 180° through a *vertical* axis passing through its center.'

10.2 Symmetry in physics

If all physical laws can be distilled down to a few simple rules, what are they? When I say 'simple,' I mean simple when expressed in the appropriate mathematical language—a language that may be as yet unknown, and almost certainly difficult to learn. The search for the answer to this basic question has been one of the primary concerns of theoretical physics over the past 100 years.

And the notion of symmetry—invariance under a transformation—plays a crucial role. It has been known for centuries that there are conserved physical quantities in nature—calculable quantities that stay the same as other measurable quantities change in complex ways. Two important examples are *momentum* and *energy*. The conservation of momentum as a principle was at least partially understood even before Isaac Newton; the conservation of energy was established as a principle by the mid 1800s, and hinted at much earlier.

And so the law of conservation of energy and the law of conservation of momentum are two examples of eight known conservation laws. Each conservation law specifies a conserved quantity that can be calculated exactly. And even as other measured quantities (velocity, position, acceleration, etc) change with time in complex ways, the conserved quantity stays the same. These conservation laws are some of the most powerful tools available to a physicist when she applies physical principles to the real world.

If one thinks of a conservation law as a statement that a particular quantity stays the same as everything else changes, it is not too difficult to see that this idea seems related in some way to the concept of symmetry. A symmetry is, after all, an invariance (something that stays the same) under a particular transformation (something that changes).

The two concepts do seem similar, but they are clearly not the same. For one thing, a symmetry involves a *particular* transformation (change), while a conservation law involves a quantity that stays the same even as *all other quantities* change.

But despite this difference, symmetries and conservation laws *are* intimately related, a fact first clearly established by the German mathematician Emmy Noether (figure 10.4) in 1915 (published in 1918). She proved what is now known as *Noether's First Theorem*, which states that *for every conserved quantity, there is an associated symmetry.*

This means, for example, that if the quantity *energy* is conserved for all physical laws, then there must be some symmetry that all of those physical laws share. Furthermore, she showed how to determine precisely what symmetry is associated with each conservation law. For the example of the conservation of energy, the symmetry is this: the very laws of physics themselves are invariant (remain the same) under the transformation of a translation in time.

Another way to say this is that as time passes, the laws of physics themselves do not change. The inevitable result of this symmetry is that a particular quantity (the total energy of a system) is conserved. The reason then, that all of our physical laws seem to conserve energy, is that they all share this basic symmetry.

Two other important cases are momentum and angular momentum. Both are conserved quantities, and each is associated with a different symmetry. The conservation

Figure 10.4. Emmy Noether (1882–1935) first established the mathematical connection between conservation laws and symmetry.

of momentum arises because—so it appears—the laws of physics do not change from one place to another. Angular momentum, on the other hand, is conserved because the laws of physics do not depend on what direction one is pointed in space.

What if, for example, one proposed a physical law that *did* change in some way as time passed? Well, then that law would violate the conservation of energy. And so the fact that, in practice, the conservation of energy works—it seems to hold up to careful measurement in all circumstances—is evidence that nature itself does have the particular symmetry of invariance under a translation in time.

Whatever the basic rules at work that nature uses, it seems that they do not change as time passes, or from one place to another, or when pointing in different directions. If they did, then we would not observe energy, momentum and angular momentum to be conserved quantities.

When we say that physical laws have symmetries such as invariance under translations in space and time, we are making claims about Nature. But how do we test whether or not these claims are true? Noether's First theorem shows us the way. The conserved quantities associated with these symmetries are measurable, and so we can measure them in experiments that test whether or not they are conserved. As we build more and more experimental evidence for a particular conservation law, Noether's First theorem demonstrates that we also provide evidence for that particular symmetry of Nature.

10.2.1 Symmetry and mirrors, again

I have described two particular types of symmetry that seem to be related to mirrors: rotate *a flat object placed parallel to a mirror* half way about either a horizontal or vertical axis, and it looks the same as before it was rotated. And this means that the reflection in a mirror of the rotated surface also looks the same. But we have also

seen that this kind of symmetry really has nothing to do with mirrors, *per se*. The transformation is our act of rotation such that the object that was facing us, now faces the mirror. And so 'mirror symmetry' is a misleading term for invariance under such a transformation. But clearly, a mirror does perform *some* kind of transformation. So what is it? The answer is obvious only if we consider three-dimensional objects.

We have seen that up, down, left and right are preserved when looking in a mirror. The reflection in the mirror of an object to the left of us also appears to the left, and the same is true for up and down. But what about the other dimension—distance? This is where the reflection in a mirror makes a real transformation. Consider figure 10.5; objects *closest* to the camera are *the most distant* in the reflection. This is a real transformation of nearer and farther; notice that the nearest example to the camera of *Castor canadensis* is in focus, but its reflection is not. That is because I set the camera focus on the nearby rodent, while its reflection is the most distant part of the picture.

And so, by a mirror transformation we really mean this: *a mirror transformation is a reflection about a plane*, rather than a half rotation around a line. Thus *mirror symmetry* is invariance under this definition of a mirror transformation.

This gives us another way of thinking about the symmetry of the word WOW. Instead of transforming the word by rotating it about a vertical line through its center, we can accomplish the same end result by putting a vertical *plane* through the center of the 'O' and reflecting it about that plane. So WOW *does* have mirror symmetry—but it is not the mirror of figure 10.3.

Figure 10.5. The image in a mirror is a transformation of nearer and farther. If an object is 1 m in front of a mirror, its image is 1 m behind the mirror.

We can easily extend this kind of mirror symmetry to three-dimensional objects. Place an imaginary vertical plane through the center of a three-dimensional object. If what is on the right side of the plane is exactly the same as the *reflection* in the plane of the left side, then we say the object has *bilateral symmetry*.

Mirror transformations are even more important when we consider motion. Imagine that I stand in front of a mirror and a bee flies from my left shoulder to the mirror and back, then landing on my right shoulder. And so the bee started from the left side, moving both towards the right and also further away. When it reached the mirror, it then—while still moving toward the right—moved closer and closer to then land on my right shoulder. But what about the bee's reflection in the mirror? The reflection of the bee also moved from left to right. But the reflection first moved closer and closer to me, and then, after reaching the mirror, moved farther and farther away.

We can consider an even more-dramatic example of a mirror transformation. Let us imagine that the bee instead flies in a horizontal circle in front of the mirror, and that as seen from above, it travels clockwise around its circular path. Clearly, the bee's reflection in the mirror also travels in a circular path. But as seen from the same vantage point, the bee's reflection in the mirror moves *counterclockwise* around its circular path. And so *a mirror transformation turns clockwise motion into counter-clockwise motion*.

10.2.2 Mirror symmetry and P-invariance

The relation between the motion of a particle and the motion of its reflection in a mirror is called by physicists a *P transformation*, and it points to another symmetry and associated conservation law. Symmetry under a P transformation is called *P invariance*, and it leads to what physicists call *the conservation of parity* (Sachs 1987 section 2.3). As is the case with the symmetries associated with the conservation of energy, momentum and angular momentum, P invariance applies not to the motions of particles themselves, but rather to the *physical laws that describe those motions*.

Let us now reconsider the bee flying in a clockwise motion in front of the mirror. Let us pretend to use Newton's laws of motion to analyze all of the forces acting on the bee in order to make it fly in its circular path. To say that Newton's laws obey the conservation of parity means that we can apply those same laws to the bee's reflection, and the reflected motion of the bee will be accurately predicted.

To do this, however, we must also apply the P transformation to the forces acting on the bee, and to its initial motion. When we do this, Newton's laws do in fact predict the reflected motion of the bee. This particular example is simple, for if the bee is traveling in a perfect circle at a constant speed, the forces acting on the bee must add up to a force of constant magnitude that always points toward the center of the circular motion. Clearly, if the force acting on the bee is toward the center of the circular path, so too is the reflection of that force in the mirror; the reflection of the force on the bee points toward the center of the reflection of the bee's circular path.

The conservation of parity means that, regarding fundamental physical law, there is nothing special about clockwise versus counterclockwise. It is all simply your point of view. This seems rather obvious; whether a clock runs clockwise or counterclockwise,

for example, is only a matter of from which side one views it. A clock hanging on the wall runs clockwise when seen from the vantage point intended—in front of the clock's face. But go around the wall to the room on the other side and use your superpower x-ray vision to view the second hand of the clock through the wall, and you will see it move counterclockwise instead.

Considerations of conservation of parity have played an important role in physics, giving meaningful clues to the nature of fundamental physical law. And even though conservation of parity seems, on the surface, to be an obvious truism, it is not always true. We consider this odd fact further in section 10.4.

10.3 Symmetry in art

Physicists have their own operational definition of symmetry, but the word is often used by artists in a somewhat different way—usually to specifically signify what we have here called mirror symmetry (for two-dimensional works) or bilateral symmetry (for three-dimensional works). But there are other forms of symmetry evident in two-dimensional art as well that can be described as an invariance under a transformation. I will first describe some examples of this *formal symmetry*, and then describe the related idea of *balance*.

10.3.1 Formal symmetry in art

Although photography—an example of two-dimensional art—is the main concern of this book, no discussion of symmetry can omit the example of a clay vase formed on a potter's wheel. *Azimuthal symmetry* is defined according to a particular axis in space, called the *symmetry axis*. If an object has azimuthal symmetry it means that one can rotate the object about the symmetry axis by any angle, and it still looks the same. For the potter's vase, it is the very act of forming the wet clay on a spinning platform that imposes this symmetry.

There are many examples of mirror symmetry in two-dimensional art; figure 10.6 shows a couple of my own. The grid in the background of the image on the left has a mirror symmetry when reflected about a vertical plane cutting through the center. Note that the same is *not* true for a reflection about any horizontal plane that cuts through the image, because of the uneven barrel distortion. And yet, there is still a sense that it has this type of mirror symmetry as well. Symmetry is so ingrained into our consciousness, that we 'fill in' the missing pieces with our imagination in order to give it symmetry. And so if one reflects the background of the image about a horizontal plane that cuts through the lowest horizontal line in in the grid, we can easily imagine the symmetrical form that would be created. I believe that part of the power of this particular image is that we see that cut-off part of the picture in our minds, even though it is not really there.

This basic idea that symmetrical forms can be implied or suggested is very important for art. It is one example of how a picture can evoke images that are not even in the picture. A picture is not only worth a thousand words, it is worth a thousand pictures.

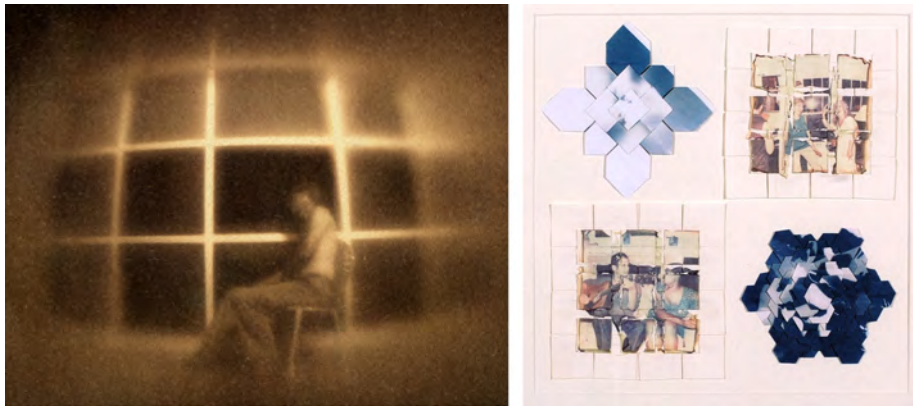

Figure 10.6. Left: *Self Portrait with Grid #2*. John Beaver, 2004. The background grid pattern in this photograph exhibits several types of symmetries. Right: *Celebration*. John Beaver, 2015 (folding by Elizabeth Breese and John Beaver). The overall arrangement of the pieces has mirror symmetry about both diagonals. The upper-left piece has four-fold symmetry while the piece on the lower right has six-fold symmetry.

There are many forms of symmetry evident in the right-hand image in figure 10.6. First, it is made of four parts of two different types, and they are arranged symmetrically. We can describe this particular symmetry in a couple of ways. Clearly, we could rotate the whole thing in its own plane by 180°, and we would still have the same arrangement of its four parts. But this particular symmetry is better captured by noting that it has mirror symmetry about a plane that cuts through the image *across either diagonal*.

The grid-like pattern in the origami-folded paper on the upper right and lower left clearly have many types of mirror symmetry, but what about the other two pieces? These patterns are best described by what is often called *N-fold symmetry*. Rotate (in its own plane) the pattern on the upper left by one fourth of a complete circle, and it is still the same pattern. This means we can rotate it four times to come back to its original orientation, and each of those four rotations produces the same shape. This is called four-fold symmetry, and by that same measure, it is clear that the figure on the lower right has six-fold symmetry.

There are many connections between the forms found in nature and those used by the artist, a fact well documented in Nathan Cabot Hale's fascinating study *Abstraction in Art and Nature* (Hale 1993). *N*-fold and bilateral symmetries are particularly common in European art of the middle ages—perhaps because those same symmetries are so often found in nature. And so a rose window on a medieval cathedral (see for example Duby 1992, cover) is like the petals of a flower. And the bilateral symmetry of the arrangement of figures in a 13th-century liturgical illustration (see for example Duby 1992, pp 31–2) is shared with the arrangement of appendages on vertebrates such as hummingbirds, whales and beavers.

10.3.2 Balance in two-dimensional art

Although it still finds its home in architecture, formal symmetry is much less common in the two-dimensional art of the Renaissance and later. I believe there is a

rather obvious reason that art evolved away from formal symmetry as it became more self aware: since any particular formal symmetry can be described by a set of simple mathematical rules of transformation and invariance, then where does the art come in? In more recent two-dimensional art, obvious formal symmetries most often appear as the exception that proves the rule. Its obvious presence in a contemporary work is intentional and glaring, and begs for a specific interpretation.

But there is a related concept that is still very strong—that of *balance* (Preble and Preble 1994 pp 92–3). The word balance is used in many ways by artists, but a picture has *visual balance* if it appears to be in a sort of equilibrium of weight (Barrett 2011). The idea of *visual weight* is a subjective combination of perceived size and weight that makes some forms in the picture look—well, heavier—than others. There is no operationally-defined mathematical way to capture this concept. One must instead have a bit of faith that it will be known when it is seen. And so the picture is said to have visual balance if the weights of its elements appear to be in a sort of almost gravitational equilibrium. It is almost as if one asks, 'if a shaft and frictionless bearing were passed through its center, would the painting want to twist one way or the other?'

There is a physical analogue to this. The *center of mass* of some object is the average location of its parts—except that one must calculate the average by counting higher-mass parts more in the average than lower-mass parts. For a two-dimensional object, if one allows it to freely rotate about a shaft placed through it, then it will want to rotate such that its center of mass is below that pivot point. And so if one places the rotating shaft directly through the object's center of mass, it will have no tendency to rotate either one way or the other.

An important case is when the center of mass of the object is below the geometrical center. From one point of view, it is unbalanced; the object is weighted below-center. But there is something different about this type of unbalance, because clearly the object would have no tendency to pivot either one way or the other if the pivot point is placed through its center. And if one did give it a push, it would desire to return to its original position. This is called by physicists a *stable equilibrium*, and something like it is probably also a part of the artists' concept of visual balance.

If on the other hand the object's center of mass is *above* a pivot point at its geographical center, then it might possibly be in equilibrium—so long as the center of mass is *exactly* above center. But all it takes is a gentle nudge from a fly sneezing in the Andromeda Galaxy, and it will flip one way or the other. This is also called an *equilibrium*, but it is instead an *unstable equilibrium*.

It may be that one could gather these physicists' ideas of center of mass and stable and unstable equilibrium and produce some kind of detailed mathematical model of the artists' idea of visual balance. But I doubt it, and I am skeptical that it would be worth the bother to try. The artist knows it when they see it, whether or not it can be defined in operational terms.

10.4 Asymmetry and broken symmetry

If symmetry is so important, then what about its absence? There are two important types of non-symmetry:

1. **Asymmetry:** one can identify no type of symmetry, no matter how hard one tries. A perfect asymmetry implies that even if small features are ignored, there is still no identifiable symmetry.
2. **Broken symmetry:** there *would* be an obvious symmetry, if it were not for some particular small detail.

And so an amoeba is asymmetrical, as is its larger cousin, *The Blob*. And the small mole on Marilyn Monroe's left cheek breaks the bilateral symmetry of her face.

Broken symmetries have played an important role in the search for the most fundamental physical laws. A good example is the mirror symmetry of P invariance, which seems so obvious, and which leads to the conservation of parity. But in 1956 it was discovered, surprisingly, that a certain physical process (called the weak interaction) sometimes violates the conservation of parity (Sachs 1987 p xi).

And this is not the only case. Two other well-known almost-symmetries are seen to be sometimes broken—namely the reversal of time (called T invariance), and the inversion of electric charge (called C invariance). And so these three most-of-the-time symmetries—C, P and T invariance—are broken in certain cases. And this means that the conservation laws associated with these symmetries are, in those same special cases, violated.

The gradual recognition by physicists of these broken symmetries led to much reconsideration of the so-called 'standard model,' the gathering together of the fundamental physical laws as currently understood. And it was discovered that these broken symmetries lead to a different and more subtle symmetry that (so far) does appear to hold as a fundamental property of nature. And so it seems that a particular combination of C, P and T invariance *does* hold, even though the symmetries of C, P and T are all broken when looked at separately from each other. This new symmetry is called CPT invariance, and so far although many questions remain, it does seem to be one of the fundamental symmetries of nature.

But there is another, even more important sense to the physicist's idea of a broken symmetry. CPT invariance is an example of a symmetry in nature that is deeper (and apparently more accurate) than the imperfect, sometimes-broken symmetries of C, P or T invariance alone. But physicists began to realize in the 1960s that symmetry at the level of a theory can even describe a physical reality that lacks that symmetry (Weinberg 1992 chapter 8). This may seem to be a contradiction, but it is not, and physicists call it *spontaneous symmetry breaking*. For it seems that some of the laws of physics are such—at their most fundamental level—that the theory itself has a simple but deep symmetry. When that theory is *applied in practice* to the real world, the symmetry is broken, and so produces the complex often-asymmetrical reality we experience. The elementary particle physicist Steven Weinberg (Weinberg 1992 p 195) described it like this:

> It is principles of symmetry that give our theories much of their beauty. That is why it was so exciting when elementary particle physicists started to think about spontaneous symmetry breaking in the early 1960s. It suddenly came home to us that there is much more symmetry in the laws of nature than one would guess

Figure 10.7. Left: *Self Portrait with Grid #1*. John Beaver, 2004. Compare this to the left-hand image in figure 10.6. Some of the symmetries present in that image are implied here, partly because of our expectations regarding the grid-like structure. But here those symmetries are broken. Right: *Revolution of the Tear Duct*, John Beaver, 2018. The background illumination has a very simplistic radial symmetry that contrasts with the asymmetry of the shadow forms, and the (literally) broken symmetry of the deformed eyeglasses.

> merely by looking at the properties of elementary particles. Broken symmetry is a very Platonic notion: the reality we observe in our laboratories is only an imperfect reflection of a deeper and more beautiful reality, the reality of the equations that display all the symmetries of the theory.

For the artist, the breaking of an obvious formal symmetry in a picture can be a powerful tool. It *evokes*, raising questions rather than giving simple answers. Some of my own photography makes use of simple lenses, and their pronounced aberrations impose a simple radial symmetry on the image. The seated subject in figure 10.7 (left side) breaks that symmetry, and our expectations of the symmetries of a grid play off its here-distorted depiction, calling attention to the cross-like form to the left of the subject. The image on the right side of figure 10.7 has an even more pronounced radial symmetry; it was exposed by shining a laser through a smear of red blood cells (kindly donated by a colleague). The symmetry of the diffraction pattern is broken by the shadow forms, and the expected bilateral symmetry of the recognizable eyeglasses is broken by their distorted shape.

References

Barrett T 2011 *Making Art: Form and Meaning* (New York: McGraw Hill)

Duby G 1992 *Medieval Art: Europe of the Cathedrals 1140-1280* (Paris, France: Bookking International)

Hale N C 1993 *Abstraction in Art and Nature* (New York: Dover)

Preble D and Preble S 1994 *Artforms* 5th edn (New York: Harper Collins College Publishers)

Sachs R G 1987 *The Physics of Time Reversal* (Chicage, IL: University of Chicago Press)

Weinberg S 1992 *Dreams of a Final Theory* (New York: Pantheon Books)

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Chapter 11

Two-dimensional (2D) design

A common mistake of the beginning photographer is to take the 'capture the instant' aspect of photography too literally. A picture is not really a capture of a moment of your visual experience. Rather, it is a flat surface with marks on it. And they had better be interesting marks, or no one will want to look at your picture.

I find it useful to think of what I call a *visual event*—a moment of your experience in seeing the world. This can be very powerful, and it is a natural impulse to want to capture that with a camera. But *the experience of looking at a photograph is different from the photographer's experience of the visual event that triggered its taking*.

You were there before and after the moment. You experienced motion directly, of either yourself or the subject, or both. You had cues of depth not directly connected to the two-dimensional (2D) image you literally saw. This is not to mention all of the non-visual sensory experience, and the context of your particular state of mind when it happened.

The viewer of your picture has none of this; they have only your picture. If you want them to have an experience like the one you had when you took the picture, you will have to trick them into it. And more commonly, you may want them to have a completely different experience, and there is nothing wrong with that.

Thus, there is an important sense in which every photograph is abstract, just as is every painting. In the end, the artist uses marks on a flat surface to cause things to happen in the mind of the person looking at their picture. But instead of paint applied directly by hand, the photographer uses straight-line rays of light coming from the 3D world and redirected by lenses, in an ordered fashion, onto some light-sensitive material.

This is not to say that a photograph is, in essence, the same as a painting. That a photograph is taken at a particular moment, from light coming from the real world often beyond the artist's control, can be a fact that is of the utmost importance. The knowledge of this inevitably changes the way one looks at a photograph, and thus it changes what one sees.

doi:10.1088/2053-2571/aae1b6ch11

Imagine a parallel universe in which Robert Capa had never taken his famous photograph *Death of a Loyalist Soldier*, and that instead the painter Chuck Close had produced a photorealist painting that happened to look exactly like it in every detail. Would it be the same picture? We explore these issues in more detail in volume 3 of *The Physics and Art of Photography*. Here we discuss some of the properties of a picture that relate only to the formal arrangement of its parts.

11.1 Elements of 2D design

2D design is the intentional arrangements of formal elements in a picture. A picture's 2D design may have a deep connection to its *content*; the interplay between the two may be much of the point. But the design elements can still be considered separately from that content.

First we should agree upon some terminology. There are many different but overlapping ways to choose such a terminology (see for example Ocvirk *et al* 2009, Preble and Preble 1994, Barrett 2011, Brainard 1991), but I list below terms as I will use them in this chapter. And so we can read a picture, or a part of it, as having some combination of these fundamental elements:

Point: A dot, of no particular size or shape.
Line: A line, either curving or straight. This could be simply an edge of a shape, and it may not necessarily be continuous.
Shape: A 2D shape in the geometrical sense, such as a circle or triangle.
Form: A representation of a 3D object, such as a ball or a cone. So a pyramid is a form and a triangle is a shape.
Size: The physical dimensions of a form, line or shape.
Value: The darkness or lightness of a mark.
Texture: The visual effect of a mark. Is it rough or smooth, furry or glassy?
Color: The *hue* and *saturation* of a mark.
Space: The space taken up by objects is called *positive space*, while the space in between objects is called *negative space*.
Depth: The perceived distance from the observer, often separated into *foreground*, *background*, and, sometimes, *middle ground*.

11.2 Figure and ground

A particular form is *figure*. while all that is *not* that form is *ground*. There is a formal sense to this, even when the picture does not represent anything identifiable in the world. So, for example, consider a red triangle on a blue background. Note the *ground* in 'background.' It is difficult to get around the notion that we seem to see a thing in contrast to all that is not that thing.

It is also well known that one person's figure can be another's ground, and vice versa. There are famous optical illusions whereby the figure and ground reverse, depending on the state of mind at the particular moment one sees it. A red circle on a green background can also be a green square with a red hole in it. The most famous example is the 'faces-vase' illusion made popular by the Danish psychologist (Edgar Rubin).

In representational images, and especially in photography, depth is often used to distinguish between figure and ground. Thus a *figure* can be in front of the *background* and behind the *foreground*, although a given photograph may not have all three of these elements. Nevertheless, a common problem in photography is to make a clear distinction between figure, foreground and background. There are many tricks for this, but one we will talk about in detail in chapter 12, section 12.3 is the use of *selective focus*. There are many other depth cues as well, and we discuss some of them in section 11.7.

11.3 Lines

Apparently, our brains will often see a line even when it is not there in any literal sense. Thus, for example, we often speak of a 'dotted line,' which if one thinks about it is something of a contradiction in terms. It is, after all, just a bunch of dots. But if the dots 'lie along a line,' then we see a line, that just so happens to be made of dots. Furthermore a form, such as an arm, may create what one sees as a line in a photograph or painting. Even the edges of a series of unconnected forms can form a line in an image, and the eye/brain will likely see that line (perhaps unconsciously) whether or not that is the artist's intent.

Lines are important because they connect, and thus establish a relationship between, different elements in a picture. Furthermore, our eyes tend to follow them when we look at a picture. Where, then, do the lines lead? Do they lead the eye off the picture entirely (probably an unwise choice), or do they lead from one thing to another in a way that establishes important connections between different forms? And finally, lines can form boundaries, signaling our minds to attach separate meaning to different forms.

Hale (1993) discusses the meaning of line in art, and how the Impressionists recognized that line is an abstract concept that does not really exist in nature. It is, rather, something we impose upon the natural world.

The history of astronomy provides an interesting historical example of the power of lines. In the late 19th and early 20th centuries, the planet Mars was the subject of much interest to astronomers. Photography had not yet advanced enough to be useful under the dire conditions of recording the telescopic view of the tiny planet, and so it was still the reign of the visual observer.

As seen through a telescope, Mars has barely-discernible surface features that appear with very low contrast. Furthermore, the image constantly wavers and shimmers, going repeatedly in and out of focus, due to distortions caused by Earth's atmosphere. Astronomers of the time squinted for hours attempting to map out the most subtle of details.

They saw far more than what was actually there, a fact later proved by close-up photographs from orbiting spacecraft. In particular, many astronomers saw an intricate network of lines on the red planet. This fit in with, and added to, an idea that was popular at the time—that Mars had once been inhabited by an ancient civilization, now perished, that had established a grand network of irrigation canals to bring water from the poles to the more temperate regions.

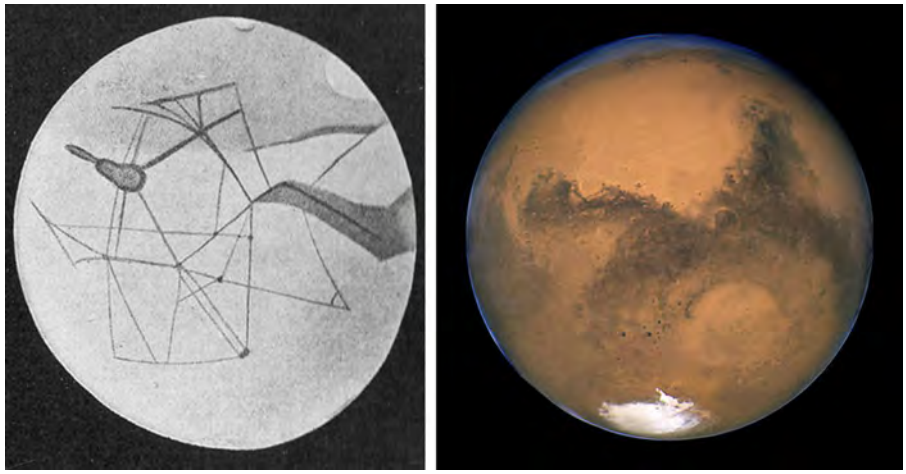

Figure 11.1. On the left is a drawing of Mars made in the early 20th century by Percival Lowell; on the right is a late 20th-century photograph of Mars from the Hubble Space Telescope (NASA, J. Bell (Cornell U.) and M. Wolff (SSI)).

Figure 11.1 shows an early 20th century drawing of Mars made by the astronomer Percival Lowell, along with a late 20th century image from the Hubble space telescope.

11.4 Geometric shapes

Just as we easily see lines, we also naturally see geometric shapes, especially simple ones such as circles, triangles and rectangles. Often a picture can be broken up into a series of geometric shapes independently of the forms they may represent in the world. These shapes have their own formal qualities in addition to the subject of the picture.

11.5 Value and contrast

Marks (lines, shapes, points, etc) may be distinguished by a difference in value from the surroundings. And so a line may be represented by a lower value (darker) linear mark on a higher value (brighter) background. The term *contrast* refers to the difference in value between a mark and its surroundings. A mark of high contrast shows a large difference in value between it and its surroundings, while a mark of very low contrast has a value barely distinguishable from its surroundings.

11.6 Hue and saturation

Hue (for example, is it red or green?) does not tell the whole story of the color of a mark. Even when one allows for differences in value along with hue, there is more to say. That is, one can have two marks that are the same hue and the same value, but have different levels of *saturation*. A pure color has the highest possible saturation. For the case of pigments, one can lower the saturation by adding grey. If the right

Figure 11.2. Left: *Road to Peace*, John Beaver 2007. Most of the features in this picture are delineated by low contrast marks. The exceptions are the figure at the center and the areas of low-value background peeking through the trees. Right: *Waiting for the Jaguar*, John Beaver 2018. The forms in this photogram are set off from the background with high contrast in both value and hue.

shade of grey is chosen, one then obtains a color of the same hue and value, but with 'less color,' and this is what we mean by less saturation.

Elements of a photograph can be distinguished by hue and saturation, as well as value and contrast. And one sees a figure differently if it is against a ground of contrasting hue as opposed to adjacent hues. Furthermore, strongly saturated colors stand in stark contrast to colors of low saturation, even if of the same hue. See figure 11.2.

11.7 Depth cues

A sense of depth can be added to a picture by many different cues. Here are just a few:

- *Atmospheric perspective*, as discussed in chapter 3, section 3.5. Because of scattering of haze in the atmosphere, more distant forms appear at lower contrast and saturation, and may have a more bluish hue.
- The perspective of *converging lines*. Place in your photograph railroad tracks converging into the distance, and a sense of depth is ensured. Your photograph may be mocked as cliché but at least it won't appear 'flat.' See the left-hand picture in figure 11.3.
- Forms may appear more or less distant depending on their *vertical placement* in the picture. This is especially true for pictures of landscapes. Objects closer to the top of the picture usually appear as more distant, while those closer to the bottom of the picture seem to be closer. If there is a horizon line, however, then the opposite is true for the parts of the picture that are above the horizon. See figure 2.4 for an example.

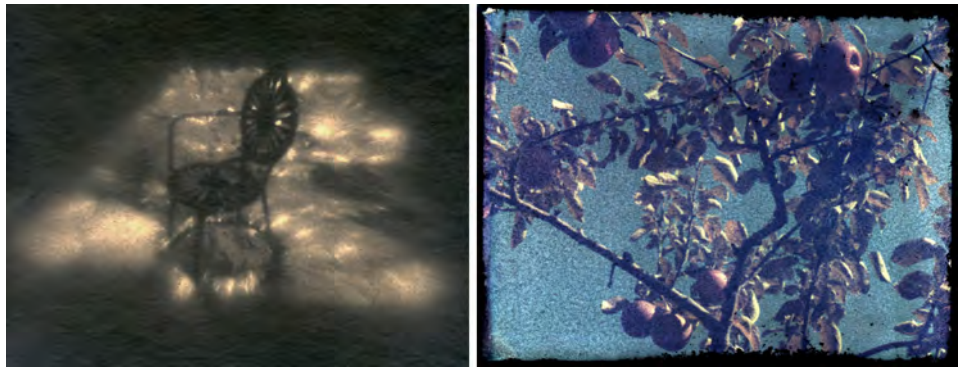

Figure 11.3. Left: *Chair*, John Beaver 2003. The converging lines formed by the edges of the shiny tarp give a sense of depth to the photograph. Right: *Apples*, John Beaver 2007. The sense of depth arises from the overlapping of elements in the image.

- Nearby objects may *overlap* more-distant objects, but not the reverse. This may seem like an obvious statement, but *intentionally allowing figures to overlap* is one way of showing depth. In the picture on the right side of figure 11.3, the sense of depth arises almost entirely from the multiple overlapping of the apples, branches and leaves.

11.8 Unity and repetition

The way design elements relate to each other is an important part of composition. A set of diagonal lines made by one set of forms can be echoed in another form in a different part of the picture, implicitly linking the two. An image has *unity* if all of the formal elements of design are in *harmony* with each other.

One way to establish unity is with *repetition*. When one formal element is repeated in a different context elsewhere in the picture it is a bit like a *canon* in music; one melodic line is altered in some simple way. The altered line is then played at the same time and the two (or more) lines harmonize with each other. Elements in a photograph can do this too. As an example see figure 11.4; diagonal lines are repeated on different scales in the backs of the chairs, the floor decking, the pattern on the shirt and, on a very fine scale, in the hair of the figure.

One can also look for unity in the color palette, something we will consider more fully in volume 2 of *The Physics and Art of Photography*. And one can attribute other design elements to the picture *taken as a whole*. Look again at figure 11.4. The placement of the figure makes the picture rather top-heavy, which gives it a sense of uneasiness. The planking in the bottom two frames seems to form 'supports' for the figure at the top. But the planks are skewed at odd angles, and so it looks as though the whole structure would fall apart. This effect is partly balanced by, and in contrast to, the very stable support of the chairs directly supporting the figure.

Like other aspects of 2D design, there can be too much of a good thing. Many bad photographs lack unity, but the presence of perfect unity does not, by itself,

Figure 11.4. *Self portrait*, John Beaver 2005. The diagonal lines, repeated in the backs of the chairs, the decking, the rail of the deck, and the pattern on the shirt, and the hair of the figure, provides a sense of unity to the photograph. But the regular arrangement of four photographs provides a different kind of unity that contrasts with the diagonal lines, and this adds tension.

make a good picture. In fact, one may intentionally subvert the unity of a picture to add tension to a photograph.

And so in figure 11.4, there is a second kind of repetition in the uniform rectangular arrangement of the four individual image transfers themselves. This is a very stable arrangement, much at odds with the unstable-looking pattern of skewed lines. It does, however, harmonize with the relaxed expression of the figure. And so one might say this photograph, taken as a whole, *lacks* unity. Some elements relate to an angular and unstable, almost vibrating picture, while other design elements are of its opposite, with a sense of regularity, peace and stability.

But there is another way to look at it. Perhaps rather than a lack of unity, this particular photograph has *two different kinds of unity*. It is saying two things at once, and these two ideas play off each other to make a more interesting photograph that evokes conflicting emotions in the viewer. That complexity of interpretation does not necessarily mean that it is a good picture. The point is that one can intentionally subvert the unity of a photograph; but to subvert something, one must understand it.

11.9 Rhythm

A regular, repeated pattern that moves from one part of a picture to another is called *rhythm*. It may be simple and *regular*, such as the boards of a fence or a pattern of bricks on a wall. But rhythm may also be more organic and *flowing*, like a pattern of

Figure 11.5. Some examples of rhythm. Left: the rails of the deck and the planks of the floor have *regular rhythm*. Center: the lines made by the whitecaps have *flowing rhythm*. Right: the birds show *progressive rhythm*. They are separate birds, but they look like multiple images of the same bird, moving from lower-left to upper-right, as time passed.

Figure 11.6. *Celeste with Cat*, John Beaver 2008. The tree on the right side of the picture can either be a vertical form or it can be part of a frame, depending on how the photograph is cropped. The example on the right is simpler, with a clear center of interest. In the more mysterious version on the left, the eye sometimes wanders away from the center of interest, to travel down the path at the right edge of the frame. Which is better?

choppy waves in the ocean. Or instead rhythm can be *progressive*, showing a migration, not necessarily in a straight line, of forms from one place to another. See figure 11.5 for examples of all three.

11.10 Framing

At some point, one must decide what is part of the picture and what is not. And the overall shape of the picture can have a large impact. Is it a long, skinny rectangle or square? If rectangular, is it oriented vertically (portrait) or horizontally (landscape)?

The inclusion, or not, of a single element can drastically affect the composition of a picture. Does that tree on the right side have space to its right, and so looks like a vertical form in the picture? Or instead does the edge of the picture cut through the middle of the tree, and so the tree acts instead as part of a frame? See figure 11.6.

11.11 Composition: some useful rules of thumb

Are there rules of design that should not be broken? Probably no one would make such a strong claim. Many rules, however, do have reasons, and it is probably wise

to understand them. It is often said that a rule cannot be successfully broken until it is first understood.

I would not argue with that, but I would put it differently. There are fundamental principles—facts if you will, as best we understand them. If you do such and such in your picture, then particular things are likely to happen when someone looks at it. Do you want those things to happen or not? Thus we are talking not about rules from an authority, and we either obey or disobey them. Rather, we are trying to understand the consequences of our actions so we can act accordingly.

Note that this list is neither exhaustive nor definitive. Some would include others, but not include some of these, or use different names for similar ideas.

11.11.1 The rule of thirds

A common way to make a boring picture is to either center a single form, or to bisect the frame with forms either vertically or horizontally (or both). Place a flat horizon line right in the middle of a picture where the only strong forms are the bright sky and the dark foreground, and one likely has a dull picture indeed. The rule of thirds proposes dividing the frame into thirds, both horizontally and vertically, with imaginary lines. Then place the forms at the intersections of these lines, thus avoiding bisecting the image.

11.11.2 The rule of odds

An odd number of similar forms in an image is easier to compose than an even number of the same forms. This is because the two outermost forms provide a frame for the symmetric arrangement of inner forms. Put two artichokes in a picture and your brain says, 'Which one do I look at? What am I supposed to do? What's the right answer?' Put three artichokes and your brain sees it as *one symmetrical form* comprised of three elements, and the outer two frame the one between.

11.11.3 The rule of space

For a form that is moving, leave some space in front of it. Otherwise, it will 'move' right out of the picture when one looks at it. If a figure is looking in a particular direction, leave some space toward where they are looking. Otherwise the viewer will follow the figure's eyes right off the picture. Intentionally violating this rule adds an element of tension (see the right-hand side of figure 9.14).

11.11.4 The rule of simplicity

Keep it simple. There should be a *center of interest* in the photograph, a subject if you will. A corollary of this rule is that one should avoid clutter. A common reason camera-phone pictures are often uninteresting is that the subject is lost in a confusing background clutter. What is one supposed to look at?

But on the other hand, there are many successful images for which there is no identifiable subject at any particular location in the picture. In the left-hand image in

Figure 11.7. Left: *Sumac* (John Beaver, 2007). This photograph has no particular center of interest, and so it violates the so-called rule of simplicity. The photograph is more challenging than those that have an obvious center of interest. Right: *The Ties that Bind* (John Beaver, 2013). Diagonal lines move the viewer's eye around the picture more than do horizontal and vertical lines.

figure 11.7, it is the geometrical patterns within the textures, combined with larger shapes arising from differences in hue and value, that form the composition.

11.11.5 The rule of diagonals

Diagonal lines are more interesting than horizontal lines. They add a dynamic element to the picture; in extreme cases they are unsettling. See the right-hand image in figure 11.7.

11.11.6 The rule of triangles

Triangular forms with the base at the bottom add a peaceful and stable quality to an image, while inverted triangular forms can result in a sense of tension and uneasiness in the viewer.

11.11.7 The golden rectangle and the rule of the golden mean

There is a particular proportion of a rectangle that has held fascination for many, for many centuries. The ratio of the lengths of the sides of the *golden rectangle* are such that, if one sections off a square, the remaining section forms a rectangle of the same proportions. See figure 11.8. The small rectangle has the same proportions as the larger rectangle it has been sectioned from by the square.

With a little algebra (see appendix B), it is easy to show that the *golden ratio*—the ratio of the sides of the golden rectangle—must be approximately 1:1.618.

The mathematical elegance of the golden rectangle has inspired some artists to prefer over the rule of thirds, a *rule of the golden mean*. Instead of thirds, we split up the picture plane into segments of 3/8 and 5/8, which approximates the golden ratio ($\frac{5}{8}$: $\frac{3}{8}$ = 1.67 ≈ 1.618). The portion of $\frac{5}{8}$ itself, compared to the whole, also approximates the golden ratio: 1: $\frac{5}{8}$ = 1.60 ≈ 1.618. Since one of the purposes of the rule of thirds is to avoid bisecting the image, the rule of the golden mean accomplishes this as well, but with a different feel.

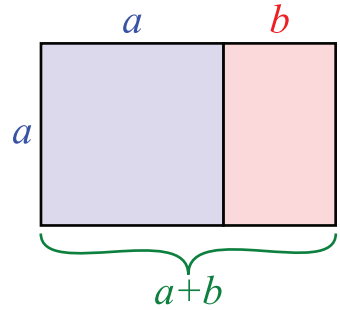

Figure 11.8. A golden rectangle has the dimensions such that, if one sections off a square, another golden rectangle remains. And so in this example $\frac{b}{a} = \frac{a}{a+b}$. This leads to a ratio of approximately 1:1.618 between the sides.

11.12 Some examples of 2D design in photography

So let us look at a few specific examples from only the first few pages of *The Photography Book* (Cooke and Kinneberg 2014). Here I point out only some of the most obvious elements of composition in these photographs; it is a useful exercise to search for additional examples.

11.12.1 *The Lambeth Walk* by Bill Brandt

Notice the converging lines of the apartment building, which provide a strong depth cue. But also, they are echoed in the converging lines formed by the line of faces and the row of arms And so one cannot look at the picture without connecting the girls to the row of working-class tenement houses. This repeated form appears once again in the dancer's blouse.

11.12.2 *Child with Toy Hand Grenade* by Diane Arbus

Converging lines from the tree shadows provide depth. The two vertical forms of the twin tree trunks are echoed, almost, in the single suspender strap, calling attention to the fact that one strap is missing. The two sets of out-of-focus figures in the distance, frame the picture. The placement of the subject does not obey the rule of thirds, but neither is the picture bisected.

11.12.3 *Marilyn Monroe, Hollywood* by Eve Arnold

Forms in the chair back are echoed canonically in the figure. The rule of space is used, giving her room to see in the direction her head is pointing (although her eyes are pointed at us). The line of the arm leads you right to the central figure, as do the triangular shape of the exposed foot and leg brace on the chair. The picture is framed to allow a small amount of space beyond the backdrop, emphasizing the posed nature of the event being photographed.

11.12.4 *Dovina with Elephants* by Richard Avedon

The rule of threes is used to frame the central figure. Violating a standard 'rule' of composition, the body and arms of the central figure bisect the picture plane with a strident, high-contrast form. Symmetrical forms are also used; note the bend in the front legs of the two elephants framing the figure. The line of the white scarf leads right to the neck and head of the figure, which then leads they eye directly to the eye of the elephant. The right-hand elephant is framed so as to add tension to the picture, as it leads off the edge. The chain on the foot forms a line also leading the eye off the picture. The central figure itself forms an inverted triangle, adding instability to that form, but it is made stable by the anchors of the two elephants.

11.12.5 *Andean Boy, Cuzco* by Werner Bischof

One can see use of the rule of space, and an avoidance of bisecting the image (although the rule of thirds is not exactly followed). The position of the legs turns the figure into a stable triangle, and this form is echoed by the triangle made by the flute, hands and back edge of the sack. Differences in contrast are used both to differentiate between figure and ground (both foreground and background), and also to add depth (through atmospheric perspective). The curved pattern of lines in the cloth is repeated in the agricultural terraces in the background.

References

Barrett T 2011 *Making Art: Form and Meaning* (New York: McGraw Hill)

Brainard S 1991 *A Design Manual* (Englewood Cliffs, NJ: Prentice Hall)

Cooke T and Kinneberg C (ed) 2014 *The Photography Book* 2nd edn (London: Phaidon Press Limited)

Hale N C 1993 *Abstraction in Art and Nature* (New York: Dover)

Ocvirk O G, Stinson R E, Wigg P R, Bone R O and Cayton D L 2009 *Art Fundamentals: Theory and Practice* (New York: McGraw Hill)

Preble D and Preble S 1994 *Artforms* 5th edn (New York: Harper Collins College Publishers)

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Chapter 12

The view camera

The lens of the most familiar type of camera has an axis that is centered on, and perpendicular to, the plane of the light detector. One can adjust the distance between the lens and the detector, in order to focus. But that is the only way in which the lens and detector can move relative to each other. A *view camera*, on the other hand has *movements*, adjustments, other than focus, for changing the relation between the lens and the detector.

A common way to accomplish this is to dispense with the dark box altogether. Instead, we mount the lens and detector separately, each on its own *standard*: a *front standard* for the lens and a *rear standard* for the detector. If we mount both standards separately on a common rail, this allows each to separately shift and swivel, while allowing us to move the two standards closer or farther apart in order to focus. We then use a flexible bellows to connect the two and keep out the light.

What I have described is a special type of view camera called a *monorail*, as shown in figure 12.1. Although relatively heavy and bulky, it allows for the greatest flexibility in movements. There are more compact versions, called *field cameras* or *technical cameras*, that fold up for portability, the trade-off being somewhat greater limitations in the movements (see figure 12.2).

The name 'view camera' comes not so much from the movements themselves, but from the way focus is achieved. The lens is allowed to project onto a ground-glass screen mounted in the rear standard. This screen can be examined directly from the outside in order to compose and focus the image (see the right-hand image in figure 12.1). The screen is then replaced by the light detector, and the picture is taken.

Nearly all view cameras have removable lenses, so different focal lengths can be used. Although some have used a single large shutter at the focal plane itself (right in front of the film), most have the shutter mounted right on the lens. This means that each lens has its own separate, built-in shutter, and also its own adjustable iris for changing the focal ratio.

doi:10.1088/2053-2571/aae1b6ch12

Figure 12.1. A monorail view camera that uses 4 × 5 inch sheet film. The lens is attached to a *front standard* while the focus screen and film are mounted on a *rear standard*. Both are attached so they can slide along a rail, and a flexible bellows keeps light out. This arrangement allows for standards that can accomplish a wide variety of *movements* (section 12.1). The image to the right shows the ground-glass focus screen. On this particular monorail the entire focus screen can rotate in order to accommodate both portrait and landscape formats without needing to tilt the camera itself.

Figure 12.2. A vintage German-made Linhof Technika III *technical camera*, that uses 4 × 5 inch sheet film. This particular example was originally made in 1950, but it has had some recent modifications. A technical camera has many, but not all, of the same movements (left image) found on a monorail. Others can be 'faked.' But unlike a monorail, a technical camera folds up (right image) into a small box. A *field camera* (not shown) is similar, but made of wood, somewhat bulkier but lighter and yet more rugged, and usually with more limited movements than a technical camera.

Since the addition of mechanical movements requires a certain amount of physical space to accomplish, a view camera is usually also a *large-format camera*: a camera with a detector that is 4 inch by 5 inch or larger. Typical sizes are 4 × 5, 5 × 7 and 8 × 10; a single frame of 35 mm film, on the other hand, is less than 1 inch by 1.5 inch. Cameras with detectors even larger than 8 × 10 are called ultra-large format[1].

[1] Film formats with sizes in-between 35 mm and 4 × 5 are called *medium format*.

Figure 12.3. Examples of film holders for different formats of sheet film. Clockwise from upper left (all sizes in inches): $3\frac{1}{4} \times 4\frac{1}{4}$ film holder, with sheet film partially inserted; 4×5 holder with film loaded; dark-slide cover for the 4×5 holder; 8×10 film holder with dark slide and sheet of film; a special 'graphmatic' $2\frac{1}{4} \times 3\frac{1}{4}$ holder that can hold six sheets of film, with an internal mechanism to allow rapid shuffling from an exposed sheet to the next unexposed sheet.

Since it is very difficult and expensive to make a digital detector even as large as 35 mm film, large format cameras almost always use traditional film, most typically in the form of individual sheets (rather than rolls).

The sheets of film are most-commonly placed in a two-sided film holder, each side with its own light-tight cover that can be removed once the holder is in the camera, just before taking the picture. The film holders are loaded with film individually in the darkroom. See figure 12.3 for some examples.

A large-format view camera with movements opens up new possibilities and raises new issues:

- Since the picture must be composed on the view screen before the film is even put in place, a view camera is impractical for photographing moving subjects.
- Since sheet film is used instead of roll film, each individual exposure can be processed differently to better control for contrast.
- Since the view screen shows exactly what will be on the film (the screen is removed and the film put in its place), one can very carefully control every detail of the composition of the photograph.
- The camera movements allow one to alter the geometrical perspective of the image. One can control what lines are parallel and what lines converge.
- The camera movements allow one to alter the orientation of the focal plane. The plane of focus no longer must be parallel to the film. It can now be, for example, at an angle stretching from the nearby left to the distant right.
- The elaborate setup and the careful focusing, arranging of movements and composing ensures that one takes fewer photographs, but spends much more

time on each one. Thus use of a view camera emphasizes quality over quantity.
- The larger format view cameras (8 × 10 for example) are so big and heavy that one finds that the most interesting subjects are often within 20 feet of the car.

12.1 Description of movements

A view camera with *full movements* allows for the following adjustments of *both* the front and rear standard (except for *rotation*, which applies only to the back):

- **Focus:** This is the basic movement that most cameras possess, so ubiquitous that is sometimes not called a movement at all. But for a view camera, focus means that the distance between the front and rear standard can be changed. On some view cameras the rear standard is fixed while the front standard can move forward and back, on either a rail or a set of tracks. On others (monorails especially), both standards can be moved independently of each other. Sometimes the tripod mount can be moved independently as well. The focus movement is usually limited in several ways:
  - The monorail or focus tracks are only physically so long. Many monorail view cameras have rails that can be extended.
  - There is usually also a *minimum* distance at which one can physically place the front and rear standard before their mechanisms touch each other.
  - The bellows will only stretch so far, and will only compress so tightly. For this reason many view cameras have interchangeable bellows, to make it adaptable for both very short focus or very long focus lenses.

  In addition, the other movements listed below both affect and are affected by the focus range.
- **Rise:** The standard moves vertically upward.
- **Fall:** The standard moves vertically downward.
- **Shift:** The standard moves to the left (left shift) or right (*right shift*), as seen from behind the camera facing the subject.
- **Swing:** The standard pivots about a vertical axis. As seen from above the camera, a clockwise rotation is called *right swing*, while a counter-clockwise rotation is called *left swing*. Some view cameras are not equipped with swing, but in my opinion it don't mean a thing if it ain't got that swing.
- **Tilt:** The standard pivots about a horizontal axis. If the top of the standard is tilted forward (toward the subject) it is called *forward tilt*. If the top of the standard is tipped rearward, it is called *backward tilt*.
- **Rotation:** The view screen and film holder rotates so as to change the orientation of the rectangular piece of film. Cameras that have this feature are said to have a *rotating back*. In some cases, the back does not fully rotate, but the view screen and film holder can be removed and re-positioned at a 90° angle.

Figure 12.4. Front movements on a 4 × 5 monorail view camera. Left: front rise. Center: front right swing. Right: front forward tilt.

Table 12.1. The possible movements for a view camera. Most view cameras have only some of these movements.

Front rise	Front fall	Front left shift	Front right shift
Rear rise	Rear fall	Rear left shift	Rear right shift
Front right swing	Front left swing	Front forward tilt	Front backward tilt
Rear right swing	Rear left swing	Rear forward tilt	Rear backward tilt
Front focus	Rear focus	Tripod mount movement	Back rotation

See figure 12.4 for illustrations of some common examples of front movements, as demonstrated with a 4 × 5 monorail camera.

The full list of movements can be found in table 12.1. Most view cameras have only a subset of these movements. But it is often the case that two movements can be combined to give the same effect as a third movement. For example, one inch of front rise, while lowering the tripod by one inch, has the same effect as one inch of rear fall. And if one uses both front and rear forward tilt of 15°, while also using the tripod to point the entire camera upward by 15°, it has almost the same overall effect as front rise. Thus many view cameras save on complexity (and thus expense, weight and bulkiness) by choosing a strategic base of movements, with the intention that the rest can be 'faked.'

12.2 Movements and the image circle

The movements of a view camera are necessarily limited by the physical mechanism. The front and rear standards will shift only so many inches, or swing by only so many degrees. But the lens itself provides another constraint. Recall the concept of *vignetting* from chapter 8 in the context of a pinhole camera. Vignetting is a feature of lenses too, although it can be mitigated somewhat by clever lens design. Even so,

every lens has a maximum angle of view, outside of which no light is able to pass through the lens.

Thus the image made by a given lens lies within a circle formed by the intersection of the maximum angle of view and the image plane at the light detector. This is called the *image circle*, and it is an important feature of any lens used for large format photography with a view camera.

For one thing, the diameter of the image circle must be larger than the diagonal of the film format. The diagonal of 4×5 inch film, for example, is (by the Pythagorean theorem) $\sqrt{4^2 + 5^2} \approx 6.4$ inches, and so a lens for this format must have an image circle at least that large. If it does, the lens is said to *cover* the format.

See figure 12.5 for an example of a picture taken with a lens that does *not* cover the format. Notice the clear presence of an image circle, outside of which there is no image at all. But near the edge of the image circle, there is an image but it is darker. And so in this case there is an image circle that does not cover the format, and it shows partial vignetting at the edges of the image circle.

Movements make the issues of coverage and vignetting even more important. A lens that covers a particular piece of film without movements may not do so when movements are employed. As an obvious example, consider the case of a lens that has an image circle only slightly larger than the diagonal dimension of the 4×5 film being used. With no movements, all is well. But clearly, if one employs two inches of rear shift, part of the detector will be moved right out of the image circle.

This is most obvious for rear rise, fall and shift, but *all* movements (even focus) can move part of the film out of the image circle, thus causing vignetting. For this

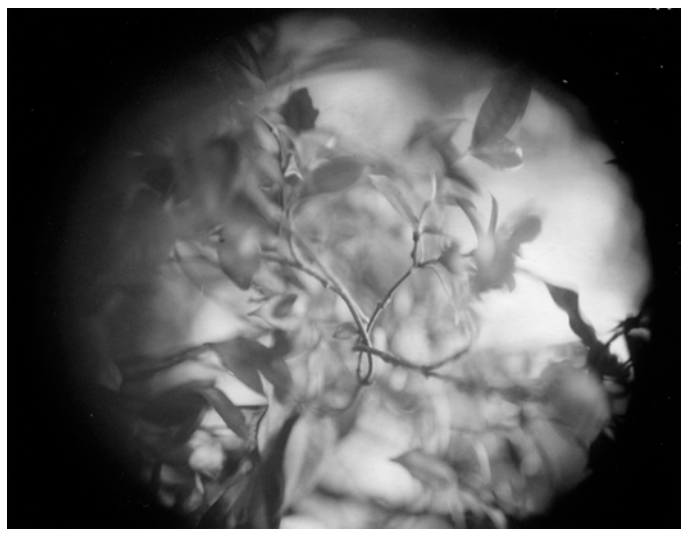

Figure 12.5. An example of a picture taken with a lens that does not *cover* the particular film format used. The image circle is smaller than the film, and the picture shows pronounced vignetting.

Figure 12.6. Views from the clipped corners of a view camera focus screen. On the left the lens is set for a small focal ratio. As seen from one of the clipped corners of the view screen, the lens appears as a crescent. This means that although some light will still get to the corner of the picture, it will be dimmer there than at the center of the image (from where the lens would appear as a full circle). And so the image would cover the entire negative, but it would show vignetting at the corners. In the view on the right, the lens has been stopped down to a relatively large focal ratio, and the circular profile means that the light would be undiminished at the corners in this case, and a picture so taken would show no vignetting.

reason most view cameras have a view screen with *clipped corners*; the corners of the ground glass screen are literally cut off. The focus screen in the right-hand image of figure 12.1 shows an example. When composing the picture the photographer looks through those holes at the lens. If the full aperture of the lens is visible from all four corners, then no vignetting will occur.

This can be a more-or-less thing rather than an either-or thing. For it is possible that, as seen from a particular corner of the film, *some but not all* of the lens aperture is visible. Figure 12.6 shows two views through the clipped corner of a view camera screen, showing partial vignetting and no vignetting.

Partial vignetting most often occurs with large lens apertures, and this can lead to a fading out of the image at the edges. For a very small aperture, the lens is pretty much either blocked or not (as seen from the film). And so vignetting at small apertures tends to give a sharp-edged circular cutoff of the image.

Clearly, in order for a view camera lens to allow for movements without vignetting, its image circle must be significantly larger than the film being used. For a given focal length it is usually more expensive, sometimes considerably more, to produce a lens with a large image circle than with a small image circle. This is especially true for a wide-angle lens, since a wide-angle lens already, by definition, has a large angle of view; to allow for movements the angle of view must be even larger still.

Figure 12.7 shows for comparison two lenses with very different image circles and angles of view. The lens on the left—although it is physically much smaller—has a *larger* aperture by which light enters the camera (it has both a longer focal length and a smaller focal ratio). The lens on the right is physically so much larger because it is better corrected for aberrations and has a much larger image circle and angle of

Figure 12.7. Left: 114 mm f/4.5 Voigtländer Anastigmat Skopar. Right: Schneider 90 mm f/5.6 Super Angulon. Even though the Skopar on the left has an aperture that is physically over 50% larger, the lens is much smaller and lighter. This is because the Super Angulon on the right is better corrected for aberrations, and has a *much larger image circle*. The Skopar was meant to cover film with a diagonal of only 5" (58° angle of view), without movements. The Super Angulon, on the other hand, has an image circle of 9.25" with a 105° angle of view. This allows it to easily cover 4 × 5 inch film with room for substantial movements.

view, allowing for its use as a wide-angle lens for large format with movements. The lens on the left, on the other hand, was intended for only medium format, and with no movements. It takes a lot of glass to get all of those light rays, coming from such a wide range of angles, to their proper places on the film.

12.3 Selective focus

In section 9.3 we considered the use of a large lens opening (small focal ratio) to achieve a shallow depth of focus. This can be useful to ensure that only the subject is in focus, while rendering other parts of the picture blurry. In a traditional camera, this *plane of best focus* is perpendicular to the film, and lies all at the same distance from the camera.

But the use of movements on a view camera introduces new possibilities. Look carefully at the images in figure 12.8. These two photographs were taken with large apertures, and so for each there is a well-defined plane of best focus—an imaginary flat plane out in the world, for which objects are rendered in good focus, while objects on either side of this imaginary plane are rendered in poor focus.

The difference is that the plane of focus for these images is not all at the same distance, and it is not parallel to the film. In the left-hand image, the plane of best focus is vertical, but it runs from nearby on the left (putting the plant in the window in focus, while the distant background is blurry) to the distant right (putting the right side of the window out of focus while the distant trees on the right side are sharp).

A similar technique was used for the right-hand image in figure 12.8, which shows what seems to be an in-focus path going almost straight away from the camera. The

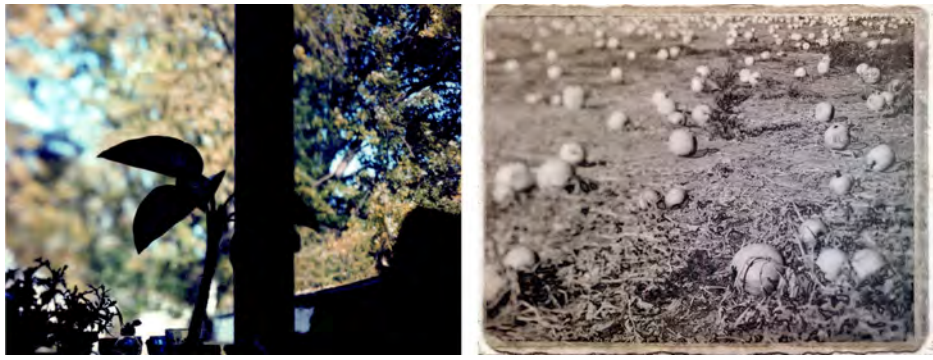

Figure 12.8. Left: *Out the Window*. John Beaver 2008. Front swing allows one to put the foreground in focus on one side of the picture while the background is in focus on the other side. Right: *Creative Nonviolence (for Kathy Kelly)*. John Beaver 2008. I used front swing to make an in-focus path running from lower left to upper right.

Figure 12.9. Left: *Pomegranate*. John Beaver 2006. Front forward tilt was used to place the plane of best focus parallel to, but above, the table. Right: *Mask*. John Beaver 2011. Front swing was used to connect the forehead of the subject to the corresponding part of her angiography mask.

left-hand image in figure 12.9 is more subtle. But close inspection shows that while the table cloth is blurry, the top of the candle and the top of the pomegranate and spoon are in perfect focus. Although the picture was taken at a downward angle, the plane of best focus lies almost parallel to, but slightly above, the table.

These alterations of focus can be done only with a view camera. The trick is to angle the lens, relative to the film. The images in figure 12.8 were produced with front swing, and for the left-hand image in figure 12.9 I used front forward tilt. The geometry of these movements is shown in figure 12.10, and it demonstrates that the plane of the lens lies at an angle somewhere in between the planes of the film and the plane of best focus.

Control of the angle of the plane of best focus opens up many exciting possibilities, and can solve many technical problems. Say for example that one wants to photograph several people in a line, among a lot of clutter. It would be nice to use selective focus to isolate the people from the clutter around them. But without

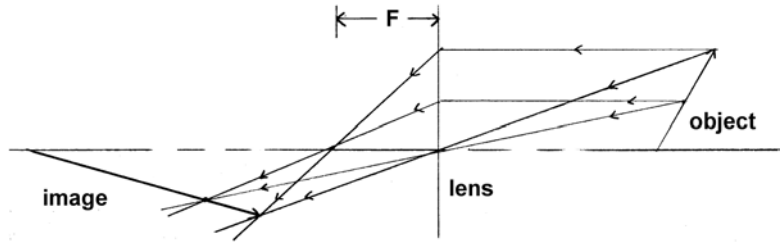

Figure 12.10. If a view camera is used to angle the lens with respect to the film, then the plane of best focus out in the world is also at an angle. In this case the plane of the lens lies at an angle in-between the planes of best focus (marked by the 'object' in the diagram) and the film (marked by the 'image' in the diagram). This technique of *selective focus* can be used to solve technical problems or as an important part of the composition of a photograph (see figures 12.8 and 12.9 for examples).

camera movements, the only way to do this would be to photograph them from directly in front. A large aperture could then put foreground and background objects out of focus.

But what if one wants to compose the picture, for dramatic effect perhaps, from an angle instead? From an angle, without movements, one cannot use selective focus to isolate only the people in focus, since they are all at different distances. With a view camera, however, one simply swings the lens until the plane of best focus lies along the line of people.

Large amounts of front swing or tilt (or both) can sometimes be cleverly used, in conjunction with large apertures for shallow depth of focus, to draw the eye in such a way to focus on two widely separated elements in the photograph, forcing comparison. For the right-hand image in figure 12.9, I used this to draw the viewer's eye to both the forehead of the subject and the in-focus geometric pattern on the part of her angiography mask (for treatment of a brain aneurysm) corresponding to her forehead.

12.4 Controlling perspective

A rectilinear lens takes straight lines in the world and projects them onto the image also as straight lines. But this does not mean that *parallel* lines in the world also appear parallel in the picture. For example, look upward from near a tall building. The sides of the building are (presumably) parallel, but they appear to you as converging off into the distance.

This is just an inevitable consequence of *linear perspective*—the mapping of lines in the three-dimensional (3D) world onto a 2D surface. But without access to all of the depth cues that one unconsciously employs while looking at that building in the real world, a photograph of the same scene can give the impression that the building is falling over backwards. This is yet another example of how the experience of looking at a photograph can be very different from seeing the same scene in real life. Sometimes an 'unaltered' photograph is less realistic than an altered one.

For both of the wide-angle images in figure 12.11 I used ordinary cameras that were incapable of movements. For the image on the left I centered the building, and

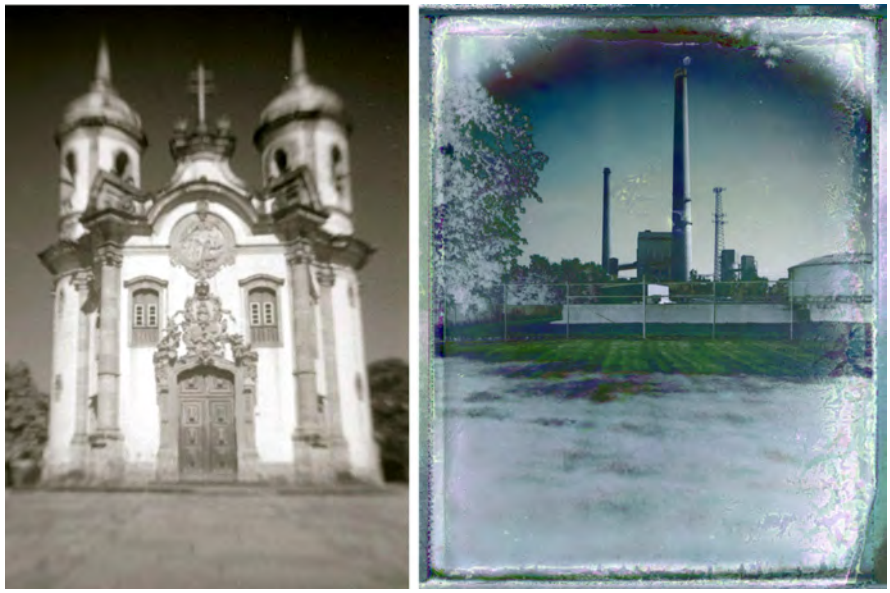

Figure 12.11. Left: *San Francisco*, John Beaver 2012. Right: *Power*, John Beaver 2006. Both of these wide-angle images were taken with cameras incapable of movements. If the tall building is centered (left) it appears as if it is falling over backwards. If on the other hand, the horizon line is centered (right), then the building appears normal, but half of the picture is foreground. Employing rear fall (or front rise) would have allowed me to center the buildings while correcting the perspective.

it looks like it is falling over backwards. For the image on the right I centered the horizon line instead, and so the perspective appears normal. The trade-off is, however, that half of the picture is foreground. That works for this particular composition, but for many pictures it would not.

Thus it is often desirable to alter the linear perspective as seen by the camera, and this can only be done with movements. While front swing and tilt are the preferred ways to alter the angle of the plane of best focus, perspective is best altered by the use of rear-standard movements.

The example of photographing a tall building from close up (with a wide-angle lens) is a textbook case. If one points the camera exactly horizontal, then the parallel vertical sides of the building will also appear parallel on the film. But then the building will only be in the top half of the picture; the bottom half will be foreground (and perhaps the top of the building will be out of the frame). If one tilts the camera upward, to precisely frame the building, then normal linear perspective will force the parallel sides of the building to converge, and the picture will look like a building falling over backwards.

The solution is simple. Point the camera horizontal, so the sides of the building appear parallel, and then simply move the rear standard (and thus the film) vertically until the building is centered in the picture. Since the image is upside down, one must use rear *fall* to recenter the building. If the camera lacks rear movements, then front rise will accomplish the same thing as rear fall. If it matters that the lens is then

positioned a couple of inches higher than it was before, and so it is now seeing from a slightly different vantage point, then simply use the tripod to lower the entire camera by the same amount. Thus front rise, one of the mechanically simplest movements to incorporate into a camera, can be used to solve one of the most common technical problems. As such, if a view camera has only one movement, it is likely to be front rise.

12.4.1 Altering perspective with a pinhole camera

A pinhole camera is an excellent tool for photographic experiments that alter geometrical perspective in unusual ways. It is the simplest of cameras—a dark box with a tiny hole. And since the exposure times are measured in seconds or minutes, a complex mechanical shutter is unnecessary. But also, the most commonly used light detector for a pinhole camera is black-and-white enlarging paper meant for producing prints in the darkroom from black-and-white negatives. Since the light detector is a piece of paper, it can be angled, folded, or curved at will, as it is placed in the pinhole camera. And thus one can do all that a view camera can do, and more, regarding alterations of perspective.

Recall that a pinhole camera does not have a particular best-focus distance, and so whatever the distance between the detector and pinhole, the focus is essentially the same. This means that one can twist and bend and tilt the light-sensitive paper to the heart's content, and all of the image will be in the same focus. For a pinhole camera this is, admittedly, an imperfect focus—but it does mean that one can explore perspective effects independent of effects of selective focus.

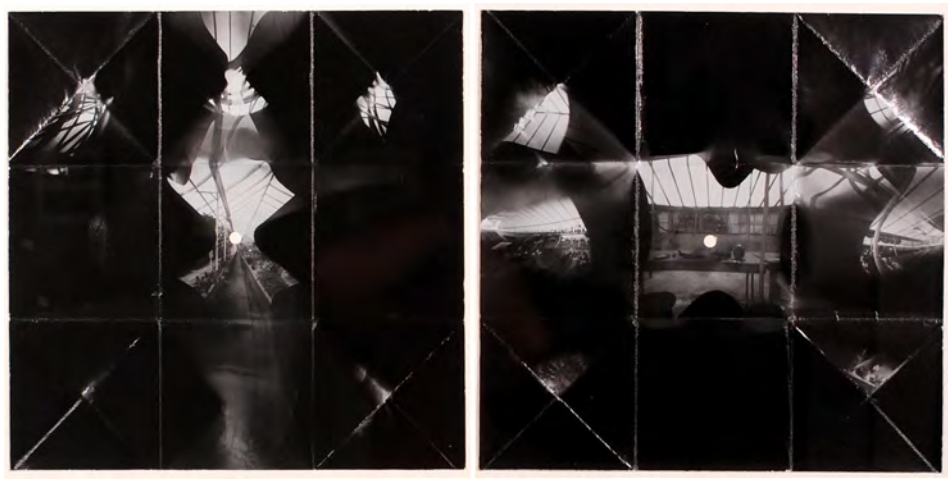

Figure 12.12. *The Greenhouse*. John Beaver 2015. These two pinhole photographs were made simultaneously in opposite directions. Two sheets of light-sensitive paper were folded together, using origami techniques, into a cube. A pinhole was placed on each side of the cube, and placed inside an identical origami cube made from opaque black paper. Once developed and flattened, the images show a perspective that is not possible to see from any particular vantage point.

Whereas the typical use of view-camera movements is to *correct* perspective to a 'natural' look, the pinhole photographer is more likely to take advantage of the everything-goes nature of a pinhole camera to produce an image with a strikingly *un*natural perspective.

For figure 12.12 I attempted the most extreme example I could think of. I took two square sheets of light-sensitive enlarging paper, punched a hole in the center of each sheet, and taped over each hole a small piece of sheet-metal with the proper-size pinhole for a pinhole camera. I then used origami techniques to fold the two sheets of paper into a cube, thus forming a pinhole camera[2]. But it had *two* pinholes, each

Figure 12.13. Top: this ephemeral process pinhole image was made with the light detector curved around the inside of a cylinder. Bottom left: the pinhole camera, with the just-exposed negative in place. The pinhole can be seen as the small piece of brass on the front of the cylinder. Bottom right: the same view as imaged with an ordinary wide-angle lens. The pinhole camera made a wider view, and the straight street seems to bend.

[2] I placed this origami pinhole camera inside a similar origami cube made with opaque black paper, so that light could enter only through the pinholes, and not through the backside of the not-entirely-opaque light-sensitive paper.

forming an image on the opposite side of the camera—and the camera was formed from the light-sensitive paper itself. To make the exposure, I uncovered both pinholes simultaneously, and each made an image on the opposite side of the cube. But since the light-sensitive paper was folded in a complex way, each part of the paper formed a different perspective of the same subject once it was unfolded and flattened.

A more typical example can be seen in figure 12.13, where the detector was curved around the inside of a cylinder, with the pinhole placed opposite. The edges of the picture are closer to the pinhole than they would be if the detector were flat. This means the part of the image at the edges is at a smaller scale than it would be with an ordinary camera—and thus objects in the center seem overly large. If the camera is pointed horizontally, horizontal lines curve around the center, while vertical lines remain straight.

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Appendix A

Make your own photograms

A photogram is a shadow print made by exposing a light-sensitive material to light while it is in contact with opaque or translucent objects. I describe here two photographic processes, suitable for photograms, that are easy, inexpensive, and require no darkroom or other specialized equipment.

A.1 Cyanotype photograms

Cyanotype was one of the earliest photographic processes. It is a *printing out* process; the light sensitive paper turns dark as light interacts with it. No chemical developing agent is used. A liquid sensitizer is brushed by hand onto watercolor paper, and allowed to dry. Once dry, it will turn a deep Prussian blue when exposed to ultraviolet light. The Prussian blue pigment is insoluble, but the unexposed sensitizer will dissolve in water. Thus the remaining unexposed sensitizer can be simply washed away in order to make a permanent image.

Cyanotype is most sensitive only to ultraviolet (UV) light, of wavelengths shorter than 400 nm. Sunlight is the best easily-available source of this light. An ordinary incandescent light bulb emits very little UV, and so such a light source can be used as a 'safelight' in order to work with the sensitizer and sensitized paper both before and after exposure. The procedure is as follows:

1. Prepare the liquid cyanotype sensitizer. It should be a bright yellow–green color.
2. Choose a suitable piece of paper.
3. In subdued incandescent light, coat the paper with the cyanotype sensitizer. Coat only the amount of paper that you expect to use that day. Already-sensitized paper does not keep!
4. Allow the paper to dry in the dark for at least 30 minutes. If it feels cool to the touch, then it is not yet completely dry.

doi:10.1088/2053-2571/aae1b6ch13

5. Assemble your photogram in subdued light. The paper can be taped to a board, and a sheet of glass can be used to hold objects in place.
6. Cover your photogram assembly with something opaque and carry it into the sunlight. When uncovered and exposed to sunlight, the uncovered parts should immediately begin to turn blue. A full exposure, producing the deepest possible blue, will cause the sensitizer to first turn deep blue-green, but then begin to fade to a lighter shade of blue, as if it is being bleached by the sunlight. In direct sunlight on a clear day, expect that you will need to expose your photogram for at least a few minutes or longer (it depends in part on the particular cyanotype sensitizer formula). It could take hours on a day of heavy overcast (wait for a sunny day).
7. Under subdued light, wash the photogram gently in several changes of water. There should be no hint of yellow–green sensitizer left. The blue color will become more intense as the paper dries and the sensitizer oxidizes. If you need instant gratification, put a tiny bit of hydrogen peroxide in your first rinse bath to cause the sensitizer to oxidize immediately. The final result is the same either way.
8. Allow the paper to dry.

With this basic background in mind, further details and advice on preparing the sensitizer solution, choosing, coating and drying the paper—as well as sources for the chemicals (pre-mixed or in powdered form)—are easily available. Detailed descriptions for this and many other techniques can be found in James (2016), and online at http://www.alternativephotography.com. The most comprehensive source on cyanotype is Ware (2016).

A.2 Ephemeral process photograms

Ephemeral process (EP) (Beaver 2017) is in many ways simpler and less expensive even than cyanotype. But unlike cyanotype, EP results in an image that is still sensitive to light. The sensitivity of the finished product, however, is low enough that it can be scanned, and the image captured digitally. Or as an alternative, it can be stored in a light-tight container, and viewed for short periods under subdued lighting. The image can also be chemically 'fixed' to make it permanent. This has both advantages and disadvantages that I consider in the last section.

Ephemeral Process uses black-and-white silver gelatin enlarging paper, intended for use in the darkroom to make permanent prints from projected negatives. Ordinarily, one would expose the paper to very dim light, which produces an invisible *latent* image. This latent image is then chemically amplified with a *developer* to make a visible 'blatant' image. This image would then ordinarily be treated with a chemical *fixer* to render it no longer sensitive to light.

But EP uses this paper in a manner more like cyanotype—it is exposed to enough light that it *prints out* with no use of a chemical developer. Thus, the image becomes visible as exposed, and photograms can be made in very much the same way as

cyanotype. We then leave the image unfixed. And so neither a developer nor a fixer is used.

The trick is that the printing-out process can be greatly enhanced by brushing on a simple *accelerator solution*, rendering the paper temporarily up to hundreds of times more sensitive to light. And so the paper when dry may be much less sensitive than cyanotype, allowing for easy handling in subdued lighting. But when 'accelerated,' it may be rendered many times *more* sensitive than cyanotype. Exposure times in full sunlight may be as little as 10 seconds.

A hand-brushed effect can be easily achieved if the accelerator is brushed on to only part of the paper; the un-brushed parts will remain relatively unaffected by the exposure to light. Once the exposure is finished, the paper is simply taken to dim lighting and washed and dried, rendering it back to its very low sensitivity state. It can then be stored in the dark, viewed briefly in dim light, or scanned to permanently capture a digital image.

A.2.1 Accelerator formulae

So what is this magic 'accelerator?' The mechanics of the accelerating process are described in detail in Volume 3 of *The Physics and Art of Photography*, but the key ingredient is water. In fact water alone has a significant accelerating effect on most enlarging papers. The accelerating effect is, for most papers, greatly enhanced over plain water by adding an oxygen scavenger such as sodium sulfite or ascorbic acid (vitamin C).

I recommend ascorbic acid, as it is safe and can be easily purchased in powdered crystals, sometimes even at the grocery store. Sodium sulfite is also an excellent, inexpensive and easily-available oxygen scavenger, and it is considered to be mostly non-hazardous. It has some advantages over ascorbic acid, but it does cause an allergic response in some people, and direct contact to the skin (or inhalation of the dry powder) should be generally avoided.

Finally, in order to make the accelerator brush more easily onto the paper, I use xanthan gum as a binder. It works better (for this purpose) and is far less expensive than a traditional art-medium binder such as gum arabic. It can be easily found online or in the gluten-free baking section of many grocery stores; the smallest package will last a lifetime for this purpose. My preferred formula is this:

1. Mix together dry:
   (a) 1/8 tsp xanthan gum powder.
   (b) 1/8 tsp ascorbic acid (dry powdered crystals).
2. Mix the dry ingredients with 1/2 cup water and shake well. The xanthan gum will want to form lumps. This can be mitigated somewhat by carefully sprinkling tiny bits onto the surface of the water, shaking, and then repeating the process. But even if lumps form, they should dissipate within 24 h. The accelerator solution should work for at least a couple of weeks.

Ascorbic acid will stain some papers brown, especially if either the exposure time or concentration is too high. The ascorbic acid solution also dries fast, and so it can

be problematic for very long exposures (not a significant issue for most photograms). And it can turn into something like a glue as it dries, and so the acetate (see below) must be carefully peeled off under running water.

Sodium sulfite has none of these problems, but it has the disadvantage that its use is much less benign. It will fog some types of papers, and the citric acid in the recipe below is to counteract that tendency. But like ascorbic acid, citric acid may stain some papers brown. So far, I have not found a paper that both *needs* citric acid to prevent fogging, but also is stained by it. My working sodium sulfite formula is as follows:

1. Mix together dry:
   (a) 1 tsp sodium sulfite powder, Na_2SO_3. Note that this is *not* sodium sulf*ate* (Na_2SO_4).
   (b) 1/8 tsp xanthan gum powder.
   (c) Optional: 1/4 tsp citric acid (dry crystals).
2. Mix the dry ingredients with 1/2 cup water and shake gently. This mixes more easily, with fewer lumps, than the ascorbic acid formula. The large amount of sodium sulfite keeps the xanthan gum particles separated from each other when the water is added.

One should consider these recipes as starting points for experimentation. Some papers require more (or less) of the ascorbic acid or sodium sulfite, and the amount of xanthan gum can be adjusted to make the solution either thicker or more watery. In my experience, the ascorbic acid recipe brushes onto the paper more smoothly than does the sodium sulfite version.

A.2.2 Choosing the paper

There are many varieties and sizes of black and white enlarging paper that can be purchased online, or possibly at a nearby camera store. The cost is usually about $1 per 8 × 10 sheet—less if bought in larger quantities, more if bought in larger sizes. There are two basic categories:

1. Resin coated (RC) papers: The light-sensitive silver gelatin emulsion is coated onto paper that is waterproof, as it is sealed with a plastic resin. This is usually the least expensive type of paper, and it is the easiest to use. But for this purpose it may be less satisfying, as the paper surface has a plastic-like perfection. Furthermore, of the RC papers I have tested, they all are *more* sensitive while dry than the fiber based papers I describe below. But they are, in general, no more sensitive when the accelerator is applied. And so they tend to show less contrast between the unaccelerated and accelerated parts of the paper. RC papers also tend to have a very high unaccelerated sensitivity when the humidity is high, and this means one must be very careful when scanning the still-light-sensitive photogram whenever the humidity is high.
2. Fiber based papers (FB): FB papers are the go-to choice for the art photographer. The silver gelatin emulsion is applied directly to good quality

paper, with all of its subtle micro-texture. It is usually more expensive and difficult to handle than RC paper. For EP photography, I have found FB papers, in general, to be more interesting and useful than RC papers. But the results vary widely from one type of paper to another.

For EP photography, it is not necessary to use newly-purchased enlarging paper. My favorite papers, in fact, have been unavailable for decades outside of the used market. Papers that are long expired and nearly useless for their original purpose may give outstanding results for EP photography. Almost any black and white enlarging paper will produce results that are at least interesting in *some* way. Experiment!

A.2.3 Preparing the photogram

The paper should be handled for as little time as possible, and in light that is just barely bright enough to work under. Incandescent lighting will produce less exposure than daylight, fluorescent or white LED lighting of the same brightness. A *red* LED headlamp, on the other hand, can be used with no fear of exposure at all.

Brush the sensitizer onto the emulsion side of the paper; any type of paintbrush can be used. Gently lay on a thin sheet of acetate, and smooth out the bubbles. In order to determine how the final image will be affected by air bubbles and unevenness in the application of the sensitizer, you will have to experiment. The answer depends on too many details to describe here, but that is much of the fun.

It is important to keep in mind that once the accelerator is applied, the paper will be far more sensitive to light. And so one can be much more casual (sometimes *very* casual) about unwanted exposure *before* the accelerator has been applied. After that, however, one should work both quickly and under light that is as dim as possible.

If you are using FB paper, you may want to tape down the corners of the paper onto a hard, flat surface; FB paper will begin to curl once it is dampened. Lay your objects onto the paper, and if you want them to be held flat, then lay a sheet of glass on top.

A.2.4 Exposing, washing and drying

Unlike cyanotype, enlarging paper is sensitive to blue and violet visible light in addition to ultraviolet. and so EP photograms can be exposed by diffuse daylight, direct sunlight, or artificial lighting. But recall the discussion in chapter 7 on the relation between the light source and the geometry of the photogram. Depending on the choice of paper and accelerator, you may find that different colors of photogram objects, and different sources of light, produce somewhat different *colors* in the printed-out image. This intriguing complication is explored in more detail in volume 2 of *The Physics and Art of Photography*.

If you are to expose your photogram in sunlight, then cover everything with something opaque, and carry it outside for the exposure. If you are using artificial lighting, then it is convenient to have some way to position the light source directly over the photogram in the same dim room you use to prepare it before exposure.

Although cyanotype must be exposed to ultraviolet light, almost any bright light source can be used to expose your EP photogram, so long as it contains some light in the blue-violet part of the spectrum. An ordinary old-style incandescent light bulb is probably the worst choice. Bright photo flood lights work well, both the hot quartz type and the newer and much-easier-to-use cool LED versions. An inexpensive LED 'blacklight' bulb may also work well, even if it does not look very bright to the eye. They emit light mostly at 400 nm, the very wavelength to which the enlarging paper is *most* sensitive. Bright compact fluorescent lights will also work. The shortest exposures will almost certainly be with direct sunlight, whenever it is available.

With a bright source of artificial light, you will likely find that an exposure time of only a minute or two (or even less than a minute) is required. For some combinations of paper and accelerator, only a few seconds may be required when exposed to the direct light of the Sun when it is high overhead on a clear day. Since, like cyanotype, this is a printing out process, you can monitor the exposure as it happens.

Once the exposure is complete, carefully separate, under running water, the acetate from the paper (this is most important with the combination of FB paper and ascorbic acid accelerator). A smooth, clean surface and a shower-stall squeegee can be very helpful for removing most of the water from the paper. RC paper can be simply hung to dry from a corner, with a spring clothespin. FB paper should be left to dry upside down on a clean porous surface (a plastic window screen works well).

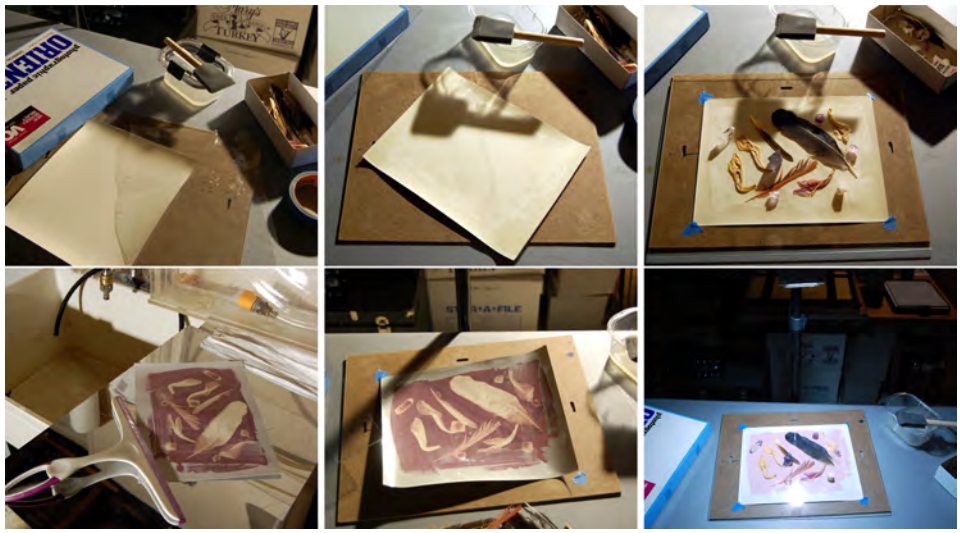

Figure A.1. The EP photogram process. Except for the actual exposure, the entire process was carried out in a room with dim, indirect lighting from a single light bulb. Clockwise from upper left: (1) materials needed. (2) The accelerator is brushed onto the paper, and it is covered with a thin sheet of acetate. (3) Objects are placed on the paper and held down by glass (some objects were placed on top of the glass in this example). (4) The photogram was exposed to an LED photo-floodlight for 90 seconds. (5) The objects are removed. (6) The acetate is carefully removed and the photogram is washed under running water, squeegeed, and dried.

Washing should be carried out in dim light, and the paper should be left to dry in total darkness. See figure A.1 for a step-by-step illustration of the process.

A.2.5 Scanning, and the option of fixing

Once your EP photogram is dry, it will be much less sensitive to light than when the accelerator was applied. But it will still gradually turn dark if exposed to ordinary room light. At what rate this happens depends mostly on the particular type of paper, the brightness and spectrum of the light it is exposed to, and the humidity of the air. But even under the worst of circumstances, a high-quality digital version of the image can be captured with any good-quality scanner.

The original EP image will almost certainly be darkened—damaged if you will—by the light of the scanner. A higher-resolution scan usually means a greater exposure by the scanning light, as does repeated scanning. And so it is prudent to practice first before scanning that one-of-a-kind perfect EP photogram you just made.

One consequence of using a scanner to make a digital image from your photograph is that you then have two other options available. You can 'improve' it by digital manipulation and you can invert it from a negative to a positive. We will consider the aesthetic implications of these choices in volume 3 of *The Physics and Art of Photography*, but see figure A.2 for an example.

It is possible to make your original EP photogram permanent by applying the same chemical fixing process that is used in the darkroom. Instructions for fixing black and white silver gelatin prints are widely available, both online and in print, so I will not go into those details here. For a black and white print in the darkroom, exposed and chemically developed in the way intended by the manufacturer, the fixing process barely alters the image. Not so for EP photography. Results vary widely, but in general, the act of fixing your EP photogram is likely to greatly lighten the image, and the overall color will likely shift to brown. Subtle differences in hue and value that are easily visible in the original, may be lost when it is fixed.

Figure A.2. Left: a digital scan of the finished photogram from figure A.1. Right: the same scan with the values digitally inverted from a negative to a positive and the levels and white balance adjusted.

A fixed photogram made with this same type of enlarging paper, but without an accelerator (and thus with *much* more light) is usually called *lumen process.* And so what I call 'ephemeral process' could also be called 'unfixed, accelerated lumen process.' These and other so-called alternative photographic process are discussed in more detail in volume 3 of *The Physics and Art of Photography*.

References

Beaver J 2017 Using new-antiquarian photographic processes to integrate art and science, *AGU Fall Meeting Abstracts*

James C 2016 *The Book of Alternative Photographic Processes* 3rd edn (Boston, MA: Cengage Learning)

Ware M 2016 *Cyanomicon II–History, Science and Art of Cyanotype: Photographic Printing in Prussian Blue* http://www.mikeware.co.uk/mikeware/downloads.html.

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Appendix B

Notes on the golden rectangle

The lengths of the sides of the golden rectangle are in the proportion of approximately 1:1.618034, a proportion known as the *golden ratio*. It has been claimed that this ratio of lengths is inherently pleasing to the eye, and that a rectangle proportioned so is the most aesthetically pleasing. Part of the fascination with the golden ratio and the golden rectangle arises due to its intriguing mathematical properties.

We can easily work out, with simple algebra, what must be the proportions of such a rectangle. Let us label the rectangle as in figure 11.8, with the short side of the original rectangle a and the long side $a + b$. The ratio of the long to the short side of the original rectangle is $\frac{a+b}{a}$, while the same ratio for the sectioned-off rectangle is $\frac{a}{b}$. Thus, for the two rectangles to have the same proportions, we must have:

$$\frac{a}{b} = \frac{a+b}{a} \tag{B.1}$$

Doing a bit of algebra on this, we have:

$$a^2 = b(a+b) \tag{B.2}$$

$$a^2 - ab - b^2 = 0 \tag{B.3}$$

$$\frac{a^2}{b^2} - \frac{ab}{b^2} - \frac{b^2}{b^2} = 0 \tag{B.4}$$

$$\left(\frac{a}{b}\right)^2 - \left(\frac{a}{b}\right) - 1 = 0 \tag{B.5}$$

doi:10.1088/2053-2571/aae1b6ch14

If we define a new variable, ϕ (the Greek letter phi), to be $\frac{a}{b}$, the ratio of the sides of the golden rectangle, then the last equation becomes:

$$\phi^2 - \phi - 1 = 0 \tag{B.6}$$

This is called a *quadratic equation*, and as such, it has two solutions:

$$\phi = \frac{1 \pm \sqrt{1+4}}{2} \tag{B.7}$$

This gives, to five decimal places, $\phi = 1.618\,034$ or $\phi = -0.618\,034$. The positive root says that the length of the long side of the Golden Rectangle is 1.618 034 times the length of the short side. If you want to look at this the other way, the length of the short side is $\frac{1}{1.618\,034} = 0.618\,034$ times the length of the long side (this is the meaning of the negative root).

It is an odd thing that ϕ is a number for which its inverse is equal to itself minus one. That is, $\frac{1}{\phi} = \phi - 1$; it (and its negative sibling) is the only number that has that peculiar property. In any event 1.618 034, or it's inverse, 0.618 034, is known as the *Golden Ratio* or the *Golden section*.

There are many other peculiar and fascinating mathematical properties of the golden rectangle, but perhaps even more interesting for 2D art are its geometrical properties. If one starts with a golden rectangle, it can be subdivided with a square, and this leaves another golden rectangle. This smaller rectangle can also be subdivided with a square, and it too will leave a golden rectangle, etc.

If one connects the corners of these squares, it makes a special type of spiral, known as the Golden Spiral. This spiral sometimes appears in nature, most directly and famously in the shell of the chambered nautilus; the nautilus and the golden spiral can be seen together in figure B.1.

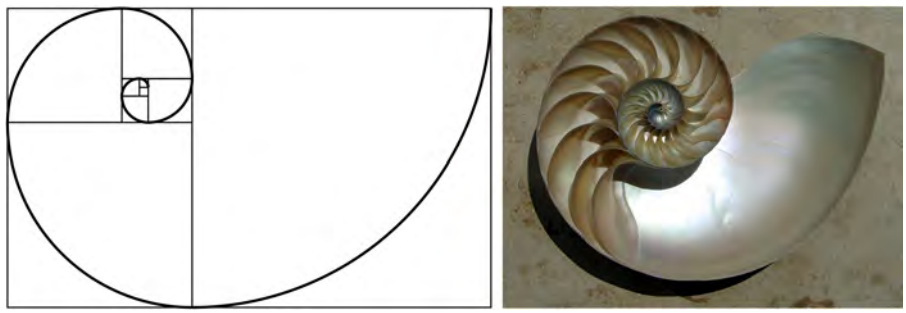

Figure B.1. Left: the golden spiral can be constructed from a series of golden rectangles subdivided into squares and golden rectangles. Right: the shell of the chambered nautilus shows an example of such a form in nature. (graphic: Chris 73, CC BY 3.0).

IOP Concise Physics

Appendix C

Optimal pinhole size for a pinhole camera

A pinhole camera uses simple geometry to restrict rays of light coming from the object to their proper position at the image. This is accomplished with simply a tiny hole on one side of a dark box. The idea is that since it is only a tiny hole, only one ray from each part of the object can get through the hole to the image. There are two problems with this:

1. This could work perfectly only if the hole were infinitesimally small, in which case, only an infinitesimally small amount of light would enter the camera.
2. If the hole is small enough that its size is comparable to the wavelength of the light, then the laws of diffraction dictate that the light will spread out as it goes through the hole. And furthermore, the laws of diffraction say that a *smaller hole* will make the light spread out *more*.

And so we are faced with the fact that the laws of geometry alone would indicate that a smaller hole would give a sharper image. But the laws of diffraction say the opposite. This means that there is a trade-off between these two factors. And so there is a particular size hole—neither too large nor to small—that yields the sharpest image.

Let us see what this best pinhole size would be. Imagine that a point source of light, located very far away, is imaged in our pinhole camera. The perfect image would be a point, since the source of light is a point. But how large would that image be in practice?

A full analysis is beyond the scope of this book, but the simpler analysis I present here gives nearly identical results. It turns out that there is more than one way to define the problem in the first place; when one looks in fine detail, it is not obvious what one means by 'the best.' It is not such a bad thing to have a slightly-vague answer when the question itself is imprecisely defined.

doi:10.1088/2053-2571/aae1b6ch15

One simple way to approach the problem is to add the effects of both geometry and diffraction. For a point source of light very far away, geometry would indicate that the size of the image would be the same size and shape as the pinhole. And so, if our pinhole has a diameter, D, then the image in our pinhole camera would also be a circular spot of that same diameter.

But diffraction plays the opposite role. From chapter 3, section 3.7 we have seen that a smaller hole produces a larger diffracted image. The exact relation, for the *angular* diameter, θ, of the brightest part of the diffracted image, is given by equation (C.1):

$$\theta = 2.44\frac{\lambda}{D} \tag{C.1}$$

where λ is the wavelength of the light. But this is the angular diameter, not the physical diameter, x, (in units of length) on the detector of our pinhole camera. To calculate that, we must use the small-angle formula:

$$x = \theta F = 2.44\frac{\lambda F}{D} \tag{C.2}$$

where F is the focal length, the distance between the pinhole and the light detector. And so, combining all of this and adding the effects of both geometry and diffraction, the size, x, of the image on the light detector would be:

$$x = D + 2.44\frac{F\lambda}{D} \tag{C.3}$$

The question is, for a given value of D, how big is x? And what is the particular value of D that gives the *smallest* value of x? We can approach this in a couple of ways. The most direct and accurate way is to use calculus: finding the value of a variable that gives the minimum (or maximum) value of some quantity is one of the fundamental uses of calculus.

But we can also make a graph of x for different values of D and look for the minimum value. This approach is especially illustrative if we also graph the two parts—geometry and diffraction—separately, to see how they individually affect the sum. One complication is the focal length, F, of our pinhole camera. How might that affect the answer? We will need calculus to answer that question, but we can still use a simple graph to see why there is a best value of D for a particular choice of F.

Figure C.1 shows the size, x, of our pinhole-camera image, for different values of pinhole diameter, D. The three curves show, respectively, the result for geometry alone, diffraction alone and the sum of the two. It is clear from the graph that both effects combined result in a particular value of D that yields the smallest value of x—and thus the sharpest image. For figure C.1 I have chosen a pinhole-camera focal length of 1 m, and for this example, it appears that the best-possible pinhole size (the lowest point in the graph) is a diameter of about 1 mm.

It would be much more useful, however, if we had a formula that would allow us to calculate the best pinhole diameter for *any* given focal length. To find the value of

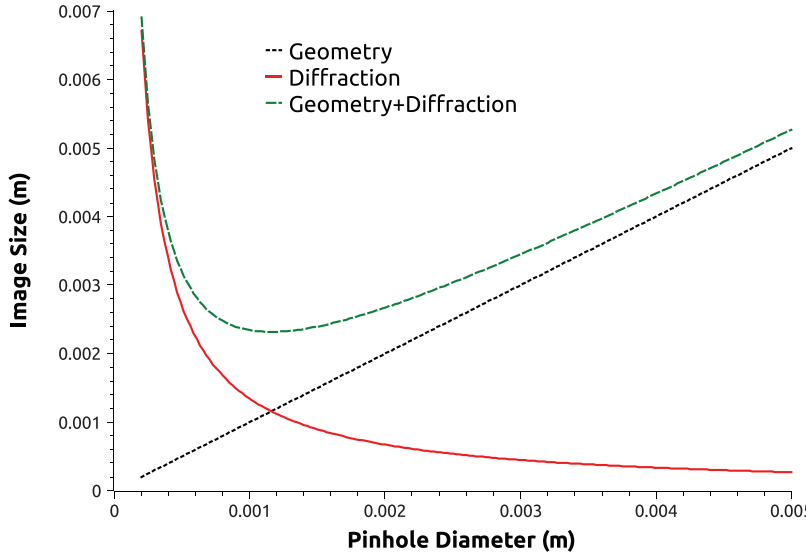

Figure C.1. The two factors, geometry and diffraction, that contribute to the size of the image of a point in a pinhole camera. Geometry alone indicates that a larger pinhole results in a larger (and thus blurrier) image of a point source. Diffraction indicates the opposite; a smaller hole yields a larger and blurrier point image. The two effects combined yield a particular pinhole size that gives the smallest (and thus sharpest) image. For a camera focal length of one meter (assumed for this graph), it is clear from the minimum in the graph that the best pinhole diameter is about 0.001 m, or 1 mm. To determine a formula for the best pinhole diameter for other focal lengths, one must analyze the problem with calculus.

D that gives the minimum value of x, we take the derivative of D with respect to x, dD/dx. We then determine for what value of D is this derivative equal to zero. Taking the derivative of equation (C.3), we find:

$$\frac{dD}{dx} = 1 - \frac{2.44F\lambda}{D^2} \tag{C.4}$$

If we set this derivative equal to zero, and solve for D, we get:

$$1 - \frac{2.44F\lambda}{D^2} = 0 \tag{C.5}$$

$$D = \sqrt{2.44F\lambda} \tag{C.6}$$

If we pick a wavelength of 550 nm $= 5.5 \times 10^{-7}$ m, in the middle of the visible part of the spectrum, and put in a conversion factor to convert from meters to millimeters, then we get for the best value of D:

$$D(\text{mm}) = \sqrt{2.44 \times F \times 5.5 \times 10^{-7}\,\text{m}\left(\frac{1000\ \text{mm}}{1\ \text{m}}\right)} \tag{C.7}$$

$$D(\text{mm}) \approx \frac{\sqrt{F(\text{mm})}}{27} \tag{C.8}$$

We can compare the result of this formula to our graph. Putting in $F = 1000$ mm gives $D = 1.16$ mm

In practice, when making a pinhole it is prudent to err on the side of a larger-than-best-sized hole, rather than a smaller one. A hole that is slightly too large at least has the advantage that it admits more light into the camera, and so the exposure time is correspondingly shorter. This is not the only possible way to analyze this problem, and other approaches yield slightly different formulae. The most sophisticated approach uses a more accurate model of diffraction, and a slightly larger best size results. In practice however, the process of making a pinhole is trial-and-error. And so my advice is to make a pinhole that has a size that is *no smaller than* that given by equation (C.8); your best attempt will likely be slightly larger, and close to the best, best value.

IOP Concise Physics

The Physics and Art of Photography, Volume 1
Geometry and the nature of light
John Beaver

Appendix D

Units, dimensions and scientific notation

D.1 Units and dimensions

When we refer to a physical quantity, it must always have associated with it a set of *dimensions*, and also in many circumstances, a set of *units*.

In this context the word 'dimension' refers not to spatial dimensions, but rather to the *type* of physical quantity. For example, length is a fundamentally different type of quantity than time. One cannot add a length to a time, nor can one subtract one from the other, because that would equal nonsense. Note that this is not the same thing as apples and oranges. Unlike length and time, one *can* add apples and oranges (it equals fruit salad).

But on the other hand, it is just fine to multiply or divide a length by a time. This produces something with different dimensions, that are a combination of the two. For example, if one divides a length by a time, the result is something that has dimensions of length/time ('length per time'). Often these combined dimensions have special names. This example of length/time has the special name of velocity or speed. And so whenever one divides a length by a time, something with dimensions of length/time results.

But what about the actual numbers one plugs into the calculator in a specific case? What if one has a specific length, and a specific time, and wants to calculate a specific speed? Whenever actual numbers are involved, there must also be *units*.

A length of 12.0345 is ambiguous. Is it 12.0345 meters or 12.0345 furlongs? The meter and the furlong are examples of *units*, which are agreed-upon standards for attaching a numerical value to a particular physical quantity. And so the meter is a unit of the dimension of length, and so is a furlong. One can convert between units of the same dimension, by establishing an equivalence between them. And so 1 m = 3.280 feet = 39.37 inches = 0.00497 furlongs, etc.

doi:10.1088/2053-2571/aae1b6ch16

In the physical sciences we mostly use a particular international system of units, called *SI*, which stands for 'International System' (in French). The SI unit of length is the meter, while the SI unit of time is the second. Every SI unit has an official abbreviation. The abbreviation for the meter is *m*, and for the second it is *s* (it matters that they are lower-case). Table D.1 lists some common SI units, with their dimensions and official abbreviations.

Just as we can derive new dimensions by multiplying or dividing dimensions by each other (length/time, for example), we can do the same for units. And so we can divide meters by seconds to get a new derived unit, which we write m s^{-1} (called 'meters per second'). What if we want to divide m s^{-1} by seconds? We can do that just fine, and we get m/s/s = m/s^2 (called 'meters per second squared'). Many of the units in table D.1 are actually derived combinations of other units. For example, the newton is actually a combination of kilograms, meters and seconds:

$$1\ \mathrm{N} = 1\ \mathrm{kg}\ \frac{\mathrm{m}}{\mathrm{s}^2} \tag{D.1}$$

These base units can be modified by any one of a number of official prefixes, which then multiplies the unit by some power of 10. These prefixes and their abbreviations are listed in table D.2, although some are more commonly used than others. For example, 'milli' means '×1/1000'. And so a millimeter (abbreviated mm) is one thousandth of a meter.

D.2 Scientific notation

We have used scientific notation for the values in table D.2. Physical quantities in nature can vary by many powers of 10. And so for example the light given off by the Sun, it's power, *P*, is many times greater than the light given off by a 60 W light bulb:

$$P_{\mathrm{sun}} = 667\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 000 P_{\mathrm{lightbulb}} \tag{D.2}$$

After the 667, there are 24 zeros there. What if I had mistyped (or you miscounted) and you found 23 zeros instead? Well that number would be *ten times*

Table D.1. Common SI units.

Dimension	Unit	Abbreviation
Length	Meter	m
Time	Second	s
Mass	Kilogram	kg
Temperature	Kelvin	K
Force	Newton	N
Energy	Joule	J
Power	Watt	W

Table D.2. Prefixes for SI units.

Prefix	Abbreviation	Meaning
Femto	f	$\times 10^{-15}$
Pico	p	$\times 10^{-12}$
Nano	n	$\times 10^{-9}$
Micro	μ	$\times 10^{-6}$
Milli	m	$\times 10^{-3}$
Centi	c	$\times 10^{-2}$
Deci	d	$\times 10^{-1}$
Hecto	h	$\times 10^{2}$
Kilo	k	$\times 10^{3}$
Mega	M	$\times 10^{6}$
Giga	G	$\times 10^{9}$
Tera	T	$\times 10^{12}$

too small. And so clearly, when dealing with numbers like this, we need a better way. And so we use what is called scientific notation. Written this way, the above equation becomes:

$$P_{\text{sun}} = 6.67 \times 10^{26} P_{\text{lightbulb}} \tag{D.3}$$

The $\times 10^{26}$ part means, $\times 100\,000\,000\,000\,000\,000\,000\,000\,000$. But in practical terms this also means, 'take the decimal point in 6.67, and move it 26 places to the right, filling in with zeros as needed.'

Raising something to a negative power means the same thing as dividing 1 by that same thing, but raised to the same *positive* power. For example:

$$27^{-3} = \frac{1}{27^3} \tag{D.4}$$

And so we can also use negative numbers in scientific notation; it means simply *divide* by the power of 10 instead of multiplying by it. And as with positive powers, we can also express this as a decimal equivalent:

$$3.27 \times 10^{-5} = 3.27 \times \frac{1}{10^5} = \frac{3.27}{10^5} = 0.000\,032\,7 \tag{D.5}$$

Here we can see that 3.27×10^{-5} means, 'take the decimal place in 3.27 and move it five places to the *left*, filling in with zeros as needed'.

This has a couple of advantages. For one thing, we can see at a glance the most important part numerically: how many powers of ten. Secondly, when we write it this way, we don't need the zeros for place holders. And so if I put them there, it means I believe that they are significant.

And so, 6.67×10^{26} and 6.670×10^{26} are not really the same number, although they will both appear the same on a calculator. 6.67×10^{26} could possibly be

6.673×10^{26} or even 6.668×10^{26}. If I do not include any more decimal places, then I am making a statement that, based on my uncertainty in the measurement of that quantity, I have no idea what the value of the next decimal place would be. If on the other hand I write 6.670×10^{26} then I am saying that I believe (even if with some uncertainty) that it really is 6.670×10^{26} and not, say, 6.673×10^{26}.

Note that one *could* use scientific notation to write the same number in several different ways. You should verify for yourself that the following is true:

$$9.75 \times 10^{7} = 975 \times 10^{5} = 0.009\,75 \times 10^{10} = 97\,500\,000\,000 \times 10^{-3} \qquad \text{(D.6)}$$

Clearly, the last two possibilities look a bit silly, but we try to avoid even the second version. When using scientific notation, it is customary to pick whatever power of 10 is needed in order to have one and only one digit to the left of the decimal place.