IOP Concise Physics | A Morgan & Claypool Publication

The Physics and Art of Photography

Energy and color

John Beaver

VOLUME TWO

IOP | ebooks

The Physics and Art of Photography, Volume 2

Energy and color

The Physics and Art of Photography, Volume 2

Energy and color

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

Morgan & Claypool Publishers

ISBN 978-1-64327-344-0 (ebook)
ISBN 978-1-64327-341-9 (print)
ISBN 978-1-64327-342-6 (mobi)

DOI 10.1088/2053-2571/aae504

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 Teresa

Contents

Part II The art and science of color

Appendices

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 basic understanding of physical units, dimensions and scientific notation; a brief review can be found in appendix D. 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). A familiarity with the material in Volume 1, *Geometry and the nature of light*, is not essential prior to reading this volume, but it is helpful.

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 show 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 to learn 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. Practical details of the technique can be found in appendix C.

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 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 ImageJ were used for many of the illustrations. All photographs and illustrations are by the author, except the following:

Figure 9.1: graphic by Vanessaezekowitz at en.wikipedia CC BY-SA 3.0.

Figure 8.10: graphic by Mix321 - Own work, Lewiński, Holak, Biologia dla liceum, wyd. Operon, GFDL,.

Figure 10.4: graphic by Figure design by Kasuga jawiki; vectorization by Editor at Large; 'The dress' modification by Jahobr, CC BY-SA 3.0,.

Figure 11.6: graphic by P.wormer - Own work, CC BY 3.0.

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

Energy and photography

IOP Concise Physics

The Physics and Art of Photography, Volume 2
Energy and color

John Beaver

Chapter 1

The nature of energy

Energy has many forms. It appears that although energy can transform from one form to another, the total amount remains constant as stuff happens. This is called the *conservation of energy*, and it is one of the most important principles in physics.

There is no good one-sentence definition of energy that does not raise more questions than it answers. As with other physical quantities, to be thorough we must use an *operational definition*, describing all of its properties in different circumstances. To do so would require a book of its own; here I give only a brief overview.

There are many types of energy, some of which have familiar names; electrical energy, solar energy, chemical energy and nuclear energy, are examples. These different types of energy manifest themselves in different ways and under different circumstances (and so they have different names). But when it comes down to it, all types of energy can be described as some variety of four *basic* types, of which I give very brief descriptions here:

- **Kinetic energy**—This is the energy due to an object's mass and motion, compared to some other frame of reference.
- **Gravitational energy**—This is energy associated with the force of gravity. A glass gets nudged off the table and then falls to the floor and shatters. The energy to make the glass shatter came at the expense of a decrease in gravitational energy.
- **Electromagnetic energy**—The energy associated with electric and magnetic forces is the basis for solar energy and chemical energy, among many other forms. Since light carries electromagnetic energy, it is of particular importance to photography.
- **Nuclear energy**—This is the energy associated with the enormous forces at work in the nuclei of atoms.

doi:10.1088/2053-2571/aae504ch1

All but the first of these are forms of energy associated with the three fundamental *forces* of nature[1], and I have listed them in order of the relative strength of these forces. This likely fits with your prior knowledge; compare the power[2] output from an old-fashioned water wheel (gravitational energy of the falling water) to typical sources of electrical power. And then compare these to nuclear energy.

In addition to these explicit forms of energy, there is a direct correspondence between energy and *mass*. Albert Einstein discovered, as part of the theory of Special Relativity, that there is a direct equivalence between mass and energy, embodied in the most famous of all equations:

$$E = mc^2 \tag{1.1}$$

Here c is the speed of light, and equation (1.1) says that a tiny bit of mass is equivalent to an enormous amount of energy. And so matter can manifest itself as energy and vice versa. This is one of the key discoveries of modern physics, and it is the foundation for many important phenomena, including the fact that the Sun shines. But it has little *direct* relevance for photography, and so we will not consider it further in *The Physics and Art of Photography*. In most circumstances of interest to photographers, mass acts like mass and energy acts like energy.

Energy can take many forms, and it can change from one form to another (or from one form to several others), but it does not disappear or appear from nothing. From a fundamental standpoint, we now know that there is a connection between this odd principle of the conservation of energy and a basic *symmetry* of the Universe itself. This symmetry is called the symmetry of time translation, and it means that whatever are the fundamental laws of physics, they are the same today as they were last Wednesday and as they will be next Thursday. This and other symmetry principles are now recognized as some of the most important ways for a physicist to look at the Universe. Indeed, physicists often look to symmetry principles in their search for new, as yet undiscovered physical laws, most strikingly in the ongoing search for a *unified field theory* of physics. I point out this somewhat arcane fact just to illustrate that one can find many metaphors for art in science itself; for symmetry is equally important in art as it is in physics.

But in any event, light has energy, and as light travels from place to place, it carries its energy with it. When light does stuff, it is in part because some transfer of energy has occurred. For example, when sunlight is absorbed by a dark surface, the energy carried by the light is transformed to other forms of energy. The light disappears, but the energy it carried does not; it changes to an equal sum of other forms of energy.

[1] The number of fundamental physical forces in nature depends in part on how one counts. There are, for example, really two fundamental types of nuclear forces, and thus two fundamental types of nuclear energy. I have lumped these together as simply ‘nuclear energy.’ Furthermore, most physicists believe that all of these forces should be understand (in ways that are, at present, only partially understood) as just different manifestations of one even more fundamental force.

[2] We will define power precisely in section 2.1.1.

The joule (J) is the official SI unit for energy; the calorie is another familiar energy unit. As an example of how much energy a joule represents, to lift one gallon of milk from the floor to the kitchen counter top requires an increase of gravitational energy of about 40 J.

1.1 Energy transfer

Since the total amount of energy remains constant, whenever some type of energy is seen to increase or decrease, it is because energy was *transferred* from one kind to another. It is when energy transfers from one form to another that stuff happens, and so let us consider some examples of this process. Although there are four *basic* forms of energy, there are many other names to describe common cases of energy transfer that are often complex combinations of the four basic types:

1. **Thermal energy:** The individual atoms and molecules in a gas, liquid or solid are constantly in motion, and so these molecules—each individually—have kinetic energy. Thermal energy is related to the total kinetic energy of these individual motions. The transfer of thermal energy is called *heat*, and if an object increases its thermal energy, it undergoes an increase in *temperature*.
2. **Solar energy:** Light carries electromagnetic energy, and when it arrives at some object, it can transfer some of that energy. In most circumstances, whatever energy is not redirected (by reflection) is transferred to thermal energy.
3. **Electrical energy:** In a conducting material some of the electrons are free to move easily from atom to atom. It is then possible to use electric forces to move these electrons from one part of the conductor to another, in a continuous fashion. Electrons have an electric charge, and the motion of this charge under the influence of electrical forces results in a change in electromagnetic energy.
4. **Hydroelectric energy:** A given volume of water has mass, and so gravity acts upon it. If one allows water to flow downhill, the mass of the water moves closer to the center of the Earth. This is a decrease in gravitational energy, and so other forms of energy must increase. In the normal flow of a stream downhill, this gravitational energy is turned mostly to thermal energy. But if one builds a dam and allows the water to undergo a sudden decrease in height, a turbine-generator can be used to transfer this decrease in gravitational energy to an increase in electrical energy (a form of electromagnetic energy).

   But if one looks at the big picture, there is more going on. How does the water get back up to higher elevations so this process can continue? If water from lower altitudes is to somehow arrive at the higher-elevation source of a river, its gravitational energy must increase—and that means other forms of energy must decrease.

   It is, of course, the natural process of evaporation and precipitation that carries the water to higher elevations. But evaporation requires an increase in thermal energy—and this is provided from the absorption of sunlight. And so in this example, solar energy (electromagnetic energy from the Sun) is transferred to thermal energy, which is ultimately transferred to gravitational energy, which is then transferred by the dam's turbine-generator to electrical energy—another form of electromagnetic energy.

5. **Chemical energy:** Atoms are comprised of positively-charged nuclei and negatively-charged electrons, and it is the electrical force between opposite charges that holds an atom together. But when atoms are placed near each other, the situation is far more complex, and some of the electrons may feel significant electrical attraction to more than one atomic nucleus at a time, as well as repulsion by the electrons of more than one atom.

   And so the rearrangement of the electrons in neighboring atoms results in a change in electromagnetic energy. Whenever atoms rearrange from one configuration to another, the resulting changes in electromagnetic energy are known as *chemical energy*.
6. **Elastic energy:** Certain solid materials can be stretched or compressed, but energy is required to do so. If the material is elastic, that energy can be efficiently transferred to the electromagnetic energy of the inter-atomic and inter-molecular bonds. When the material goes back to its un-stretched (or uncompressed state), that electromagnetic energy stored in those bonds can be transferred to other types of energy.
7. **Potential energy:** Potential energy is a *category*, rather than a specific type of energy. As an example, consider gravitational energy. If I lift a particular weight so it is 1 m higher, its gravitational energy increases. If I then lower it by 1 m, its gravitational energy *decreases by the same amount*. In this particular case, the gravitational energy depends only on the weight and its height. And so a particular change in height (of a particular weight) results in a particular change in gravitational energy *no matter how that change in height came about*.

   And so for example I could raise the weight 12 m, then lower it by 27 m, and then raise it again by 16 m. Since the end result is to raise the weight by 1 m higher than its original position ($12 - 27 + 16 = 1$), the gravitational energy increases by exactly the same amount as if I had simply raised the weight by 1 m in the first place.

   For any case in which a transfer of energy depends on only the initial and final orientation—and not how we got there—we can define a *potential energy*. The particular definition is mathematical and it depends on the details of the physical system being considered. Besides gravitational energy, elastic energy can also be described as a form of potential energy.

Each of these types of energy can be explained in terms of the basic energies described earlier. And these are not the only kinds of energy. Energy is partly a matter of context; one's definition of the total energy of a given situation—what a physicist calls a *system*—depends on the precise definition of that system. But this does not mean that energy is subjective, that anything goes.

Let us consider, as an example, hydroelectric energy. A dam holds back water so that a reservoir forms, the top of which is much higher than the stream below. And thus, as water goes through the dam, it decreases in height by so many meters, h. Since the water moves closer to the center of Earth, its gravitational energy decreases, and the law of conservation of energy says that this must be accompanied by an equivalent increase in other forms of energy. The point is that it is the *transfer* of energy that is meaningful, rather than the quantity of energy. An energy of 137 J could mean anything or it could mean nothing. But the statement, '137 J of electrical energy were transferred to thermal energy,' means that something physical has happened.

Chapter 2

Energy and exposure

When light falls on a light *detector* such as traditional photographic film, a digital detector or the retina of the human eye, energy must be transferred in order for anything to happen; some physical change must occur in the detector, and this requires a transfer of energy. Thus, if no energy is transferred, then the light is not detected.

The amount of energy per unit area that arrives at the detector is related to what photographers call *exposure*. The amount of physical change that the detector undergoes as a result of exposure to light is called *density*. And so density is the effect—it is what makes the picture. And exposure is the cause. In Volume 3 of *The Physics and Art of Photography*, we discuss the *physical* details of the interaction between light and different kinds of detectors more carefully. In this chapter we explore the different factors that affect exposure.

2.1 Defining our terms

We must untangle several interrelated concepts. I have listed them in table 2.1, along with their SI[1] units and the symbols I will use in this book to represent them, their units, and some of the relations between them. We will consider each in detail, and how they relate to each other, in the following sections. But here I will give a brief summary of each.

But before you look at table 2.1, I must give a warning. There is more than one way to lay out these relevant definitions, and the exact terminology that is used in practice depends, unfortunately, on the context and the particular field of study. In particular there is the distinction between the physics of the light itself, and our human perception of that light, a subject we will take up in more detail in Part II of this book. Also, we often define quantities in a practical way that takes into account

[1] SI is the abbreviation for the official, international system of units used in the physical sciences. See appendix D if you are unfamiliar with physical units.

doi:10.1088/2053-2571/aae504ch2

Table 2.1. Summary of physical terms and their SI units, relating to energy, power and flux.

Term	Symbol	SI units	Equivalent to	Relation
Energy	E	joule (J)		
Time	t	second (s)		
Area	A	square meter (m^2)		
Power	P	watt (W)	$J\ s^{-1}$	E/t
Intensity	B	$W\ m^{-2}$	$J\ s^{-1}\ m^{-2}$	P/A
Illuminance	I	$W\ m^{-2}$	$J\ s^{-1}\ m^{-2}$	P/A
Specific intensity	I_0	W/m^2/solid angle	J/s/m^2/solid angle	B/solid angle

the particular technical methods by which we measure them. For now, I concentrate on the physics of the light itself, with no special regard to our perception of it or the technical details of the devices we use to measure it. And for that purpose I choose a terminology that seems to me the most straightforward.

Power, P, is a transfer of energy per unit of time, and it is measured in watts, or joules of energy transferred per second. The total emitted brightness of a source of light is an example of power, and that is why a 100 W light bulb is brighter than a 40 W light bulb[2].

Intensity, B, is the brightness of the light itself as it travels through space. We know that if direct sunlight passes through a tiny window, it has less ability to light a room than if the same light passes through a large window. And so the intensity *of the light itself* (rather than what makes it through the window) is not power, but power per unit area crosswise to the direction the light travels. A more physical name for this basic concept is **energy flux**.

Illuminance, I, is the power per area arriving at a surface. This has the same physical units as intensity, but there is a subtle difference. Intensity is the power per area *crosswise to the direction the light is traveling.* Illuminance, on the other hand, is power per area *of the surface the light arrives at*. These two areas may not be the same because the surface may be tilted with respect to the direction of the light. Thus, illuminance often contains a separate factor that includes this angle.

Specific Intensity, I_0, is the intensity of light traveling along an individual ray. It is an important but subtle (and often confusing) concept. We will use the basic idea in this book, but we will sidestep its mathematical definition by finding other routes to the answers we are looking for.

2.1.1 Power, *P*

Although it is the total energy absorbed, at a particular location on the detector, that determines exposure, it is the *rate* at which energy is transferred that relates to the

[2] I am assuming the two light bulbs are of the same type. We will later consider some of the complexities of the relation between the power of a light source and its *perceived* brightness.

brightness of light. And to complicate matters, we could mean a few different things when we use the term *brightness*:

- The rate at which energy is *emitted* by a source of light.
- The rate at which electromagnetic energy of light is carried through a given area (an open window, for example) as it travels through space.
- The rate at which energy, carried by light, is transferred to some surface—a solar panel, for example.

And so instead of 'brightness,' I will use other terms instead that distinguish between these cases. But for all of these, it is the *rate* of energy transfer, rather than the total energy transferred, that is the important concept.

The rate of transfer of energy is so important that it has its own name—*power*. In SI units *power is the amount of energy, measured in joules, that is transferred per second*. And so, since power is energy per time, the SI units for power are joules/second (J s^{-1}). This has its own name, the watt (W). And so 1 J s^{-1} = 1 W. When a light bulb is labeled with '100 W,' it means that every second, 100 J of electrical energy are transferred to other forms of energy (thermal energy and electromagnetic energy in this case).

There is an equation hidden in here. When we say that power, P, is energy, E, per time, t, it means this:

$$P = \frac{E}{t} \tag{2.1}$$

It also means this:

$$E = Pt \tag{2.2}$$

. . . and this:

$$t = \frac{E}{P} \tag{2.3}$$

These three relations are also true for our official units for these three quantities. A watt is a joule per second. A joule is a watt times one second. And a second is a joule per watt.

Since the brightness of the light relates to power, *a given total amount of energy is transferred only when the light of a particular brightness is absorbed for a particular length of time*. And the product of the two, the power multiplied by the time, tells us the energy transferred, and thus the total physical effect on the light detector.

As an example of power in watts, consider the following. To lift a weight requires one to transform other forms of energy into gravitational energy. One can put the watt into context by imagining at what rate would one have to lift a certain amount of weight by a certain height, in order to transfer energy at a rate of, for example, 100 W?

You may find the answer surprising. Energy transfer at a rate of 100 W is equivalent to raising 10.2 kg by a height of 1 m every second. This is about the same as lifting 10 liters (about 2.7 gallons) of water from the floor to the kitchen counter *every second*.

2.1.2 Intensity, *B*

The total power *emitted* by a source of light can be measured directly in watts. The power emitted by the Sun, for example, is 6.67×10^{28} W. And so one can say 'the Sun has an emitted power of 6.67×10^{28} W.' But what about the *intensity* of sunlight passing through a window on a sunny day here on planet Earth, nearly 100 million miles away?

The *intensity of light passing through space* and the power of a *source* of light are related, but somewhat different, concepts. Emitted power is energy per time. But the *intensity* of the light itself is an *energy flux*. A flux is a flow of something *through a surface*. And so the energy flux of light is the rate at which energy travels *crosswise to some (maybe imaginary) surface*, and it is measured *per surface area*.

For example, as sunlight streams through an open window, one could calculate how many joules of energy are carried by light through the window per second, and that would be a calculation of power, in watts. But clearly, the answer one gets would depend on the size of the window, even if it is the same sunlight. And so the intensity of the light itself is not just energy per time, it is energy per time per area, and is thus measured in *watts per square meter* (W m^{-2}).

A photo-voltaic solar panel provides an excellent example; it converts the energy carried by sunlight directly into electrical energy. The sunlight itself, on a particular day, carries a certain energy flux—a certain number of joules per second that it can transfer *per square meter*. And so the power that is intercepted by the solar panel depends directly on the number of square meters of sunlight that it intercepts. If the sunlight has an intensity—an energy flux—of 1000 W m^{-2}, then one needs at least one square meter[3] of solar panel in order to transfer energy at rate of 1000 W.

Go to Calacali, Ecuador on the first day of Spring, at local noon, and the Sun is directly overhead. On a clear day the intensity of this sunlight is about 1300 W m^{-2}. And so place a 1 m^2 piece of cardboard horizontally on the ground, and it will intercept 1300 W of power from the sunlight. Place a bigger piece of cardboard, say 2 m on each side, in the same place and it will intercept proportionally more power. Since this new piece of cardboard would have an *area* of $2 \times 2 = 4$ m^2, it would intercept $1300 \times 4 = 6700$ W of power. And so there is a simple relation between the intensity, B, of the light, and the power, P, it carries through a given area, A. Equation (2.4) shows this relation.

$$P = B \times A \tag{2.4}$$

2.1.3 Illuminance, *I*

Light has a direction, and the number of watts of power that arrive at a given surface depends not only on the intensity of the light, but also at what angle the light strikes

[3] In practice, not all of the energy is transferred to electrical energy—some is always transferred to other forms as well.

that surface. And so there is yet another distinction we must make: there is a difference between the intensity of the light itself—how many watts per square meter that are carried by the light in whatever direction it is traveling, and the number of watts per square meter that actually strike a particular surface the light is falling on. We call the former intensity, while we call the latter *illuminance*.

A key fact in the previous example of intensity is that the thought experiment is performed on the equator, and so the noon-time Sun is directly over head, and the sunlight travels in a direction that is normal (perpendicular in every way) to the piece of cardboard. For comparison, let us do this same thought experiment in Suring, Wisconsin, at a latitude of 45°. At local noon on the first day of Spring, the Sun is not straight overhead; rather, it makes a 45° angle with the horizon. It is, however, still the same sunlight (assuming the air is just as clear as in Calacali), and so it still has an energy flux of 1300 W m^{-2}.

Nevertheless, a 1 m^2 piece of cardboard placed horizontally on the ground in Suring will intercept *less* than 1300 W of power from that sunlight. For the sunlight carries 1300 W of power only through every 1 m^2 surface that is *normal to the direction the light is traveling*. When striking a surface at an angle, this 1300 W is spread out over an area *greater* than 1 m^2. And so the energy flux striking the piece of cardboard is *less* than 1300 W m^{-2}, simply because it is 1300 W divided by an area greater than 1 m^2. See figure 2.1.

Simple trigonometry shows that the relation between the energy flux of the light itself and the energy flux intercepted by a surface oriented with its normal at an angle θ to the direction of the light is given by a simple cosine. This fact, combined with equation (2.4), gives us equation (2.5) for the power actually intercepted by a surface, even if it is not oriented perpendicular to the direction the light is traveling.

$$I = B\cos\theta \tag{2.5}$$

$$\mathrm{P(transferred)} = I \times A \tag{2.6}$$

$$\mathrm{P(transferred)} = BA\cos\theta \tag{2.7}$$

And so the 1 m^2 piece of cardboard in Suring would intercept only 1300 $\mathrm{Wm^{-2}} \times 1\ \mathrm{m^2} \times \cos 45° = 919$ W.

2.1.4 Specific intensity, I_0

With the exception of parts of our image that may be unresolved (such as stars in the night time sky), *the brightness of the light along a particular ray from the surface of something in the world is what it is*. And so the surface of the Sun is blindingly bright, while the surface of black asphalt in the moonlight is very dim. Moving closer to the blacktop will not make its surface appear brighter, and going to Mars so you are farther from the Sun will not make the Sun's visible *surface* appear any less bright. We refer to the inherent brightness of a ray of light coming from a particular surface as the *specific intensity* of the light ray.

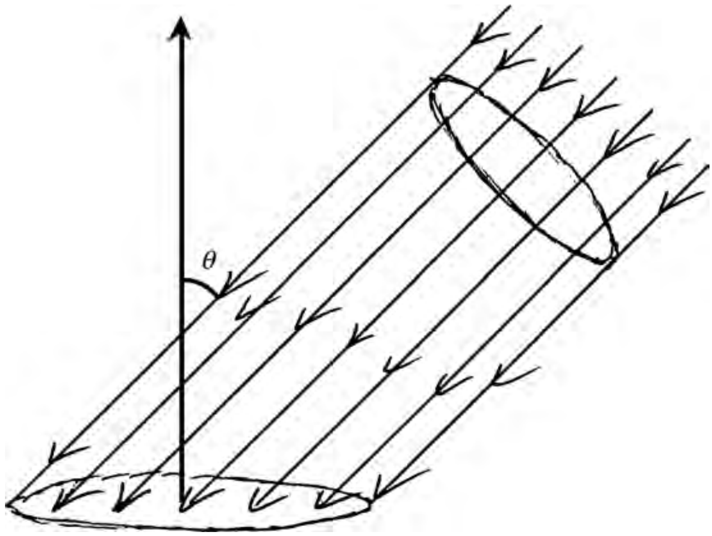

Figure 2.1. An example of the relation between the transfer of power and angle. Light with an intensity of 1300 W m^{-2} strikes a surface at an angle. The energy flux of 1300 W m^{-2} means that the light carries 1300 W of power through a 1 m^2 surface oriented normal to its direction of travel. But that 1300 W is then spread out over a *greater* area when it strikes a surface at an angle. And thus the power transferred is *less* than 1300 W for every square meter of surface.

Technically, specific intensity is an *energy flux per solid angle*[4]. In *The Physics and Art of Photography* we will not consider in detail how to calculate specific intensity numerically. But we will use the concept; it is the appropriate term to describe the 'brightness' of an individual ray of light coming from a particular point on the visible surface of some object.

Intensity decreases with distance as rays of light spread out from some source. But the *specific intensity* along an individual ray stays the same. Similarly, the apparent *total* brightness of a source of light decreases as one moves farther away from it. But its *surface brightness* remains the same. It is this surface brightness of the object that is recorded by the camera, so long as the object is not so small or so distant that it appears as only a point of light.

There is an important exception to the statement that the specific intensity of a ray of light is constant, whatever the distance. For if the ray of light passes through a diffuse, absorbing medium—a cloud of smoke for example—then the medium itself can remove energy from the light ray as it passes through, deflecting it off to other directions or converting it to thermal energy. If that happens, then the specific intensity of the ray becomes smaller and smaller as it passes through the absorbing medium.

[4] A solid angle is related to area at the end of a cone, whereas a 'normal' angle is related to the length at the end of a triangle.

2.2 Tracing the energy from source to camera

And so let us use these terms in a particular example, tracing the light all of the way from its source to its interaction with the detector of a camera. Let us consider light emitted by the Sun arriving at Earth, striking a 2.5 cm diameter flat gray disk at an angle of 37° to its normal, with 18% of the light reflecting off the disk, some of which travels toward the camera, which is 3.5 m directly in front of the disk. This light is then focused by the camera onto the detector with a 135 mm, *f*/5.6 lens, for an exposure of 1/60 of a second. **How many joules of energy per unit area strike the detector?** The answer to this question is important because it is the physical basis of *exposure*.

We will determine this answer by going step-by-step, following the light, and using our basic relations between power, area, time, intensity and illuminance. Along the way, we will have to think some more about how light reflects off of objects.

2.2.1 The power of light emitted by the Sun

The Sun *emits* light with a *power*, P, of 3.85×10^{26} W, or 3.85×10^{26} J of energy every second.

2.2.2 The intensity of sunlight at Earth

The 3.85×10^{26} W of light emitted by the Sun spreads out equally in all directions. And so by the time it arrives here at Earth, it has spread out over a sphere with a huge radius equal to the Sun–Earth distance, d, of nearly a hundred million miles (1.50×10^{11} m). The *intensity or energy flux*, B, of the light that arrives here is thus equal to that enormous power—but divided by an equally-enormous area. And so we are left with a modest *intensity* of sunlight, here at Earth:

$$B = P/A \tag{2.8}$$

$$A = 4\pi d^2 \tag{2.9}$$

$$B = \frac{P}{4\pi d^2} \tag{2.10}$$

$$= \frac{3.85 \times 10^{26}\ \text{W}}{4 \times 3.141\,59 \times (1.50 \times 10^{11}\ \text{m})^2} \tag{2.11}$$

$$= 1362\ \text{W m}^{-2} \tag{2.12}$$

Here we have used the fact that the surface area of a sphere is 4π times the square of its radius. Thus the intensity of light arriving at Earth, and so also at our gray metal disk, is 1362 W m^{-2}

The inverse square law
We derived equation (2.10) for our special case of sunlight reaching Earth. But it should be clear that this equation would apply equally to any case in which a single source of power, P, emits light *isotropically*—uniformly in all directions. Equation (2.10) gives the intensity of light at a distance, d, from any such *isotropic emitter*.

This is called the *inverse square law*, and it is easy to see why; the intensity varies with the inverse of the square of the distance from the source of light. And so if one were to move three times farther away, the intensity would not reduce to one third as bright, but rather to only $1/3^2 = 1/9$ as bright.

2.2.3 Illuminance of the light on the subject

Since sunlight of intensity $B = 1360 \text{ W m}^{-2}$ strikes the disk at an angle of $\theta = 35°$ to its normal, then the illuminance on the disk is given by:

$$I = B \times \cos\theta \tag{2.13}$$

$$I = 1362 \text{ W m}^{-2} \times \cos(35°) \tag{2.14}$$

$$I = 1115 \text{ W m}^{-2} \tag{2.15}$$

And so the illuminance of sunlight on our metal disk is only 1115 W m^{-2}.

2.2.4 The power of the light intercepted by the metal disk

The power, P, of light intercepted by the metal disk is given by,

$$P = I \times A \tag{2.16}$$

where I is the illuminance on the disk and A is the surface area of the 2.5 cm-diameter disk. And so:

$$P = I \times A \tag{2.17}$$

$$= I \times \pi r^2 \tag{2.18}$$

$$= I \times \pi \left(\frac{D}{2}\right)^2 \tag{2.19}$$

$$= 1115 \text{ W m}^{-2} \times 3.141\,596 \times \left(\frac{0.025 \text{ m}}{2}\right)^2 \tag{2.20}$$

$$= 0.5475 \text{ W} \tag{2.21}$$

And so 0.5475 W of sunlight are intercepted by the metal disk.

2.2.5 The power of the light reflected by the metal disk

The disk does not reflect all of the light that it intercepts; a certain percentage is absorbed. In this example, we have chosen a disk that reflects only 18% of the light that falls on it. And so instead of reflecting 0.5475 W, it only reflects 18% of that, or 0.098 55 W. Incidentally, a reflectivity of 18% is considered by photographers to be a mid-tone gray.

2.2.6 The intensity of the reflected light when it reaches the camera

Here is the complicated part. If the disk reflects a total of 0.098 55 W of light, how much of that is directed towards the camera, and what is its intensity by the time it reaches the camera 3.5 m away?

To answer this question we must make some assumptions about how the partially-reflective surface of the disk spreads out the reflected light in different directions. Obviously, one can imagine different possibilities. If it were a mirror, it would send all of the light in the same direction, according to the law of reflection for mirrors. We instead assume here that it is a *diffuse reflection*; each ray of light striking the disk is spread out into rays reflecting in all directions.

The simplest possibility to calculate is if the disk were to reflect light with an intensity that is equal in all directions, what we could call an *isotropic reflector*. In that case, the 0.098 55 W reflected by the disk would, by the time it reaches the camera a distance $d = 3.5$ m away, spread out uniformly over one half of a sphere of that radius. And since intensity is power per area, we would have:

$$B = \frac{P}{A} \tag{2.22}$$

$$= \frac{P}{2\pi d^2} \tag{2.23}$$

$$= \frac{0.098\ 55\ \text{W}}{2 \times 3.141\ 59 \times (3.5\ \text{m})^2} \tag{2.24}$$

$$B = 1.280 \times 10^{-3}\text{W m}^{-2} \tag{2.25}$$

Here I have used the fact that the area of a hemisphere of radius d is given by $2\pi d^2$. And so if our metal disk were to produce a diffuse reflection in this particular way (isotropic), the intensity of the reflected light, by the time it reached the camera, would be 1.280×10^{-3} W m^{-2}.

The problem is that this assumption is very unrealistic, and it flies in the face of everyday experience. Consider that as seen from a sharp angle, the flat disk would appear as only a thin sliver. And so for it to reflect the same amount of light at a sharp angle as it does straight forward, it would have to *appear* much *brighter* per surface area, as seen from that angle. Experience shows that this is not true for most objects that have a matte finish—flat gray paint for example. Instead, the *surface* *brightness* appears approximately the same from all angles. This type of reflecting surface is called a *Lambertian reflector*, after Johann Heinrich Lambert, who in the 18th century first described its properties mathematically.

Since the more-realistic Lambertian reflector appears the same brightness *per surface area* in all directions, it must send more light directly forward than it does at an angle. And this means that by making an assumption of an isotropic reflector, we must have *under*-estimated the amount of light that is sent directly forward. And so our calculation of 1.280×10^{-3} W must be too small.

The question is, too small by what factor? Calculus is required to determine that answer, and I show how to calculate it correctly for a Lambertian reflector in appendix A. The answer turns out to be simple: *the Lambertian reflector sends twice as much light as the isotropic reflector in the forward direction.* And so to correct our calculation, we need simply to double it. And thus the light arrives at our camera with an intensity of $B = 2.561 \times 10^{-3}$ W m^{-2}.

Another way to distinguish between the (very unrealistic) isotropic reflector and the (much-more-realistic) Lambertian reflector is as follows. The isotropic reflector reflects light with the same *intensity* in all directions. But in order to do this, it would have to send light off at angles with a greater *specific intensity*. The Lambertian reflector sends off light at all angles with the same *specific intensity*. The consequence is that it sends light off at an angle with a smaller *intensity*.

2.2.7 The power of the light that enters the camera lens

The light arriving at the camera passes through the aperture of the lens. The power, P, that makes it through the aperture is simply given by,

$$P = B \times A \tag{2.26}$$

where B is the intensity of the light at the camera lens (2.561×10^{-3} W m^{-2}), and A is the area of the circular lens opening. This is a simple calculation once we know A. But we are not given that—only that it is a lens of 135 mm focal length, with a focal ratio of *f*/5.6. But the focal ratio, f, of a lens is simply its focal length, F, divided by its diameter, D:

$$f = \frac{F}{D} \tag{2.27}$$

$$D = \frac{F}{f} \tag{2.28}$$

$$A = \pi\left(\frac{D}{2}\right)^2 \tag{2.29}$$

$$A = \pi\left(\frac{F}{2f}\right)^2 \tag{2.30}$$

$$A = 3.1416\left(\frac{0.135\ \text{m}}{2 \times 5.6}\right)^2 \tag{2.31}$$

$$A = 4.564 \times 10^{-4}\ \text{m}^2 \tag{2.32}$$

And so the power that makes it through our camera aperture is then:

$$P = B \times A \tag{2.33}$$

$$P = 2.561 \times 10^{-3}\ \text{W m}^{-2} \times 4.564 \times 10^{-4}\ \text{m}^2 \tag{2.34}$$

$$P = 1.169 \times 10^{-6}\ \text{W} \tag{2.35}$$

And so 1.169×10^{-6} W of power makes it through our camera lens.

2.2.8 The illuminance of the light on the camera detector

The light that reflected off our gray metal disk, traveled to the camera and passed through the aperture of the lens has a power of 1.169×10^{-6} W. But to determine the exposure we need the illuminance of this light as it is focused onto the detector. This light does not spread out uniformly; rather it is focused onto the detector by the lens (or so we assume).

The lens will focus this light to a circular image, and if we know the area, A, of that image, we can easily calculate the illuminance. If the lens is focused properly, then *all* of that $P = 1.169 \times 10^{-6}$ W spreads out over only the area, A, of the image. And so we have:

$$I = \frac{P}{A} \tag{2.36}$$

But what is the area of the circular image on the detector? To determine this we must use the relation between the size of an image focused by a lens, the distance to the object, and the focal length of the lens. If we let D_O and D_I represent, respectively, the distances of the object and the image from the lens, and S_O and S_I represent the sizes of the object and image, then simple geometry gives the following:

$$\frac{S_I}{S_O} = \frac{D_I}{D_O} \tag{2.37}$$

$$S_I = S_O\frac{D_I}{D_O} \tag{2.38}$$

The size of the image is what we want; the area is then simple to calculate. We have already said that the camera is 3.5 m from the metal disk, so we know that $D_O = 3.5$ m. But what about D_I? How far is the image from the lens? There is a formula to exactly calculate the image distance, but we don't really need that. If the object is located much farther away than the focal length of the lens, then the image distance very closely equals the focal length of the lens, F (135 mm in our example). And so we have:

$$S_I \approx S_O \frac{F}{D_O} \tag{2.39}$$

$$S_I \approx 0.025\ \text{m} \frac{0.135\ \text{m}}{3.5\ \text{m}} \tag{2.40}$$

$$S_I \approx 0.009\,643\ \text{m} \tag{2.41}$$

$$A = \pi \left(\frac{S_I}{2} \right)^2 \tag{2.42}$$

$$A = 3.1416 \left(\frac{0.009\,653\ \text{m}}{2} \right)^2 \tag{2.43}$$

$$A = 7.303 \times 10^{-7}\ \text{m}^2 \tag{2.44}$$

Now that we know the area of the image on the detector, we can calculate the illuminance:

$$I = \frac{P}{A} \tag{2.45}$$

$$I = \frac{1.169 \times 10^{-6}\ \text{W}}{7.303 \times 10^{-7}\ \text{m}^2} \tag{2.46}$$

$$I = 1.600\ \text{W m}^{-2} \tag{2.47}$$

And so the image on the detector has an illuminance of 1.600 W m^{-2}.

2.2.9 The exposure imparted to the detector

Now that we know the illuminance of the image on the detector, all that is left to calculate the exposure is to realize that what we want is not the *power* per square meter, but rather the total energy, in joules, of light that strikes the image per square meter. Since $E = Pt$, we simply have the following basic relation between exposure, illuminance and time:

$$\text{exposure} = \text{illuminance} \times \text{time} \tag{2.48}$$

And so, since we stipulated that the light is exposed to the camera detector for only 1/60 s, we have an exposure:

$$\text{exposure} = 1.600\ \text{W m}^{-2} \times (1/60)\ \text{s} \tag{2.49}$$

$$= 1.600\ \text{J s}^{-1}\ \text{m}^{-2} \times (1/60)\ \text{s} \tag{2.50}$$

$$= 96.0\ \text{J m}^{-2} \tag{2.51}$$

And so, the image on the detector receives 96.0 J m^{-2} of energy from the exposure to light. This is typical for the real case of taking a picture in the daytime with a real camera.

It is important to note that this is a *tiny* amount of energy. A square meter is a lot of area, and the image on the detector in our example is very tiny in comparison—only 7.3×10^{-7} m^2. And so a total of only 7.0×10^{-5} J of light strikes the image area during the exposure, in our example. This is roughly the amount of energy required to lift a one-ounce first-class-mail envelope to a height of one quarter of a millimeter. In Volume 3 of *The Physics and Art of Photography* we will consider how it is that such a tiny amount of energy can still produce a measurable effect on a light detector.

2.2.10 Summary of steps

So now let us retrace these steps, from the Sun to the camera, and list them in order.

1. The Sun emits light with a power, $P = 3.85 \times 10^{26}$ J uniformly in all directions.
2. This light spreads out over a huge area by the time it reaches Earth, and so the Earth, and the subject, is illuminated by light with an intensity of $B = 1362$ W m^{-2}.
3. Our subject, inclined at an angle of 37° to the incoming sunlight, receives an illuminance of $I = 1115$ W m^{-2}.
4. Our subject, only 2.5 cm in diameter, intercepts a power of $P = 0.5475$ W.
5. Our subject reflects only 18%, or $P = 0.098\,55$ W of the total that it intercepts.
6. The subject reflects the 0.098 55 W in all forward directions as a Lambertian reflector. By the time this light reaches the camera, directly in front of the subject and 3.5 m away, it has an intensity of $B = 2.56 \times 10^{-3}$ W m^{-2}.
7. The particular lens on our camera intercepts some of this light, and a power, $P = 1.168 \times 10^{-6}$ W enters the camera.
8. This light is focused by the lens as an image of a particular size (and thus area) on the detector, giving the image an illuminance on the detector of 1.600 W m^{-2}.
9. Multiplying the illuminance by the exposure time (set by the camera shutter) of 1/60 s, gives the exposure of 96.0 J m^{-2}

2.3 The Jones–Condit equation

In appendix A, I go through this process more formally, and derive an equation that relates the illuminance, I_D, of the image on the detector inside the camera, directly to the illuminance, I_{SUB}, of the light shining on the subject. The result is as follows:

$$\boxed{I_{SUB} = \frac{4f^2}{R} I_D} \tag{2.52}$$

where f is the focal ratio of the lens and R is the reflectivity of the subject (%-reflectivity/100). For our example, the illuminance on the subject was 1115 W m^{-2}, the illuminance on the detector was 1.600 W m^{-2}, the reflectivity was $R = 0.18$, and the focal ratio of the lens was $f = 5.6$. Given the reflectivity and focal ratio in our example, equation (2.52) says that the illuminance of the image on the subject should be 697 times larger than the illuminance on the detector; $1.600 \text{ W m}^{-2} \times 697 = 1115 \text{ W m}^{-2}$, in agreement with our step-by-step calculations.

Notice that, when all is said and done, the illuminance of the image on the detector (and thus the exposure time needed to achieve the correct density) *does not depend at all on the size of the subject or on its distance.* We did use those values in our step-by-step calculations, but if we instead carry them all the way through the algebra (as I do in appendix A), we find that both of those quantities cancel out in the end. Instead, for a subject reflecting light as a Lambertian reflector, the illuminance on the detector depends on only three things:

1. The illuminance on the subject.
2. The reflectivity of the subject.
3. The focal ratio of the camera lens.

For equation (2.52), I *have* made the additional assumption that the camera is far from the subject, compared to the focal length of the lens. This is usually true, but it is decidedly *un*true for close-up macro photography. For that case the distance to the subject *does* affect the image illuminance, sometimes greatly so. We consider this complication more fully in chapter 3, section 3.2.1. I have also ignored vignetting and the reduction of the transmitted light through the lens by reflection off the lens surfaces and absorption by the glass. The Jones–Condit equation (Ware 2016, p 305) takes all of these factors into account, but equation (2.52) is a good approximation.

I have here laid out the foundations for the relations between light and exposure in terms of physical quantities such as illuminance and power. But photographers take a very different approach to these same goals, as we shall see in what follows. From a physical standpoint, the photographer's approach is roundabout and obscures somewhat the physical principals involved. But from a practical standpoint of taking pictures with real cameras in the real world, it makes a lot of sense.

Reference

Ware M 2016 *Cyanomicon II–History, Science and Art of Cyanotype: Photographic Printing in Prussian Blue* http://www.mikeware.co.uk/mikeware/downloads.html

Chapter 3

Shutter speed and aperture

In chapter 2, we saw that exposure = illuminance × time. Here we describe how the photographer makes two different adjustments *to the camera itself* in order to alter exposure. One setting—*shutter speed*—allows the photographer to adjust the time over which the image interacts with the detector. The other setting is *aperture*, and this allows the photographer to adjust the illuminance *given the amount of light arriving at the camera from the subject*. We leave the other factors that affect exposure, and ultimately the density of the image, to chapter 4.

There is an important feature about the way photographers have chosen to make adjustments that affect the density of the image: adjustments are made in *steps*, and the difference between steps is a *factor of two* in exposure. In this light, we consider first shutter speed and then aperture.

There is another important basic distinction between the photographer's approach to setting exposure on the one hand and the perspective of the physicist on the other. Photographers measure and set exposure in *relative* terms, without the use of physical units such as W m^{-2}. The idea of the exposure step is key. If all factors are counted in steps that affect exposure by a factor of two, then all we need is some agreed-upon starting point for each of these factors. After that, it is just counting the number of steps. We lay out this procedure in detail in chapter 4.

3.1 Power and shutter speed

Since exposure is the product of illuminance and time, then one of the ways to change the exposure is to allow the image to interact with the detector for a shorter or longer amount of time. In most cameras this is controlled by a device called a shutter. It is usually a precise timer that controls a mechanical cover that opens, exposing the detector to light, and then closes, covering it up. In some cases (called a *leaf shutter*) this mechanism is located between the elements of the lens itself, while in others (known as a *focal plane shutter*) it covers the entire detector and is located directly in front it.

doi:10.1088/2053-2571/aae504ch3

Since the most commonly-used exposure times are fractions of a second, they are usually labeled with their reciprocals. And so 1/60 s is labeled '60,' 1/30 s is labeled '30,' etc. Furthermore, photographers generally refer to this not as 'exposure time,' which is literally what it is, but rather as *shutter speed*. This makes sense if one thinks about it; '60' means the shutter is 'moving faster' than does '30.' If the exposure time is greater than 1 s, then 's' is added to clarify: '4s' means 4 s. And an unusually-long exposure time measured in many seconds, minutes or even hours is known as a *time exposure*. Because of the use of reciprocals instead of the actual exposure times, a larger shutter speed number results in *less* exposure.

These are the shutter speed settings available on most cameras: 1, 2, 4, 8, 15, 30, 60, 125, 250, 500, 1000. In the order I have written them each step would give one half the exposure of the previous step. In addition, many modern cameras with electronically controlled shutters have settings of 2s, 4s, 8s, 15s and 30s, and additional fast speeds of 2000, and perhaps 4000 and 8000. Some also have 'half step' settings, in between the standard values. Furthermore, shutters may have settings of 'B' or 'T' that are not controlled by a timer, and so allow the photographer to take time exposures of whatever length desired. See figure 3.1 for an example of an old mechanical shutter mechanism.

Notice that the standard list of shutter speed numbers does not *exactly* represent steps of factors of two. And so, for example, $60 \times 2 = 120$, not 125. The standard numbers are tweaked slightly, for convenience. And so we start with 1 s and end with 1/1000s, rather than 1/1024s. In practice, the difference is so slight as to be unimportant.

Figure 3.1. An old mechanical leaf shutter for a large-format camera. It uses quickly-moving leaves of thin metal to briefly uncover the lens opening. The lens elements have been removed for this picture, and I am holding the shutter leaves open halfway. The timing mechanism—a complex arrangement of gears, levers and springs—is inside the outer housing. Note the shutter speed settings, listed in reciprocal seconds, on the right-hand side of the outer ring.

3.2 Aperture and focal ratio

To control the power of the light entering the camera, we use an *aperture stop*, a variable-sized hole of some sort. It is sometimes also called an *iris*, after the part of the human eye that accomplishes the same task. A camera iris is usually made of several overlapping thin metal leaves, and it is almost always located inside the camera lens, between the different lens elements. See figure 3.2 for an example. Our measure of aperture is the focal ratio, f, the ratio of the focal length of the lens to the diameter of the lens opening.

Equation (2.52) relates the illuminance of light on the subject to the illuminance of light on the detector from the image focused by the lens. We can solve this equation for the image illuminance as follows:

$$I_{\mathrm{D}} = \frac{R}{4f^2} I_{\mathrm{SUB}} \tag{3.1}$$

Equation (2.52) means that, for a given amount of light on the subject, I_{SUB}, and a given subject reflectivity, R, the illuminance of the image on the detector is proportional to $1/f^2$. And so, when pointing a camera at a particular subject, we can adjust the image illuminance simply by adjusting f. And since this illuminance scales as the *reciprocal* of the square of the focal ratio, it means that increasing the focal ratio, *decreases* the illuminance of the image.

As is the case for shutter speed, photographers choose focal ratio settings in *steps*, the difference of which results in a factor of two in exposure. And so, since the exposure is proportional to the illuminance, and the illuminance is proportional to $1/f^2$, then to increase the exposure by one step—a factor of two—we must *decrease* the focal ratio not by a factor of two, but rather by a factor of $\sqrt{2}$.

To adjust f according to the photographers' exposure steps, we need a list of settings that are the square roots of successive factors of two: $\sqrt{1} = 1$, $\sqrt{2} = 1.414$, $\sqrt{4} = 2$, $\sqrt{8} = 2.828$, $\sqrt{16} = 4$, etc. Here is a complete list that covers the possible f settings on nearly any camera lens one might come across. By tradition, the settings are rounded off to at most two significant digits:

1, 1.4, 2, 2.8, 4, 5.6, 8, 11, 16, 22, 32, 45, 64, 90

Figure 3.2. An iris aperture stop in the lens of a medium-format camera. The front lens elements have been removed to make the aperture blades more easily visible. The aperture is set, from left to right, at *f*/4.5, *f*/8 and *f*/32.

No real camera lens has all of these settings. For a given lens, the minimum value of f is determined by the glass of the lens itself, while the maximum value is set by the mechanical limitations of the iris inside the lens. Since the minimum f of the lens itself may not match up exactly to one of the standard exposure steps, the smallest f setting is often in between the standard steps. And so, for example, one can see that the f settings on the lens in figure 3.2 are 4.5, 5.6, 8, 11, 16, 22, 32.

I must point out a particular quirk of how photographers refer in print to focal ratio. To emphasize that it is a *ratio*, $f = 5.6$ (for example) is often referred to as '$f/$ 5.6.' To the mathematically-inclined this looks odd, as if one means 'f divided by 5.6,' which is of course simply 1, if $f = 5.6$. There is more quirky terminology: focal ratio is often referred to as 'f stop.' This arises from the earliest days of photography, before the advent of continuously-adjustable mechanical irises, when the focal ratio of a lens was controlled by inserting an 'aperture stop' into the lens—one of a set of metal slides with different size holes in them. As such, a photographer will often refer to increasing the focal ratio of a lens (decreasing the diameter of the aperture) as 'stopping it down.'

Since $f = F/D$, where F and D are, respectively, the focal length and diameter of the lens, a smaller f means a larger aperture diameter. And so the minimum f of a lens is when the aperture stop is 'wide open,' and not covering the lens at all. Since it is far more difficult to design and manufacture a lens of small f than large f, this *maximum aperture* (minimum f) is one of the key features of any lens. As such, lenses with a focal ratio less than 1.4 are very rare. For larger detector formats, such as the 4 × 5 inch format view cameras described in Volume 1 of *The Physics and Art of Photography*, everything is scaled up and even more expensive. And so the vast majority of lenses for such cameras have a minimum focal ratio that is greater than $f/4$.

We have appealed to equation (2.52) to see why it is $1/f^2$ that controls the illuminance of the image, but it easy to see why this is true. For a given energy flux (W m^{-2}) entering the lens, the power (in watts) that makes it through is proportional to the *area* of the lens opening (W $m^{-2} \times m^{-2}$ = W). But an aperture of twice the diameter allows *four* times as much light to enter the camera, since *the area of a circle is proportional to its diameter (or radius) squared*. And so the power of the light entering the camera must be proportional to D^2.

But illuminance of the image is not power, it is power per surface area. And so we must divide this by the surface area of the image. For a given subject a given distance away, the diameter of the image is approximately proportional to the focal length, F, of the lens. This fact is explained in detail in Volume 1 of *The Physics and Art of Photography*, but it is not difficult to see that the image is larger if it is farther from the lens. And since the *area* of the image scales with its diameter squared, the area of the image must be approximately proportional to F^2.

I say 'approximately' because the size of the image is directly related not to the focal length, but rather to the *image distance* from the lens. This is very nearly equal to F if the image is far away, but it is not true for close-up or macro photography. We consider that complication in more detail in section 3.2.1, but for now, let us make the assumption that the image area is proportional to F^2.

And so, the power (in watts) of the light from the subject entering the camera is proportional to D^2, while the area on the detector (in m^2) over which that light spreads out to form the image is F^2. Illuminance is power per area, in W m^{-2}, and so the illuminance, I_D, of the image on the detector is:

$$I_D \propto \frac{D^2}{F^2} \tag{3.2}$$

$$I_D \propto \left(\frac{D}{F}\right)^2 \tag{3.3}$$

$$I_D \propto \frac{1}{f^2} \tag{3.4}$$

where $\propto$ means 'is proportional to,' and we have used the fact that $f = F/D$. Thus, even without the more sophisticated analysis that results in equation (2.52), it is clear that the image illuminance must scale with the reciprocal of the square of the focal ratio.

Because of the way focal ratio is defined, a large focal ratio number results in *less* exposure. Recall that this is also true for the shutter speed numbers. And so the settings for both aperture and shutter speed are marked with numbers such that larger numbers result in less exposure.

3.2.1 The effect of focus on exposure

If an object is very far away from the lens, the image distance is roughly equal to the focal length of the lens. But when a camera is focused on a nearby object, the detector is at a distance from the lens *greater* than the focal length. This larger image distance means the same light is spread out over a greater area on the detector, and thus a *smaller* exposure results.

This means that exposure is really proportional, not to the inverse square of the ratio of the *focal length* to the lens diameter, but rather to the inverse square of the ratio of the *image distance* to the lens diameter. But the lens is marked with f numbers that correspond to $f = F/D$, not d_i/D, and so the photographer may have to apply an exposure correction when focusing on a very nearby object. This is especially relevant for macro photography, where the image distance may be as much as twice the focal length, leading to an effective focal ratio twice as great as what is marked on the lens. This results in only one fourth the exposure intended—a difference of two exposure steps—if the photographer (or the camera) does not apply an appropriate correction. See figure 3.3.

In this discussion of focal ratio and exposure, we have assumed that doubling the image distance doubles the diameter of the image on the detector (and so increases its area four times). This is true the vast majority of time, but it may not *always* be true. For what if the object is so distant that no lens can resolve it as anything but a perfect point? In that case the image on the detector would be only a point as well, no matter what the focal length of the lens.

Figure 3.3. A camera lens focused at infinity (left) is only half the distance from the film as one focused at 1:1 macro (right). The *f* settings on the lens correspond to infinity focus. For close-up work, the effective *f* is *larger* than this. For this example of 1:1 macro, a lens marked as *f*/5.6 would actually expose the detector at *f*/11 (two exposure steps less exposure).

The most obvious examples are the stars in the night sky; most appear as essentially perfect points in even the world's largest telescopes. There is the issue of diffraction; a perfect point-like object forms a diffraction pattern rather than a truly-perfect point-like image. But for most cameras this diffraction pattern is smaller than the smallest detail that can be resolved by the light detector, and so that is usually not a factor. Thus images of stars are *unresolved*; they appear as points no matter what focal length is used.

And so if one sharply focuses a camera on the stars in the night sky, a larger aperture lets in more light (and thus increases the exposure), but a longer focal length has essentially no effect on the size of the focused image of the stars. This means that *exposure for an unresolved image is simply proportional to the square of the diameter of the lens*, rather than the inverse of the square of the focal ratio.

When taking real pictures of the night sky it is usually more complicated than this, as the stars are often not all that is in the picture. Besides the unresolved, point-like stars, there may be resolved, diffuse elements in the photograph—objects in the foreground or perhaps the glow of the Milky Way. The exposure of the resolved elements are related to the focal ratio of the lens, while the exposure of the unresolved stars are related only to the lens diameter.

Chapter 4

Density and the elements of exposure

When all is said and done, there are four factors that determine how dark or light a particular part of the picture will be. The photographer's name for the degree of physical effect on the detector is *density*. And so for traditional negative film, for example, density is how dark the film gets when it is developed. For a digital detector, density is related to the amount of electric charge that builds up on a part of the detector due to its exposure to light.

The density that is produced is a result of a combination of the exposure from the light, and the properties of the detector itself. We have already considered exposure, and it has three elements, listed below. The fourth factor that determines density is a measurement of the *sensitivity* of the detector to exposure. These four factors are summarized below.

Shutter: This determines the length of time for which light enters the camera, and thus the total energy transferred. All else being equal, *a smaller shutter speed number (slower shutter speed) results in greater density*.

Aperture: This varies the power of the light entering the camera compared to how much it spreads out on its way to the detector, by adjusting the relative size of the lens opening compared to the focal length. All else being equal, *a smaller f/ number (larger aperture) results in greater density*.

Light: While the first two factors are settings on the camera itself, the third factor that determines exposure is, of course, the intensity of the light arriving at the camera. *Metering* is the process of measuring the amount of light. All else being equal, *more light results in greater density*.

ISO: ISO is the official measure of the *sensitivity* to light of the light detector (whether it is film or digital). We will discuss the physics of this in detail in Volume 3 of *The Physics and Art of Photography*. The same exposure on a more-sensitive detector produces a greater density, and thus a brighter picture. All else being equal, *a larger ISO number means a more-sensitive detector, and so results in greater density*. Typical ISO numbers are 50, 100,

doi:10.1088/2053-2571/aae504ch4

200, 400, 800, 1600. Each represents a doubling of the previous, and thus a doubling of the detector sensitivity. With twice the ISO, half the exposure is needed to produce the same density.

As a warning, we will consider factors that complicate this a bit. For example, the specific intensity of light entering the camera is different in different directions, obviously, or else our picture would be one uniform shade of gray. That is why some parts of our picture are dark and some are light. So what do we do if we are faced with a choice of either making some parts of the picture too dark or some parts too light? What if there is no single exposure that gets both the dark parts and the bright parts right at the same time? Life is complicated, and so is photography. We will have more to say about this important issue, but let us put it aside for now.

Another complication is color. Although aperture and shutter speed for the most part have the same effect on exposure whatever the color of light, the same may not be true for ISO. And so this fact, that light is comprised of a range of wavelengths, will have significant consequences for exposure and for metering. We will take up this issue in Part II of this volume.

4.1 Reciprocity and exposure

Exposure is the total energy/area that strikes the detector. Since the illuminance of the light falling on the light detector is energy/time/area, it is easy to see that exposure is given by:

$$\text{Exposure} = \text{Illuminance} \times \text{Time} \tag{4.1}$$

Notice that it is neither illuminance nor time alone that matters, but rather the product of the two. If, for example, one is quadrupled while the other is reduced by a factor of four, the same exposure results: $4 \times \frac{1}{4} = 1$.

The density that results from an exposure depends on the detector ISO, clearly. But there is another important question. Does it depend *only* on the exposure—the product of illuminance and time? Or does density depend on *both* illuminance and time separately, in some more complicated way?

The short answer is that most of the time, density is a function of exposure alone, and not illuminance or time specifically. When this is true, it is called *the law of reciprocity*, and it means that any two exposures that are the same, but have different combinations of illuminance and time, will produce the same density. And so one can reduce the illuminance by a factor of 5, while increasing the time by a factor of 5, and the detector will produce the same density.

There are occasions, however, when the law of reciprocity does not hold, and this is called *reciprocity failure*. The most common case is that some detectors have, in effect, a lower ISO value at very low levels of illuminance. We will revisit reciprocity failure specifically in Volume 3 of *The Physics and Art of Photography*, but for what follows I will assume that the law of reciprocity holds.

4.2 Camera settings

We have seen that exposure is the product of illuminance and time. Exposure is the cause, but density—what actually happens to the detector—is the effect. It is the detector sensitivity, ISO, that connects density to exposure, and so we can say the following:

$$\text{ISO} + \text{Exposure} \Longrightarrow \text{Density}$$

Taken together, these facts mean that density is determined by a combination of *three* things: detector ISO, exposure time, and illuminance of the light falling on the detector. The catch is that one of these three—illuminance of the image focused on the detector—is usually not directly measured or controlled by the photographer. Instead, image-plane illuminance is determined by a combination of aperture and light. And so the photographer has the following *four* factors: ISO, exposure time (shutter speed), aperture (focal ratio), and intensity of light arriving at the camera:

$$\begin{aligned} \text{Light} + \text{Aperture} &\Longrightarrow \text{Illuminance} \\ \text{Shutter} + \text{Illuminance} &\Longrightarrow \text{Exposure} \\ \text{ISO} + \text{Exposure} &\Longrightarrow \text{Density} \\ &\text{and so:} \\ \text{Light} + \text{Aperture} + \text{Shutter} + \text{ISO} &\Longrightarrow \text{Density} \end{aligned}$$

Aperture, shutter and ISO are settings on the camera. These three factors, combined with the light, determine the density. All three of our camera settings use steps that, taken individually, double or halve the density from one step to the next. Table 4.1 shows typical settings, each representing one step in density, for

Table 4.1. Possible camera settings to affect the density of the picture, given the light reaching the camera. Greater density is toward the top of the table. Each number in a given column represents a doubling of the physical exposure as compared to the number below it, or a halving of exposure as compare to the number above it. But we call such a doubling (or halving) a single *step*.

f/number	Shutter speed	ISO
1.4	1	6400
2	2	3200
2.8	4	1600
4	8	800
5.6	15	400
8	30	200
11	60	100
16	125	50
22	250	25
32	500	12
45	1000	6
64	2000	
90	4000	

shutter, aperture and ISO. Taking each of the three settings individually, the top of the table represents increased density while the bottom of the table represents lower densities. Most cameras have only a subset of these possible settings, and it is possible to extend the table beyond what I have reproduced here.

ISO is a special case. Traditionally, photochemical detectors were used in cameras (ordinary film is an example), and a given piece of film has a particular ISO. Thus, once the camera is loaded with film, the ISO is already chosen. Most cameras have, on the other hand, many possible settings for aperture and shutter. Digital cameras, however, are different. It is possible to internally change the sensitivity (and thus the ISO) of the light detector in a digital camera. And so ISO can be 'set' from picture to picture just like aperture and shutter.

Below I list three examples, each of which, taken alone, *increases* the density by one step.

1. Change the shutter speed from 250 to 125.
2. *or* change the aperture from *f*/5.6 to *f*/4.
3. *or* change the ISO from 200 to 400.

What if I instead did all three? What if I changed the shutter speed from 250 to 125 *and* changed the aperture from *f*/5.6 to *f*/4 *and* changed the ISO from 200 to 400? Well, since each individually would increase the density by one step, doing all three increases the density by three steps. The same logic applies if decreasing density. A decrease in any of the three decreases the density. If some increase and some decrease, well then one must both add and subtract to determine the overall change. Armed with this knowledge, you should now be able to explain why this next example results in an *increase* in density of 3 steps:

1. Change the shutter speed from 250 to 15.
2. *and* change the aperture from *f*/5.6 to *f*/16.
3. *and* change the ISO from 200 to 800.

If you are having trouble with this exercise, think of it this way. The change in shutter speed from 250 to 15 moves *up* the table by 4 steps. And so let us record that as +4. The change in aperture from *f*/5.6 to *f*/16 moves *down* the table by 3 steps; so let us record that as −3. And finally, the change in ISO from 200 to 800 moves up the table by 2 steps, or +2. To find the overall result, simply add the numbers: $+4 - 3 + 2 = +3$, indicating an increase in density of 3 steps.

4.3 Choosing between equivalent settings

It should be evident that *there are many different camera settings that achieve exactly the same density* for a given amount of light entering the camera. For example, one could …

1. Increase *f* by 3 steps, while decreasing the shutter number by 3 steps.
2. Decrease the ISO number by 2 steps, while decreasing the shutter number by 2 steps.

Figure 4.1. The left-hand image was taken at *f*/1.4 with a shutter speed of 1/2000 s, while the right-hand image was taken at *f*/16 with a shutter speed of 1/15 s. Both combinations produce the same exposure, but the smaller focal ratio results in a greatly-reduced depth of focus.

3. Increase the ISO number by 3 steps, while increasing the shutter number by 2 steps *and* increasing *f* by 1 step.

…and the exposure would remain exactly the same. So why choose one combination over another? The reason is, of course, that density is not the only issue. Each of these three settings has other consequences for the picture.

4.3.1 Aperture and depth of focus

Apart from exposure, focal ratio alters the *depth of focus* of the picture—the range of subject distances that are in acceptable focus. A smaller focal ratio leads to a more narrow range of distances that are in good focus, while a large focal ratio makes both nearby and distant objects in focus at once. And so for example, one might want to use a combination of shutter speed and ISO that achieves the desired exposure, while using a large aperture (small *f*)—because this results in a shallow depth of focus, and so isolates the subject from the background. Or perhaps one wants the opposite—to use a small aperture to get both the foreground and background in focus simultaneously. One can then pick, for the amount of light present, a combination of ISO and shutter speed that gives the proper exposure for the aperture desired. See figure 4.1 for an example.

4.3.2 Shutter speed and motion blur

It should be unsurprising that shutter speed is important in relation to time and motion. A fast shutter speed (large shutter speed number) freezes motion, both for the subject and the camera. And thus there is the rule of thumb that when holding a camera by hand, a shutter speed number lower than 30 will likely produce a blurred picture, as it will be too difficult to hold the camera steady enough to insure it does not move appreciably during the picture[1].

As is the case for selective focus, careful control of *motion blur* can be an essential part of a composition. Take two pictures of a waterfall, one at a shutter speed of

[1] This rule of thumb assumes that a lens with a 'normal' angle of view is being used. There are modified versions of this rule for wide-angle or telephoto lenses.

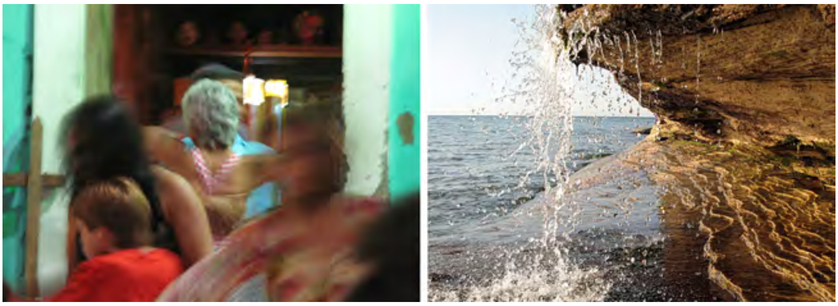

Figure 4.2. Left: The camera was panned with one of the dancers, and the shutter triggered when she was relatively motionless. A moderately slow (1/8 s) shutter speed was used to produce motion blur in the other dancers. **Right:** a very fast shutter speed (1/2000 s) was used to freeze the motion of the falling water, so it would look like a crystal sculpture.

1/4000s and the other with a shutter speed of 1/2 s (using a tripod). The former will freeze the motion of every drop of water, and the waterfall will look like a crystal sculpture. The latter will make the water look like a dreamy cloud flowing over the sharply-defined rocks. Which do you want?

Typical examples of using shutter speed as part of the composition can be seen in figure 4.2. For the photograph on the left I used a shutter speed fast enough that I could successfully pan the hand-held camera with one of the dancers, triggering the shutter at a moment when she was relatively motionless. But the shutter speed was slow enough to produce motion blur in the rapidly moving dancers surrounding her. This had the effect of isolating the subject, while reducing the other dancers to abstract swirling patterns that emphasize the motion of the dance.

For the photograph on the right, I did the opposite. I used a very fast shutter speed (1/2000 s) to freeze the motion of the falling water, so the waterfall would look like a crystal sculpture.

For a much better example, consider Carlo Bevilacqua's intriguing use of motion blur in his photograph *Catari* (see Cooke and Kinneberg 2014, p 59). A large depth of focus (from a small aperture) leaves background detail distinct so as to ground the figures in a real place. The animated face of the lead figure is seemingly frozen in time, while her clothes are blurred by motion due to the slow shutter speed. The girl's skirt seems to have moved mostly in the vertical direction, making its outline indistinct, but leaving intact its vertical stripes.

Successful use of motion blur often involves an element of luck, combined with the experience to know how to make the most of the situation. In a real scene some parts usually move faster than others, while other parts don't move at all. A slow shutter speed will cause the faster-moving parts of the scene to blur the most. Panning the camera with the moving object has the opposite effect; the stationary parts of the scene blur the most. Finally, the blurring is always in the direction of the *relative* motion between the subject and camera. In a real situation one must often react quickly, and so it is crucial to think ahead about what *might* happen, and to have already chosen a shutter speed that is appropriate for the desired effect.

Figure 4.3. From the series, *Motion in Scotland*, John Beaver 2017. I panned the camera from a moving bus, using a slow shutter speed.

The photographs in figure 4.3 are a complex example of the use of motion blur. I took the photographs out the window of a quickly-moving bus, while panning the camera with particular features as we passed them. The effect of the panning depends strongly on distance, and I now had two motions—the motion of the bus and the panning of the camera. The complexities of which features were sharp, which were blurred, and in what direction, often defies after-the-fact explanation. And so luck was a big part of the result. As such, I took over 200 pictures, but exhibited only 15 of them.

4.3.3 ISO and noise

There are at times competing reasons for wanting to set aperture and shutter speed. For example, suppose one wants to freeze motion, but also show a large depth of focus. Both choices will result in a *decrease* in exposure. And so it is nice that there is the third factor of ISO. If increasing both the shutter speed number *and* the *f* to the desired values gives too little exposure for the amount of light, one may be able to simply increase the ISO number to compensate.

But this too, has its consequences. We will take up this issue more fully in Volume 3 of *The Physics and Art of Photography*, but I point out here that an increase in ISO inevitably results in a loss of detail in the picture. This can show up as 'graininess' and a lessening of the range of subtle values and colors that can be portrayed. See figure 4.4 for an example.

Figure 4.4. **Left:** ISO 200, *f*/5.6, 1/15 s. **Right:** ISO 3200, *f*/5.6, 1/250 s. Both images have the same density, but the right-hand version, taken at higher ISO and correspondingly lesser exposure, has more noise. This is apparent in the loss of fine detail and the grainy nature of the blacks in the close-up view. But the loss of richness in the colors is apparent even in the unmagnified view.

4.3.4 Changing the light

Finally, it may be possible to change the light itself. In a studio, one can shine more light on the subject with flashes or flood lights. In the field, one can use a portable flash or a reflector to add some extra light on the subject. To reduce the light, one can use a neutral-density filter.

4.3.5 Navigating the trade-offs

I here summarize the consequences, apart from exposure, of a particular choice of aperture, shutter speed and ISO:

1. A smaller aperture (larger *f*) results in a greater depth of focus. This can be either a good thing (if one wants the viewer to compare nearby objects to far-away objects) or a bad thing (if one wants the in-focus subject to be isolated by selective focus from its surroundings).
2. A faster shutter speed (larger shutter speed number) results in a greater freezing of motion of the subject and camera, and less motion blur.
3. A more light-sensitive detector (larger ISO number) results in a picture that is more grainy and less subtle in its range of values and hues.

If we add to this the constraints of getting the proper exposure, we also have the following:

1. Shallow depth of focus, all else being equal, means one will need a *faster* shutter speed (because *f* is smaller), and thus a greater freezing of motion. On

the other hand, if one wants a large depth of focus (large *f*), the consequence is a slower shutter speed and thus more motion blur.

2. Greater motion blur (slower shutter speed) means that, all else being equal, a larger *f* is required, and thus a more shallow depth of focus. If one wants to freeze the motion (faster shutter speed), then the consequence, all else being equal, is a smaller *f*, and thus a more shallow depth of focus.
3. If we do not want motion blur from the vibrations of a hand-held camera, then we must use a tripod for shutter numbers smaller than 30 or so.

And so there are trade-offs, imposed by the constraints of exposure, on our compositional choices regarding motion blur on the one hand and depth of focus on the other. What if one wanted to freeze motion, with a large depth of focus, in dim light, and with relatively low amounts of noise and graininess? It should be clear that all of these desires are in conflict with each other. Sometimes we can't have it all; we have to make tough choices.

4.4 Exposure value (EV)

For traditional film cameras, the ISO is a pretty-much fixed property of the film loaded into the camera. Particularly when using roll film, once the camera is loaded, one is stuck with whatever ISO of film is in the camera. It is possible to alter the ISO in the process of chemical development of the film, but still one must make the choice of ISO for the entire roll, not for each individual picture.

And so, before the advent of digital photography, exposure was changed from picture to picture primarily through shutter speed and aperture alone. For this reason, it has long been convenient to talk of an *exposure value, or EV* that represents the effect of only aperture and shutter speed on exposure. *A particular EV value represents all of the combinations of aperture and shutter speed that give the same exposure*.

Like aperture, shutter speed and ISO, EV numbers represent steps that correspond to a doubling of exposure. And as is the case for shutter and aperture numbers, a *smaller* EV number represents *greater* exposure. And so EV = +1 represents, all else being equal, one step higher (greater exposure) than EV = +2. I say '+2' because EV numbers can be negative. And so EV = −3 represents an increase of 5 exposure steps over EV = +2 (because −3 is 5 less than +2). And yes, there is EV = 0; it represents an increase of exposure of one step over EV = +1 and a decrease in exposure of one step from EV = −1.

An EV of 0 is defined to be any combination of aperture and shutter speed equivalent to a 1 s exposure at *f*/1.0. Very, very few lenses have an *f*/1.0 setting, but it represents one aperture number step smaller than *f*/1.4 (which is not so uncommon). From this definition and the fact that EV numbers represent our normal exposure steps, we can make a table of EV values for some combinations of aperture and shutter speed. See table 4.2.

We can also look at this a different way, as seen in table 4.3. Here we represent different combinations of aperture and shutter speed, *all of which represent the same EV* (+15 in this example). Thus, for a given amount of light, these all represent the

Table 4.2. EV numbers for some different combinations of aperture and shutter speed. For a particular ISO, these EV values represent correct exposures for different lighting conditions. The fourth column gives lighting conditions for EV_{100} (scene descriptions adapted from en.wikipedia.org/wiki/Exposure_value 2018, table 4.1 and references cited therein).

EV	f/	Shutter speed	Scene for ISO 100
−1	1.4	4 s	
0	1.0	1	
+1	1.4	1	
+2	1.4	2	Distant view of lighted buildings at night
+3	1.4	4	Floodlit building at night
+4	1.4	8	Floodlit building at night
+5	1.4	15	Home interiors
+6	1.4	30	Home interiors
+7	2	30	Home interiors
+8	2.8	30	Office or work area
+9	4	30	Night sports
+10	5.6	30	Skyline just after sunset (clear)
+11	8	30	Skyline just after sunset (clear)
+12	11	30	Heavy overcast or open shade on sunny day
+13	16	30	Cloudy bright (no shadows)
+14	16	60	Hazy sun (soft shadows)
+15	16	125	Sunny bright (distinct shadows)
+16	16	250	Distinct shadows, sand or snow

Table 4.3. Shutter speed and aperture combinations for EV = +15.

EV	f/	Shutter speed
+15	2	8000
+15	2.8	4000
+15	4	2000
+15	5.6	1000
+15	8	500
+15	11	250
+15	16	125
+15	22	60
+15	32	30
+15	45	15
+15	64	8
+15	90	4

same exposure. Notice that if we read down the table, aperture numbers get bigger while shutter speed numbers get smaller. This leads to the *one up, the other down rule*: increase the aperture number by a set number of steps while decreasing the shutter speed number by the same number of steps, and the exposure remains unchanged. Similarly, increase the shutter speed number by a set number of steps while decreasing *f* by the same number of steps, and the exposure is unaffected.

When EV is combined with a particular ISO, it represents all of the camera settings that determine a particular density, *given the amount of light*. Thus, for a particular ISO setting, EV values can be used to represent the proper exposure for different lighting conditions. This is often given in a chart of EV_{100}, which represents EV settings combined with ISO set to 100. The fourth column in table 4.2 gives lighting conditions that correspond to EV_{100} exposures.

Looking at table 4.2, notice that entries at the top of the table (smaller EV values) represent camera settings that would result in an *increase* of exposure over choices from lower in the table (larger EV numbers). But the opposite is true for the lighting conditions listed in the fourth column. A moment's thought should render this unsurprising; a darker scene requires a camera setting that corresponds to a greater exposure, compared to the camera setting for a brighter scene.

The particular example I have chosen for table 4.3, $EV_{100} = +15$, is an important one; it represents the normal exposure for taking pictures outdoors on a clear, bright sunny day, assuming a 'normal' landscape. It is often referred to as *sunny bright*, and one can see in the fourth column of table 4.2 names for related outdoor scenes that require either more or less exposure.

Notice that one of the combinations in table 4.3 is 16 for the aperture and 125 for the shutter. And so in this case the shutter number (125) and the ISO (100) are almost the same. If we made instead a similar EV chart for ISO 400 (2 steps *greater* density for the same exposure), all of the shutter speed numbers would have to be adjusted to 2 steps *lesser* exposure (larger shutter numbers) for any given *f*, in order to give the same density, and thus still be appropriate for the same lighting condition. And thus for that EV_{400} chart the same density would result at EV = +17 instead of EV = +15. This would correspond to 16 for the aperture and 500 for the shutter, and we would again find that our ISO number (400) is about the same as our shutter speed number (500) for an aperture choice of *f*/16.

This leads to what is known as the *sunny 16 rule. When taking pictures outdoors on a bright sunny day, set the aperture for f/16 and then set the shutter number as close as possible to the ISO of the film*, and the exposure will be approximately correct. To use an aperture other than *f*/16, simply use the 'one up, the other down' rule to pick another combination of aperture and shutter speed that has the same EV.

References

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

2018 en.wikipedia.org/wiki/Exposure_value. URL https://en.wikipedia.org/wiki/Exposure_value (Collected September 20, 2018).

IOP Concise Physics

The Physics and Art of Photography, Volume 2

Energy and color

John Beaver

Chapter 5

Metering

Metering is the process of measuring how much light from the subject will arrive at the camera, in order to correctly control the exposure in order to achieve the desired density. To meter properly we must make the distinction between forms in the picture that are emitting their own light (the sky, for example) and forms that we see because of light reflected from some other source. There is also another special case, and that involves the use of flash. Furthermore, some objects reflect more light than others. These considerations lead to different approaches to metering, and we will consider each in turn:

1. **Ambient-light metering**—The scene is lit by ordinary light sources that deliver luminous energy gradually over time. And thus the *intensity*—proportional to the luminous energy *per second*, or power—is measured.
2. **Flash metering**—The scene is lit by a *flash*, which delivers all of its luminous energy at once in a short burst. And thus the *total luminous energy*, not the power, is measured.
3. **Incident-light metering**—The light *falling on the subject* is measured.
4. **Reflected-light metering**—The light *arriving at the camera* is measured.

And so a light meter may measure either ambient light or flash. And whether it does one or the other (or both), it may be either a reflected-light or incident-light meter.

In all of these cases the starting point is to render on the picture dark forms dark and bright forms bright. That may not be, in every case, the final goal; one may want to make a picture that is very different from how the forms in the real world appeared to the human eye. But even then, the starting point is usually to determine what would be the proper exposure to render the forms in the picture as they appear in real life. One may then adjust the exposure accordingly, to make either an *overexposure* (to brighten the picture) or an *underexposure* (to darken the picture).

In what follows we first consider meters that measure ambient light, leaving the special case of flash photography for section 5.4. Furthermore, many cameras

doi:10.1088/2053-2571/aae504ch5

use a built-in meter to set the exposure automatically, according to some pre-arranged scheme; we will explore these automatic systems further in section 5.2.2. We first consider meters and cameras that are used in a manual mode.

Photographic light meters do not record the intensity of light in physical units, such as W m^{-2}. Instead they display, for the amount of light measured, the required combination of camera settings in order to achieve a ‘normal’ range of densities on the detector. Most often, ISO is set first, and then different combinations of aperture and shutter (or EV) can be selected that are appropriate for the measured amount of light.

5.1 Direct-read versus null meters

Whether the light meter is a separate hand-held device or is built in to a camera, there are two basic approaches to how the meter communicates with the photographer. In a *direct-read meter*, there is some kind of indicator that directly shows the proper exposure for the measured light. In older meters, this was usually some kind of moving needle. A more modern direct-read meter typically uses a digital display to show the properly-exposed combination of shutter and aperture numbers, usually with some method for displaying alternative combinations with the same EV or with a different ISO. As an alternative, some direct-read meters indicate only the EV number (for a particular ISO setting) for proper exposure, leaving it up to the photographer to pick a combination of aperture and shutter with that EV. A direct-read meter built in to a camera may point to shutter numbers for a particular aperture, or to a focal ratio for a particular shutter speed, depending on the design of the camera.

A *null meter*, on the other hand, does not directly communicate the meter reading. Instead one adjusts aperture, shutter and ISO (or EV and ISO), and the meter tells the photographer whether that combination would be under- or over-exposed. Typically, there is some system of arrows or other marks to either side of a ‘perfect exposure’ marking. And so one adjusts the exposure by playing the higher-lower game.

Null meters are very popular for manual exposure modes on in-camera meters, because they are easy to understand. But they have the disadvantage that the photographer only knows the camera setting once the meter is ‘nulled’ to a perfect exposure (by adjusting the shutter speed or aperture or both). With a direct-read meter, on the other hand, it is possible for the photographer to point the meter at different parts of the scene and quickly determine a range of readings.

5.2 Reflected-light metering

Reflected light metering occurs *at the camera*. It measures the amount of light from the subject that actually makes it to the camera. As such, reflected-light meters are often located inside the camera itself. If a reflected-light meter measures the amount of light that actually makes it through the camera lens, it is called a through-the-lens, or *TTL* meter.

Reflected-light metering, especially of the TTL variety, is very convenient since it measures just what the camera sees (figure 5.1). And so it can easily be used to guarantee that the detector will be exposed by an amount that puts the recorded light and dark values in the middle range of what the detector is able to record. On most cameras this can be done automatically, thus guaranteeing that every exposure is right at the optimum level for the particular light detector.

There is, however, a serious disadvantage to reflected-light metering; *the light meter does not know what the picture is supposed to look like*. Maybe there is a lot of light entering the camera because one is taking a picture of a snowy field on a sunny day. In that case, the picture is supposed to be pretty much all white—but the light meter doesn't know this. And so, left to its own devices, it will decrease the exposure to make that bright, snowy field look a mid-tone gray instead. Similarly, if one takes a picture of a bunch of dark rocks, that are *supposed* to appear dark on the picture, the light meter in the camera will increase the exposure so as to make the black rocks appear a mid-tone gray.

The point is, things in the world appear dark or light partly due to the light that is falling on them, but also due to how they reflect light. The black rock appears dark to us because it reflects only a small percentage of the light that falls on it, while the

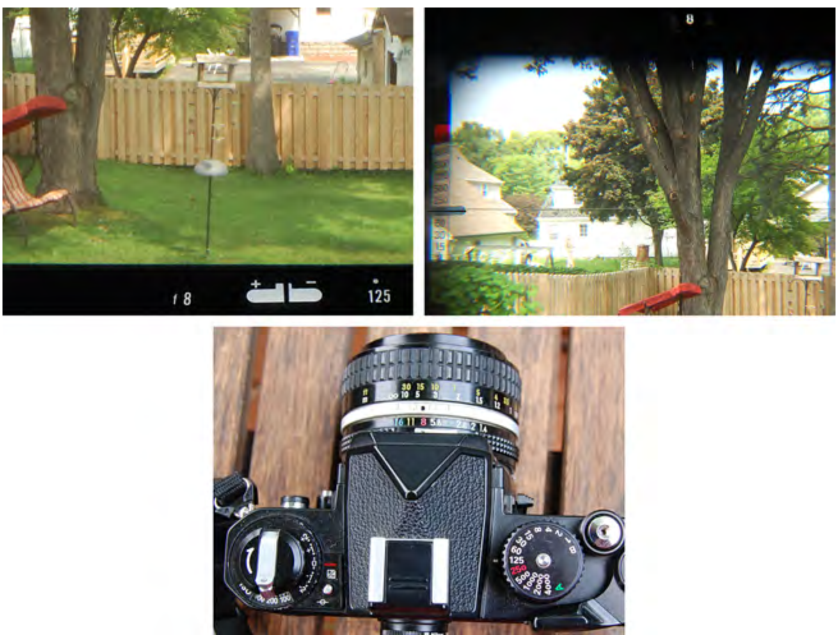

Figure 5.1. TTL metering systems in two vintage 35 mm SLR cameras. At top left is the view through a Nikon F2, which uses a null meter. The centered needle indicates correct exposure, in this case set at aperture *f*/8 and shutter 1/125 s. At top right is the view through a Nikon FM3a, which uses a direct-read meter. The blue indicator shows the shutter is set for 1/125 s, while the aperture of *f*/8 is shown by the number at the top of the frame. A black needle points to the shutter speed needed for correct exposure. One then matches the blue indicator and the black needle (or leaves them mismatched if under- or overexposure is desired). The bottom image shows a Nikon FM3a set for *f*/8, 1/125 s at ISO 200.

snow appears bright white because it reflects a very high percentage. Shine a lot of light on an object of low reflectivity, and then shine only a little light on something of high reflectivity; both situations may result in exactly the same amount of light arriving at the camera, and the meter cannot tell the difference between the two. And so the meter must make some assumptions.

By convention, the photographic metering system is designed with a reflectivity of 18% as the standard[1]. And thus any object that reflects 18% of the light that falls on it will be exposed correctly when metered with a reflected-light meter. And so if one uses a reflected-light meter to measure the light reflected off, for example, white snow, which reflects close to 80% of the light falling on it, the meter will nonetheless *assume* the snow is reflecting only 18% of the light. And so it will suggest an exposure that would make the snow look grey, not bright white. An *increase* in exposure over the meter's suggestion would be required, in this case, to make the snow look bright white.

One solution is to read the meter only while it is pointed at something that is illuminated by the same light as the subject, but that reflects about 18%. Then one sets the exposure on the camera manually, based on that 18%-reflectivity meter reading. Many photographers carry around a *gray card*, a piece of cardboard coated with a dull surface that reflects 18%. They then put the gray card in the same position as the subject (or with the same light falling on it as the subject), read the meter while it is pointed at the gray card, and then set the camera exposure manually before taking the picture. It is even possible to buy a T-shirt that is 18% gray, so one can always have at least a 'gray shirt' handy.

A variation on this technique is to simply memorize what 18% gray looks like, and find something in the scene that looks like that, and has the same light falling on it, and so use that in place of a gray card. It is interesting that when we humans look at things in the world, we see them not in terms of the light they reflect—even though that is what literally enters our eyes. Instead we see things in the world according to our unconscious assumptions about the light that is falling upon them. We know, instinctively, that a white object in very dim light reflects very little light. And so, *in context*, we still see it as white. The camera, on the other hand, does not know this; and so we must tell it.

5.2.1 Spot, center-weighted and matrix metering

When one points a light meter at a subject, just what light does it record? Does it measure only the light in one precise direction? Or does it record an average of all of the light in the general direction it is pointed?

The answer is that it depends. First off, recall that if we are talking about a TTL meter built into a camera, then clearly it is only measuring the light coming through the lens. But even in this case, we still must ask the question whether it is measuring the brightness of only one small spot on the picture, or is it measuring all of the light

[1] This is approximately true, but it is more complicated than this when other factors, such as the various wavelength response of different detectors, are taken into account.

the picture will record. So let us make a distinction between the three most-common types of reflected-light meters.

1. **Spot**—The meter responds only to the light coming from a small region in a very precise direction.
2. **Center-weighted**—The meter responds to a large angle of view, but the light coming from a small region in the center counts for the majority of the meter's reading.
3. **Matrix**—The meter makes many small individual measurements of different small parts of the scene. It then combines these into one reading according to some program.

Many modern cameras, have a built-in meter that can be switched between spot, center-weighted and matrix modes.

A *spot meter* measures only a small spot on the picture, usually at the direct center of the field of view. Thus if one points the meter at someone wearing a white shirt and black pants, it will read a low value when pointed at the person's pants and a high value when pointed at their shirt. To use this kind of meter, one takes many readings on different parts of the image in order to determine the range of exposure values. Then the exposure decision is made based on how one wants the picture to come out. Needless to say, this is a slow and meticulous way to determine the exposure, but it is also the most versatile.

Many photographers carry a separate dedicated spot meter, such as the one shown in figure 5.2. It shows a magnified view, somewhat like a small telescope, and the meter needle (visible from the inside) responds only to the light in a small circle (5° across, for this particular meter) at the center of the field of view. The needle on the inside reads EV_{100} values, and the photographer reads off aperture, shutter speed and ISO by dialing in the meter reading to the rotary dial on the outside of the meter.

One of the most common types of built-in TTL light meters is the *center-weighted meter*. A small percentage of the meter reading corresponds to the entire image (in a TTL meter) or field of view (in a hand-held meter). Most of the reading, however comes from a usually-medium-sized circle in the center. For most center-weighted

Figure 5.2. A typical hand-held spot meter. This model shows a magnified view of the subject, like a small telescope. The meter responds only to the small 5° circle at the center of the field of view. The EV value is read from inside the meter, and the aperture and shutter speed can then be calculated by the set of wheels on the side of the meter.

meters, it is about 60%–70% for the center and 30%–40% for the rest. This works pretty well on its own for many common situations, but it still allows the photographer to move the camera around and point at different objects to crudely sample the possible range of meter readings.

The most sophisticated type of meter is the *matrix meter*. It records light readings at many small places scattered (usually in some kind of grid pattern) over the image. It then makes a decision for the best exposure based upon some kind of program. Cameras with matrix meters often have many different programs the photographer can choose from. Each makes different assumptions about just what is a 'typical' scene, and thus what the light and dark values may represent. The idea is that, once one chooses the appropriate program, the photographer can just trust the meter to do the right thing. This actually works pretty well much of the time, and it is usually the best type of metering for taking pictures in a fully-automatic mode.

Matrix metering gives the photographer far less control over individual pictures since, due to its complex program, it is often difficult to know just what the meter is doing and why. Spot metering gives by-far the most control, but it is also the most difficult and time-consuming to use, and it is pretty much hopeless when combined with automatic settings. In many situations, center-weighted metering provides a nice balance for the photographer between speed and control.

5.2.2 Manual, automatic, semi-automatic and program exposure modes

Many cameras have a prominent dial with, among other symbols, four enigmatic letters: M, A, S and P. They represent, respectively, *manual*, *aperture-preferred*, *shutter-preferred* and *program*.

In most cases, the ISO is set separately from the information provided by the meter, and we can think of that as the starting point for metering. It is the first thing one chooses, based on whether there is, in general, a lot of light (use a low ISO number) or very little light (use a high ISO number). Many digital point-and-shoot cameras, as well as most cell-phone cameras, do this automatically and may not even allow the photographer to adjust the ISO at all. We consider here cameras that at least allow for the possibility of manual metering.

The *manual exposure mode*, usually symbolized on a camera by 'M,' means just what it says. The photographer adjusts both aperture and shutter speed according to the reading of the meter. This means of course, that one does not have to do what the meter recommends. For example, if the meter says the camera should be set for a shutter speed of 60 and an aperture of *f*/5.6, one is perfectly free to use an aperture of *f*/4 instead.

Reflected sunlight flickering off a flowing stream is a good example of a situation that can be difficult to meter; in this case the light (and thus the meter reading) changes rapidly with time. If one were to use any form of automatic metering, the result would be essentially random. But with a manual exposure mode, one can watch the meter for a few seconds and get a feel for the range of values, and then set it accordingly. Or one could point the camera at something else entirely to read the meter, since the processes of reading the meter and setting the camera are separate from each other.

Aperture-preferred metering ('A') allows the photographer to set the aperture manually, and the meter then automatically adjusts the shutter speed for the chosen aperture, according to the meter reading. *Shutter-preferred* metering ('S') is also an automatic exposure mode, but it works in the opposite sense; the photographer adjusts the shutter speed and the camera adjusts the aperture according to the meter reading. These are called *semi-automatic metering* modes, and one or the other is more appropriate depending on whether the photographer wants more control over aperture or shutter speed. For example, if one is using a tripod to photograph a non-moving subject, then a particular choice of shutter speed is unimportant, and an aperture-preferred mode is more appropriate, to allow for control of depth of focus. If on the other hand, one wants a particular amount of motion-blur in a moving subject, then a shutter preferred mode is more useful.

For either semi-automatic mode, once one picks the aperture or shutter speed, the possible range of the other settings limits the EV values that can be photographed. For example, if a camera has shutter numbers of only 1, 2, 4, 8, 15, 30, 60, 125, 250, 500 and 1000, then once one picks the aperture, the aperture-preferred mode will only have a range of 11 EV values to choose from. And it may turn out, for example, that there is too much light for the chosen aperture even for the highest available shutter speed. Most cameras would give some kind of warning in that case. The moral is, one must still pay attention even when using an automatic mode.

A *program exposure mode* ('P') adjusts both aperture and shutter speed automatically according to a prearranged scheme (a program). Most modern cameras actually have many different program modes. The 'P' setting is usually reserved for the simplest and least-sophisticated program that makes the fewest assumptions about the scene. It is thus, of all the program modes, the easiest for the photographer to understand.

Understanding the camera's auto exposure modes is important, as many cameras still allow the photographer some choice even with fully-automatic metering. This is usually done in one (or both) of two ways, the most common of which is called *exposure compensation*—the exposure is altered by a selected number of steps above or below that which the meter indicates. Less common, but even more useful in my opinion, is some sort of easily-accessible button that serves as an *auto-exposure lock*. One presses and holds the lock button, and the camera remains at the same exposure setting even as the camera is repositioned and fired.

With a little practice, a combination of semi-automatic exposure, spot or center-weighted metering with a direct-read meter, and use of an auto-exposure lock button can give the photographer nearly all of the flexibility of a fully manual setting—but at much greater speed. Even so, manual metering is still necessary for conditions of randomly-changing lighting conditions.

Finally, I point out that in certain cases manual meeting is, surprisingly, the *fastest* method for determining the exposure. For sometimes the situation is such that the lighting conditions are consistent from one picture to the next, over a significant period of time. In that case one need only use the light meter once to determine the proper camera settings. For subsequent pictures, the camera can be pre-set, with no metering at all.

5.3 Incident-light metering

A completely different approach to metering is to directly measure the light *falling on the subject*, rather than the light reflecting off the subject and arriving at the camera. An *incident light meter* (figure 5.3) is almost always hand-held, separate from the camera, and it is placed at the subject, so the light falling on the subject falls on the meter. This avoids altogether the issue of the light meter not knowing what the subject is supposed to look like, and so no gray card is needed. Snow will look like snow; asphalt will look like asphalt.

An incident light meter is, of course, helpless in the face of a subject that emits its own light. It is useless, for example, for metering a sunset. And if the subject is distant, it may be difficult to get the meter physically in the same position as the subject. It does work, however, to place the meter in a different (and more convenient) location, so long as the same light falling on the subject also falls on the meter.

A hand-held meter, whether it is reflected-light or incident-light (some allow for both), has all of the adjustments present on a camera—ISO, shutter speed and aperture. Some are null meters, some have a needle that reads EV values, and some read shutter speed, aperture and ISO directly with a digital display. For all of these methods of communicating with the photographer, it is possible to easily choose any of the combinations of aperture and shutter speed that give the same EV.

5.4 Flash

The use of flash dates back to the earliest days of photography. It can be quite convenient to carry around one's own source of light, especially when it is mounted directly on the camera itself. See for example, *The Endurance by Night* by Frank Hurley, in *The Photography Book*. But the use of flash requires a completely different approach to metering.

Figure 5.3. A modern digital incident-light meter. This model works with either ambient light or flash.

For most modern flash units, the energy is delivered in a burst of light that lasts for only about 1/1000 s. For this reason it is ineffective to use the shutter speed to limit the exposure when the subject is lit by a flash. Obviously if the source of light (the flash) is only giving off light for 1/1000 s, it hardly matters whether the shutter speed is 1/30 s or 1/250 s. Both shutter speeds would allow *all* of the energy of the flash pulse to enter the camera. And so *when using a flash alone to illuminate the subject, the exposure is controlled entirely by the flash, ISO setting and aperture; shutter speed does not enter into the calculation*.

This is not to say that shutter speed is completely unimportant for all flash photography. For example, there may be ambient light falling on the subject from sources other than the flash. Also, most cameras have a maximum shutter speed for which flash is usable. This is called the *sync speed* of the camera. If one attempts to use flash at a higher shutter speed, the camera will not be able to synchronize the shutter with the flash, or the shutter may not expose all of the image frame during the flash exposure. For a digital or 35 mm SLR type camera, the maximum sync speed is typically between 1/90 to 1/250 s, depending on the camera.

Recall that the exposure is determined by the total energy delivered to the light detector. For non-flash sources of light, we achieve this by using a shutter to limit the time of the exposure by a light source of a certain power. And so energy equals power multiplied by time. For flash, however, we can't do this; instead it is a pulse of light with a particular *total energy*. In effect, the short duration of the flash itself becomes our camera shutter. Thus for a given brightness of light from the flash on the subject, the exposure is set at the camera solely by the ISO and aperture. Most flash units have the ability to vary the power and duration (and thus the total energy) of their flash pulse. There may be a direct setting for this on the flash unit, but it may also be done automatically as part of the flash metering process.

5.4.1 Distance and flash

There is another complication for metering that arises whenever the flash is mounted on the camera. Recall that the *specific intensity* of light is independent of distance. And so for a particular amount of light incident on a particular subject, one sets the camera exposure the same whether the camera is close to the subject or far away. Not so for a camera-mounted flash!

Remember the inverse square law? The light coming from a small source of light decreases with the inverse square of the distance from the source. It is true that *once the amount of light falling on the subject is fixed*, the exposure setting of the camera does not depend on distance to the subject. But if the flash is mounted on the camera, *there is less light from the flash on the subject if the subject is farther away from the camera*, because the subject is also further away from the flash. Thus a given setting of flash brightness, ISO and aperture will expose subjects at different distances by different amounts.

Since the light from the flash varies as the inverse of the distance *squared*, this is a huge effect. Twice the distance means only one fourth the light, corresponding to two whole exposure steps. A common flash fail arises when someone uses flash to take a picture of a football game. The flash exposes the backs of the nearby people in the next row a *lot*, and the distant game field not at all. The result is a couple of bright, overexposed backs framing a completely dark football field.

5.4.2 Flash metering

As is the case for metering in general, one can take either an incident-light or a reflected-light approach to metering flash. But an ordinary light meter can not be used for flash at all; a special *flash meter* is needed. The reason is simple; a flash meter measures *energy*, while an ordinary light meter measures *power*, which is energy *per time*.

An incident-light flash meter is a hand-held meter placed at the subject. The flash is triggered in the same way as it would be for the picture, and the meter tells the photographer the appropriate camera settings for ISO and aperture (remember that shutter speed does not affect the flash exposure). Some flash meters can be set in a 'ready mode' whereby the flash itself triggers the meter. This is convenient, but usually slightly less accurate, as some of the energy of the flash is needed to trigger the metering process, and so goes unmeasured (the flash unit corrects for this bias as best it can). Sometimes a wire is connected between the flash unit itself and the meter, and the meter itself triggers the flash. The most sophisticated (and expensive) process uses a small radio transmitter mounted in the meter and a radio receiver mounted in the flash unit. This saves one from tripping over wires.

In a portrait studio, there may be many flash units used at once lighting the subject from different angles, and there may be no flash mounted directly on the camera at all. In this case, use of a separate, incident-light flash meter is essential. The separate flash units can be easily synchronized with each other with an inexpensive device known, most unfortunately, as a *slave unit*. A slave unit mounted on a flash triggers the flash whenever it 'sees' the light from another flash. Thus only one of the flash units need be triggered by the camera or flash meter, and all others will trigger themselves when the 'master unit'[2] flashes.

Since flash exposure depends so critically on the distance between the flash and the subject, built-in, reflected-light meters are commonly used when a flash is mounted directly on a camera, and this allows for fully-automatic flash exposure. But this arrangement can also be accomplished in a fully manual mode. If the energy output of the flash is known, it is possible to calculate the proper camera ISO and aperture settings given the distance to the subject. *Once the ISO and flash power are set, there is then a particular aperture setting for a given subject distance*[3].

[2] Another very unfortunate term.

[3] There is another important factor, and that is how well-focused is the light coming from the flash. Does the flash emit all of its energy in a narrow beam or does it instead allow it to spread out over a wide area?

The brightness of the flash can be represented in physical units related to joules, but photographers use a more convenient measure known as a *flash guide number*, GN. The guide number of a flash is defined by equation (5.1).

$$\text{GN} = \text{distance} \times f \tag{5.1}$$

Since f is simply a unit-less ratio of two lengths, the GN of a flash has units of distance. One important complication is that GN depends on ISO. That is, the GN for a flash is always expressed assuming a particular camera ISO setting (usually ISO 100). If one is using a different ISO setting, then the exposure must be adjusted accordingly. But given the ISO number, it is easy to use the flash GN to calculate the exposure.

For example, if our flash has a GN of 40 feet for ISO 100, then a subject at a distance of 20 feet will require an aperture of $f/2$. This is because equation (5.1) can be rearranged as follows, and 40/20 = 2.

$$f/\text{number} = \frac{\text{GN}}{\text{distance}} \tag{5.2}$$

Most on-camera flash units, however, employ some kind of automatic exposure using a built-in reflected-light flash meter called a *thyristor*. The thyristor, usually located on the flash unit itself, measures the accumulated energy flux of the light returning to the camera as time passes. Thus the flash can begin the flash pulse, and the thyristor can automatically turn off the flash when the correct amount of light has returned to the camera from the subject. This all happens in only a tiny fraction of a second.

Of course, what is the correct amount of light returning to the camera depends on the camera's ISO and aperture settings. And so if the flash is a separate unit attached to the camera via a cable or *hot shoe* (see figure 5.4), then the camera and flash must communicate with each other somehow. First, the camera shutter must tell the flash to fire only when the shutter has opened. And the flash must somehow know the ISO and aperture settings of the camera in order for the thyristor to know how much light must return to the camera. In some cases the camera and flash communicate to each other electronically, but it may involve manually turning a dial on the flash unit to 'dial in' the camera aperture and ISO settings.

Some cameras allow for TTL flash; the thyristor is mounted inside the camera itself, and it measures only the light returning from the subject directly through the lens. This may seem impossible; remember that the thyristor measures the light entering the lens *as the picture is being taken*. It then turns off the flash (in a tiny fraction of a second) once enough light has returned for a proper exposure. But if one is taking a picture, the light coming through the lens is going directly to the detector, whether it be film or a light-sensitive digital chip. So how then can that light also go to the thyristor? Wouldn't the thyristor just block the light and not allow it to strike the detector?

If the thyristor were placed directly in the light path to the detector, it would indeed interfere with the picture. Instead the thyristor for a TTL flash meter is

Figure 5.4. The small metal bracket on top of this SLR camera is a *hot shoe*. The flash unit shown to the right can be mounted on the camera, and they communicate with each other by means of the electrical contacts in the hot shoe which connect to those on the bottom of the flash unit (which is shown upside down). The large center contact synchronizes the flash with the camera shutter, while the other contacts convey information specific to the manufacturer of the camera and flash. For this reason a flash from one manufacturer may not work on the camera of a different manufacturer.

located inside the camera off to the side of the detector, just outside the light path from the lens. A real light detector only detects a fraction of the light that falls on it; some light is inevitably reflected off the detector surface. It is this little bit of reflected light that the TTL flash thyristor measures. And so a TTL flash meter in a film camera sees the little bit of light reflected off the shiny film surface as the exposure is made.

TTL flash metering is more complicated in that the thyristor inside the camera must communicate with the flash. This is not much of a problem if the flash is built-in to the camera itself, as with most point-and-shoot cameras. But if it is a separate flash unit mounted on the camera via a hot shoe, the flash and camera must be compatible with each other. But TTL flash has a big advantage for cameras with interchangeable lenses, since one can switch between wide-angle and telephoto modes and only the part of the returning light that corresponds to the actual picture is measured.

5.4.3 Fill flash

The most complex use of flash is the mixing of light from flash and non-flash (ambient) sources. A common example is what is known as *fill flash*. Consider, for example, the case of taking a close-up picture of a person's face with a beautiful sunset in the background. If the camera is facing the sunset, then inevitably, there is very little daylight shining on the face. And so if one makes the correct exposure for the face, the sunset will be overexposed. And if one sets the exposure for the sunset, then the face will be nothing but a dark silhouette.

Fill flash to the rescue! The key point is that the flash exposure is set (for a given ISO) only by the aperture, while the exposure for the non-flash light is set by a combination of both aperture and shutter speed. And so one can choose an aperture setting that will give the proper flash exposure of the nearby face. Then select a shutter speed that, in combination with the aperture already chosen, gives the proper exposure for the sunset.

IOP Concise Physics

The Physics and Art of Photography, Volume 2

Energy and color

John Beaver

Chapter 6

VLS detector photography

Here and in chapters 7 and 13 and appendix C, I stray off the beaten path and describe what I call *very low sensitivity (VLS) photography*—a set of techniques that are rarely used by practicing photographers. I do so for two reasons—first, VLS photography brings to the fore many of the physical principles that underly photography, and second, it is inexpensive, accessible and a lot of fun.

When thinking about the different ways in which exposure is determined in photography, it is interesting to consider eight basic categories, with exposure time as the focus:

1. A shutter speed of a fraction of a second results from a combination of a sensitive detector (ISO 6-1600) and relatively small ($< f/64$) focal ratio, combined with ordinary lighting. Sunny 16 at ISO 100 is a good example; this is the exposure regime for most photography.
2. A medium-length exposure, typically of several seconds, results from a small focal ratio combined with significant light and a lower-sensitivity detector. This is the lens-cap-as-shutter photography of processes preceding modern silver gelatin emulsions. Examples are tintype, ambrotype, daguerreotype and wet collodion photography.
3. A long time exposure of many minutes results from a combination of very dim light with a sensitive detector and small focal ratio. A night-time exposure with an ordinary digital or film camera is a good example.
4. A long time exposure of many minutes results from a combination of a very large focal ratio ($> f/100$) with a moderately sensitive detector (ISO > 1) and bright daytime lighting. This is pinhole photography.
5. A long time exposure (many minutes to possibly hours) results even with the combination of significant lighting and a small focal ratio, because a detector of *very low sensitivity* (VLS) is used.

doi:10.1088/2053-2571/aae504ch6

6. A VLS detector is used in conjunction with a very large focal ratio (such as a pinhole), under bright light conditions. This results in a days-long exposure time.
7. An *extremely-low sensitivity* (ELS) detector is used in conjunction with small focal ratios and bright light, in an exposure that takes many hours or even several days or weeks. The very first in-camera photograph is an example, as well as some other techniques we will consider in Volume 3 of *The Physics and Art of Photography*.
8. A VLS detector is used in conjunction with a very large focal ratio (such as a pinhole) under very dim light. To my knowledge, this has never been attempted.

The first and third categories involve traditional cameras, lenses and detectors (ordinary film or digital detectors). The second category is of historical interest, but as we shall see in Volume 3 of *The Physics and Art of Photography*, it is of contemporary interest as well for art photographers. The fourth category, pinhole photography, was discussed in Volume 1 of *The Physics and Art of Photography*. It has enduring popularity because the camera construction is so simple. The eighth category seems to be a non-starter, and we will have a little bit to say about the sixth and seventh categories in Volume 3. But it is the rather unfamiliar fifth category we consider here—what I call *very-low sensitivity (VLS) photography*.

We intentionally choose a detector that is of such low sensitivity that time exposures of at least several minutes are required for an exposure, even with ordinary outdoor lighting conditions and typical focal ratios. As is the case for pinhole photography, no shutter is needed because of the long time exposures. But there is the added feature that the low-sensitivity light detector is more easily handled, and the camera construction is, in some ways, even simpler than for pinhole photography.

What we want is a detector that is just barely sensitive enough that it can be used to take outdoor photographs with a moderately-fast lens, but has a sensitivity low enough that it can be handled in dim ambient light for brief periods of time without risk of significant exposure during handling. This allows for a greatly simplified camera design. In particular, *no mechanism is needed for keeping the detector in total darkness before and after the exposure*. And for brief handling of the VLS detector outside of the camera, no photographic darkroom is needed—only a 'dimroom.' This discussion would be simply academic, except that there really *are* detectors with these properties, and we will see how to use them in chapters 7 and 13 and appendix C.

6.1 An exposure benchmark for VLS photography

To set up a practical benchmark for such a VLS detector, we might assume the following:

1. We must handle the detector in subdued (but not totally dark) lighting for 30–60 s. This should provide plenty of time to prepare the detector and load

it into the camera before the exposure, and to transfer it from the camera to a dark bag after the exposure.

2. While handling the detector, we can shield it from light well enough—by working in a dim room, under a black sheet, or in a 'dim tent'—to reduce the lighting by 10 exposure steps dimmer than the lighting used to take the picture. A difference of 10 steps in lighting is equivalent to the difference between bright sunlight ($EV_{100} = +15$) and very-dim home interior lighting ($EV_{100} = +5$). One can accomplish this difference in the field on a sunny day simply by going into the shade and working under a dark sheet.
3. We want the unwanted exposure of the detector during handling to be at least 5 exposure steps less than the correct in-camera, mid-tone exposure of the subject. That is, whatever exposure it takes to bring the subject up to a mid-tone density in the camera, we want to expose the detector during handling by no more than 5 exposure steps *less* than this. This is to ensure that the dim-light handling has little detectable effect on the final picture.

And so let us consider the time, t_H, to expose the detector, while handling it in dim light, to some particular not-quite unacceptable background density. And let us compare this to the time, t_C, to expose the detector in the camera from reflected light off an 18%-reflectivity subject, and achieve a mid-tone density on the detector for our exposure.

Let us also assume we are using a camera with a focal ratio, f, and let us define a couple of new variables. Let Z be the number of exposure steps less than the subject exposure that we can accept as a barely-acceptable level of exposure during handling. And so, for example, if $Z = 8$ it would mean that whatever exposure achieves the desired mid-tone density of the subject, we would need to keep the unwanted exposure resulting from handling the detector to eight exposure steps less. This translates to $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^8 = 256$ times less exposure. For our benchmark we will choose $Z = 5$.

Also, let us define N to be the number of steps by which we can dim the light while handling the detector, by for example putting a black sheet over our head. I have already indicated that $N = 10$ is a reasonable assumption, but let us keep this as a variable so we could plug in other numbers, at the end, if we so choose. Since N refers to the number of *steps* in illuminance, then $N = 10$ means that the illuminance on the detector while handling is 2^{10} times less than the illuminance on the *subject* during the in-camera exposure.

Note the distinction between Z and N. They both refer to the difference between the case of handling the detector in dim light and exposing the detector in the camera to an image of the subject. But Z is in regards to the difference in *exposure* that results from those two cases, while N refers to the difference in *illuminance* on the detector while handling and on the *subject* while its picture is taken.

We now have everything we need to derive an equation that relates t_C to t_H. What this equation tells us in practice is the following: if I can successfully limit the dim-light handling of the detector to t_H seconds, then how long would be the exposure time, t_C, in order to correctly expose our subject?

One other crucial assumption we must make is regarding the subject. Does it reflect a lot of light or only a little? Since I have specified that our exposure of the subject should be whatever is needed to yield a mid-tone density in its image on the detector, then we should assume a subject that reflects light as some mid-tone level of gray. The photographic standard for this turns out to be a subject reflectivity of $R = 0.18$, or 18% reflectivity.

We have already seen (equation (2.52)) that the illuminance, I_D, on the detector inside the camera is related to the illuminance, I_{SUB}, on the subject by:

$$I_{\text{SUB}} = \frac{4f^2}{R} I_{\text{D}} \tag{6.1}$$

where R is the reflectivity of the subject (18% in our example). And we also have the most-basic of facts about exposure:

$$\text{exposure} = \text{illuminance} \times \text{time} \tag{6.2}$$

We can write this equation for three cases: inside the camera, handling the detector, and exposure by the illumination directly shining on the subject. I will represent these with subscripts 'D', 'H' and 'SUB,' respectively. And so we have:

$$E_{\text{SUB}} = I_{\text{SUB}} \times t_{\text{SUB}} \tag{6.3}$$

$$E_{\text{H}} = I_{\text{H}} \times t_{\text{H}} \tag{6.4}$$

$$E_{\text{D}} = I_{\text{D}} \times t_{\text{D}} \tag{6.5}$$

where E here represents exposure (not energy!), I represents illuminance, and t represents exposure time.

We also have our assumptions that the illuminance on the detector while it is being handled is less than the illuminance on the subject by a factor of 2^N, and that the exposure due to handling the detector in our subdued light is less than the exposure in the camera by a factor of 2^Z. And so we have:

$$I_{\text{H}} = \frac{I_{\text{SUB}}}{2^N} \tag{6.6}$$

$$E_{\text{H}} = \frac{E_{\text{D}}}{2^Z} \tag{6.7}$$

We can combine these relations as follows. Equations (6.6) and (6.1) give:

$$I_{\text{H}} = \frac{4f^2}{R2^N} I_{\text{D}} \tag{6.8}$$

Combining this with equations (6.4), (6.5) and (6.7) gives:

$$\frac{E_{\text{H}}}{t_{\text{H}}} = \frac{4f^2}{R2^N} \frac{E_{\text{D}}}{t_{\text{D}}} \tag{6.9}$$

$$\frac{\cancel{E_H}}{t_H} = \frac{4f^2}{R2^N}\frac{\cancel{E_H}2^Z}{t_D} \tag{6.10}$$

$$t_D = \frac{4f^2}{R}\frac{2^Z}{2^N}t_H \tag{6.11}$$

or:

$$\boxed{t_D = \frac{4f^2 2^{(Z-N)}}{R}t_H} \tag{6.12}$$

And so we now have an equation that relates the time that we handle the detector in subdued light compared to the time that the detector is properly exposed in the camera. We assume that we handle the detector in light subdued by a factor 2^N, and that this dim-light handling results in a (hopefully insignificant) exposure that is Z exposure steps less than the proper in-camera exposure of the subject.

For reasons I will explain in the next section, let us choose a focal ratio of *f*/5.6 as our benchmark for VLS photography. And, so given our benchmarks values of $N = 10$, $Z = 5$, $R = 0.18$ and $f = 5.6$ we have:

$$t_D = \frac{4f^2 2^{(Z-N)}}{R}t_H \tag{6.13}$$

$$t_D = \frac{4 \times 5.6^2 \times 2^{(5-10)}}{0.18}t_H \tag{6.14}$$

$$t_D = 21.7\ t_H \tag{6.15}$$

This means that the in-camera exposure would need to be about 22 times longer than the exposure while handling the detector in dim light. This equation is true for any detector; we have so far made no assumptions about the sensitivity of the detector. But it is only for a VLS detector that it would be of any practical use. For a typical film detector, with ISO=100 for example, the sunny-16 rule indicates that on a sunny day we would have, using an *f*/5.6 lens, an exposure time in the camera of $t_D = 1\ /\ 1000\,\text{s}$ (−3 steps from 1/125 s, and +3 steps from f/16). Under our assumptions, equation (6.15) states that we would only be able to handle the detector for 1/22 of that already-tiny amount of time. Clearly that would not be practical.

But for a VLS detector, the exposure time under the same conditions would be much longer. For example, what if the very-low sensitivity of our detector means we need an exposure time on a sunny day at *f*/5.6 of not 1/1000s, but instead much, much longer than that—say 15 minutes (900 s). Then equation (6.15) indicates that we could handle the detector in our subdued light for 900 s/ 21. 7 = 41.5 s without adversely affecting the exposure.

If the detector is even less sensitive than this, or if we do a better job of shading the detector while transferring it to the camera, or if there is less than full sunlight, then handling is even less problematic. On the other hand, if the sensitivity is *much* less than this, then exposure times begin to get uncomfortably long, and the range of conditions that can be photographed in practice diminishes greatly. There are photographic processes that have sensitivities near that of our benchmark, and we will see shortly how to use them to take pictures.

6.2 VLS photography in context

Given our benchmark sensitivity for VLS photography of a sunny day exposure in about 15 min at *f*/5.6, the 900 seconds of our VLS exposure time is 9×10^5, roughly one million times, longer than the 1/1000 s ISO 100 exposure. It is easy to verify that this is roughly a difference of 20 exposure steps ($2^{20} = 1.05 \times 10^6$). And so our benchmark VLS detector has a sensitivity of about one million times less than ISO 100, corresponding to an ISO of approximately 1×10^{-4}.

To illustrate this in a different way, consider table 6.1. Here I have calculated exposure times in a camera at *f*/5.6, and also incident-light exposure times, for detectors of different ISO sensitivities, both for sunny-16 ($EV_{100} = 15$) and dim room interior lighting ($EV_{100} = 5$) conditions. The third column gives the exposure time in a camera at *f*/5.6, while the fifth column gives the equivalent exposure time if the light-sensitive material were to be placed directly in the same incident light as the subject.

If we pick a range of sensitivities centered on our ISO 1×10^{-4} benchmark, then a VLS detector would need to have an effective ISO in the range between 1×10^{-3} and 1×10^{-5}. At the upper end of this sensitivity, it would take about 2 min to expose the detector in the camera on a sunny day, but it would take about 3 min *to give the detector the same exposure* outside of the camera in dim room lighting.

Of course, in handling the light-sensitive material outside of the camera, we want to give it much *less* exposure than this so as not to ruin our picture; I have already suggested that 5 exposure steps (32×) less is a safe level of out-of-camera exposure. And so at ISO 1×10^{-3} we would have only 180/32 = 5.6 s to handle our detector outside of the camera. This may be possible, but it would likely be difficult.

At ISO 1×10^{-4}, close to our original VLS benchmark, we would have an exposure time of 20.4 min in the camera, and it would require 30 min outside of the camera to give the detector the same exposure in dim light. And so we could safely handle the detector for 1800/32 = 56 s in our dim room. At one tenth this ISO (1×10^{-5}), on other hand, handling the detector in a dim room would be quite easy—but the exposure time outside the camera would then be measured in hours rather than minutes.

What about the other parts of table 6.1? The top of the table, with ISO between 100 and 1000, is the range for the detectors used in most film and digital cameras, while ISO sensitivities in the single digits are typical of the enlarging papers used in the darkroom to make prints from negatives. Lower sensitivities, in the range of 0.1–0.01, are more typical of some of the 19th-century photographic processes used before the advent of faster silver–gelatin emulsions. The seconds-long exposure times were often made simply by uncovering and then covering the lens by hand with a lens cap. The very bottom of the chart on the other hand could be called ELS (extremely-low sensitivity) photography. We will consider one example in

Table 6.1. Effect of detector ISO on exposure times for both sunny-16 (EV_{100} = 15) and dim room interior lighting conditions (EV_{100} = 5). The third column shows exposure times assuming an in-camera reflected-light exposure at *f*/5.6 of an 18%-reflectivity subject. The fifth column gives the exposure time to achieve a similar density if the detector is exposed directly to the same *incident* light (without a negative). Reciprocity failure has been ignored in these calculations. The rows marked 'VLS' have outdoor, in-camera exposure times at *f*/5.6 that are typical of traditional pinhole photography. But in this case the detector itself can be safely handled for a brief period in dim room lighting conditions.

EV_{100}	*f*/	*t* (in camera)	ISO	*t* (incident)	
15	5.6	1/8000s	1000	1.8×10^{-7} s	Film
5	5.6	1/8 s	1000	1/6000s	Film
15	5.6	1/800 s	100	1.8×10^{-6} s	Film
5	5.6	1.25 s	100	1/600 s	Film
15	5.6	1/80 s	10	1.8×10^{-5} s	Paper
5	5.6	12.5 s	10	1/60 s	Paper
15	5.6	1/8 s	1	1/6000s	Paper
5	5.6	125 s	1	1/6 s	Paper
15	5.6	1.23 s	0.1	1/600 s	
5	5.6	21 m	0.1	1.8 s	
15	5.6	12.3 s	0.01	1/60 s	
5	5.6	3.5 h	0.01	18 s	
15	5.6	123 s	0.001	1/6 s	VLS
5	5.6	1.5 d	0.001	3 m	VLS
15	5.6	20.4 m	0.0001	1.8 s	VLS
5	5.6	14.5 d	0.0001	30 m	VLS
15	5.6	3.4 h	0.000 01	18 s	VLS
5	5.6	145 d	0.000 01	5 h	VLS
15	5.6	1.42 d	0.000 001	3 m	
5	5.6	1450d	0.000 001	2 d	

Volume 3—the so-called anthotype. With rare exceptions, ELS detectors are used only for photograms, where they are illuminated directly by the light of the Sun.

In chapter 7, we consider two practical VLS detectors. What I call *ephemeral process* (EP) photography has an effective ISO right in the middle of the practical range of sensitivities for a VLS detector, very near that of our benchmark, while what I call *cyanonegative photography*, on the other hand, is at the low-sensitivity end of that range.

I note that the enlarging paper used in the darkroom, with ISO in the range 1–10, is exactly the same paper as used for EP photography, with ISO 1×10^{-4}. How can the same light-sensitive paper be used for both? We explore this question in detail in Volume 3 of *The Physics and Art of Photography*, but the short answer is that in its normal use in the darkroom, its sensitivity is chemically amplified by a *developer*; for VLS photography we use only its native sensitivity to light, without chemical development.

I have chosen for our benchmark example of VLS photography a focal ratio of *f*/5.6 on a sunny day—but why *f*/5.6? This may seem odd given the well-known 'sunny-16' rule, which uses *f*/16 as a starting point. But if we remember the purpose of VLS photography—to take pictures with a camera—then it should be clear why I have chosen a smaller focal ratio around which to center the discussion. Even at *f*/5.6, the exposure times are still minutes long, and if a 15 min exposure is required at *f*/5.6, then it would take 2 h to expose the detector at *f*/16. And so the VLS photographer is far more likely to choose a small focal ratio than a large one. With the high sensitivities of ordinary film, on the other hand, a large focal ratio is often needed on a sunny day, or else a shutter speed faster than the camera's highest setting might be required.

The ISO standard for describing the sensitivity of a photographic detector is clearly rather awkward for VLS photography, with resulting ISOs that are tiny fractions. In appendix B, I instead propose describing such detectors with a *slowness*, rather than a speed. The slowness is defined inversely to ISO speed; a greater slowness means a *less*-sensitive detector. The ISO speed is roughly equal to the reciprocal shutter number for a sunny day exposure at *f*/16 (this is the sunny-16 rule). And so an ISO 1000 detector is such that one sets the shutter at 1000 (1/1000 second) to properly expose the picture at f/16 on a sunny day. My slowness measure takes the opposite approach for the (in many ways opposite) case of VLS photography:

> The slowness of a VLS detector is equal to the proper exposure time (*not* its reciprocal) *in minutes* (not seconds) on a sunny day at *f*/5.6.

And so our benchmark VLS detector, for which a 15 min exposure would be required at *f*/5.6 on a sunny day, would simply have a slowness of 15. And if I am using a VLS detector with a slowness of 30, then I know it will take me one half hour to take a picture at *f*/5.6 on a sunny day.

Chapter 7

Ephemeral-process (EP) and cyanonegative photography

We here consider two examples of VLS photography, using detectors already introduced in Volume 2 of *The Physics and Art of Photography* in the context of making photograms (shadow prints). Here I describe their use to take a picture with a camera. Both of these processes have some of the elegance and simplicity of pinhole photography, but the end result is quite different, and in some ways they are even simpler than pinhole photography. Both are printing-out processes: the image becomes visible during the exposure to light, with no chemical development needed.

> *Cyanonegative photography* uses the antiquarian process of cyanotype, traditionally used only for making direct contact prints, not as a detector in a camera. A cyanotype makes a blue-tone negative of Prussian blue.
>
> *Ephemeral process photography* (EP) uses ordinary silver–gelatin enlarging paper for the in-camera detector. This is the paper that is used in a darkroom to make black and white prints directly from black and white negatives. But for EP photography, the paper is not used as intended. No chemical developer or fixer is used, but a non-developing accelerator is applied immediately before exposure. Once exposed, it is washed and dried, and the paper returns to its very low sensitivity state.

The inevitably-long time exposures give these processes a kinship with pinhole photography. No shutter is needed, and all motion disappears from the picture. Since a fast lens is used and the detector is in the camera only during the length of the exposure, small light leaks cannot compete with the large lens. *Thus the camera need not be perfectly light-tight*. The very low sensitivity means no darkroom is needed, only a 'dimroom.' This greatly simplifies storage, loading, unloading and processing.

Both of these processes are inexpensive and require no specialized facilities such as a darkroom. But there is, of course, a catch—*for both of these negative processes,*

doi:10.1088/2053-2571/aae504ch7

the most practical way to make a positive print from the negative is via the intermediate step of digital image capture. With some limited exceptions, these negatives must be scanned, and then printed (or not) digitally.

An appealing feature of both processes is that neither uses hazardous wet-chemistry processing. Cyanotype does use a chemical sensitizer that is brushed onto paper before the exposure. This wet sensitizer is toxic, but it can be purchased pre-mixed, and there is also a usable (but inferior) pre-sensitized cyanotype paper available. The accelerator solution used for EP photography can be made entirely from non-toxic ingredients. In fact, it can be made with *edible* ingredients: water, vitamin C, and xanthan gum (used in gluten-free baking). Both cyanotype and EP negatives are simply washed in plain water after the exposure.

7.1 Cyanonegative and EP wavelength response

Before describing how to take pictures with cyanotype and EP, it is helpful to understand what range of wavelengths the two processes are sensitive to. We can use a simple *spectrometer* to determine the wavelength response of cyanotype and EP paper when exposed to sunlight. We use a *diffraction grating* to disperse the light according to wavelength. A diffraction grating is a series of many evenly-spaced parallel lines, usually on some transparent medium. If the spacing between the lines is only several times that of the wavelength of the light itself, then different wavelengths are deflected by different angles. And so a narrow beam comprised of many wavelengths at once is spread out such that each wavelength can be detected and measured separately.

Figure 7.1 shows my (rather-crude) spectrometer. It is simply a light-tight box, with a place to put a small piece of the light-sensitive paper. Light entering the box passes through the diffraction grating, and so is deflected such that different wavelengths are sent off in different directions. And so, when this light arrives at the light-sensitive paper, each wavelength is measured separately.

For this to work, the entering beam must be narrow and *collimated*—all pointing in one direction. This is usually accomplished with lenses. But lenses are made of glass, and glass absorbs the very short-wavelength light we want to detect. And so

Figure 7.1. A simple spectrometer for recording the wavelength response of cyanotype and EP paper, as exposed by sunlight. Left: the sunlight is collimated by two narrow slits at either end of a tube. A small sighting-telescope is used (by projection, not direct viewing!) to keep the spectrometer pointed at the Sun. Right: A diffraction grating disperses the collimated beam according to wavelength, and the light-sensitive paper is exposed inside the dark box (here shown open).

my collimator is simply two slits, one at each end of a tube; geometry provides the collimation, with no absorption in lenses to worry about. This is a crude way to make a spectrometer, but we can still use it get at least a rough measurement of the overall wavelength response of our VLS light detectors.

When a beam of light of a particular wavelength, λ, passes through a diffraction grating, the light is deflected by an angle, θ, given by:

$$\sin\theta = m\frac{\lambda}{d} \tag{7.1}$$

where d is the (very tiny!) distance between the parallel lines on the grating and m is called an *order number*.

The order number, m can take on any of the values 0, 1, 2, 3, …. And this means there is not just one spectrum produced; there are many. For $m = 0$, the *zero-order spectrum*, the angle is $\theta = 0$ for all wavelengths. And so the light of the zero-order spectrum simply passes through undeflected. But some of the light instead goes to the *first-order spectrum*, given by $m = 1$, or simply:

$$\sin\theta = \frac{\lambda}{d} \tag{7.2}$$

And so for the first-order spectrum, every wavelength gets deflected by a different angle. If we know d for our grating, and we measure the angle θ with our spectrometer, we can then determine the wavelength by:

$$\lambda = d\sin\theta \tag{7.3}$$

A value of $m = 2$ gives the second-order spectrum ($\sin\theta = 2\lambda/d$), and so all wavelengths are deflected by larger angles than for the first-order spectrum. In most cases, less and less light goes into the spectra of higher orders, and so we will use only the first-order spectrum. Figure 7.2 shows four spectra recorded with the spectrometer of figure 7.1. The top is pre-sensitized cyanotype paper—blueprint paper essentially. The second was made by hand-brushing 'new cyanotype' sensitizer

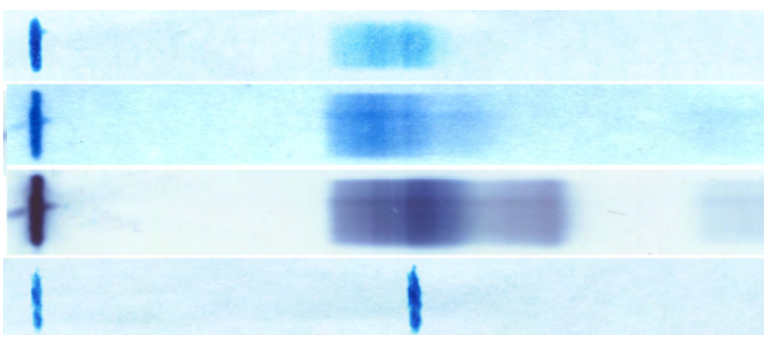

Figure 7.2. Spectra recorded with the spectrometer of figure 7.1. From top to bottom: pre-sensitized cyanotype paper (blueprint paper), hand-sensitized cyanotype paper, ephemeral process paper, and blueprint paper. The top three were exposed to sunlight, while the bottom was exposed to the single wavelength of a 403 nm laser, thus helping to calibrate the spectrometer.

(James 2016, p 212) onto paper, and the third is EP paper. The bottom spectrum is also blueprint paper, but it was exposed not to sunlight, but rather to a violet laser pointer that emitted only a single wavelength of roughly 405 nm, right at the short-wavelength edge of the visible spectrum.

The vertical line on the left of each spectrum is the zero-order spectrum; it was exposed by the portion of light that passed through the collimator tube undeflected and directly exposed the paper. The smear to the right in the top three spectra is the first-order spectrum, with the shortest wavelengths on the left and the longest wavelengths on the right. Clearly, all three types of paper are sensitive to wavelengths both longer *and shorter* than our 405 nm laser, and so they are sensitive to ultraviolet light as well as visible light.

The beginning of the second-order spectrum is just barely visible at the far right edge of the third spectrum. It corresponds to the same wavelengths as the first-order spectrum, but repeated at larger angles (so most is off the paper), and much fainter.

7.1.1 Wavelength calibration

The top image in figure 7.3 shows the geometry of the spectrum, for the case of a single wavelength from a laser entering the spectrometer. If we know the wavelength

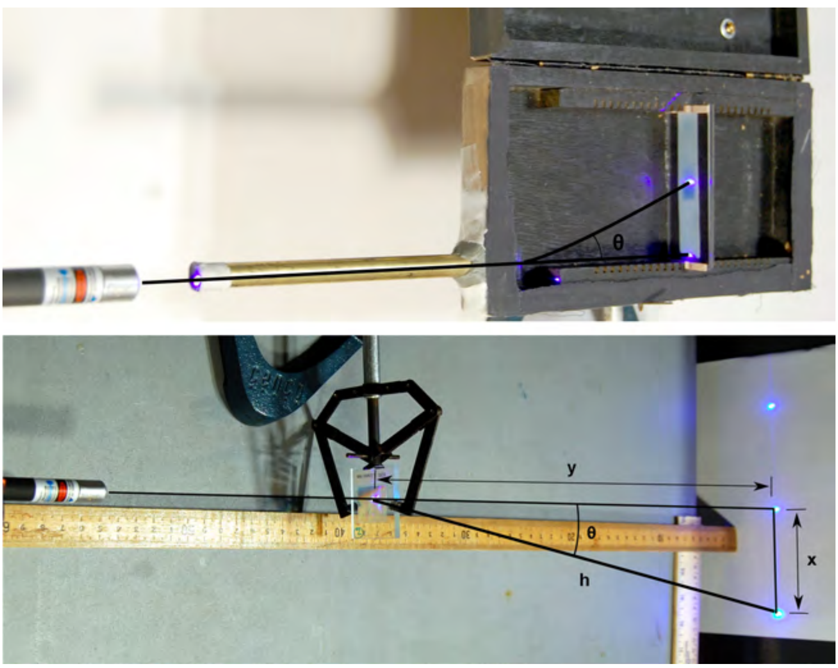

Figure 7.3. Top: a violet laser pointer, of one single wavelength, is projected through the spectrometer, and the light is deflected by an angle θ. If the wavelength of the laser is known, the measured angle can be used to determine the spacing in the lines of the diffraction grating. The spectrometer can then be used to measure unknown wavelengths. Bottom: the same laser shines through a precision diffraction grating of known spacing of 600 lines/mm. By measuring the dimensions x and y, the angle θ can be determined, and the precise wavelength of the laser pointer can then be calculated.

of the laser, then equation (7.3) and the measured angle θ can be used to find the spacing, d, of the lines in the spectrometer's diffraction grating. I used a nominally 405 nm laser pointer for these tests—but the precise wavelength of this type of laser pointer varies somewhat from one unit to the next.

For the bottom of figure 7.3 the same laser shines through a precision diffraction grating of 600 lines/mm. Thus d, the spacing between the lines, is $1/600\,\text{mm} = 1.667 \times 10^{-6}\,\text{m} = 1667\,\text{nm}$. We can turn equation (7.2) around, and use the fact that the geometry in figure 7.3 is that of a right triangle, for which the sine of the angle is the ratio of the side opposite to the hypotenuse:

$$\sin\theta = \frac{\lambda}{d} = \frac{x}{\sqrt{x^2 + y^2}} \tag{7.4}$$

$$\Longrightarrow d = \frac{\lambda\sqrt{x^2 + y^2}}{x} \tag{7.5}$$

$$\Longrightarrow \lambda = \frac{xd}{\sqrt{x^2 + y^2}} \tag{7.6}$$

where y is the perpendicular distance between the grating and the spectrum, and x is the distance along the first-order spectrum, as measured from the straight-through zero-order spectrum.

And so we can use the precision grating, with known d, to determine the wavelength of the laser, and then use that same laser to determine d for the different grating in our spectrometer. Once d is known, we can use equation (7.6) to determine λ for any measured position, x on our spectrum. For my particular example, the laser turns out to be 403 nm, very close to its marked wavelength of 405 nm.

Figure 7.4 shows the result for both the new cyanotype formula and one particular EP paper. I digitized the paper spectra by scanning them, and then averaged across many rows to output intensity of the image versus columns in the scan (the ImageJ software has a tool for this purpose).

The calibration can easily be performed with a spreadsheet. I first noted at what column the peak of the zero-order spectrum occurred, subtracting that value off so that the x coordinate would start at the proper zero. I then applied a correction factor to convert column number to a physical measurement in units of length, given the resolution at which I scanned the spectra. Finally, I applied a calibration to the data, converting x to wavelength in nanometers with equation (7.6).

The crude wavelength calibration I have described is at the appropriate level of sophistication for the crude spectrometer of figure 7.1. But how are real, non-crude spectra—taken with non-crude spectrometers—calibrated in real life? The answer is at once both simpler and more complex.

It is more complex because there are many other factors than what I have described that affect the calibration in significant ways, if one needs high precision. Many of these factors are subtle, and specific to the particular construction of the

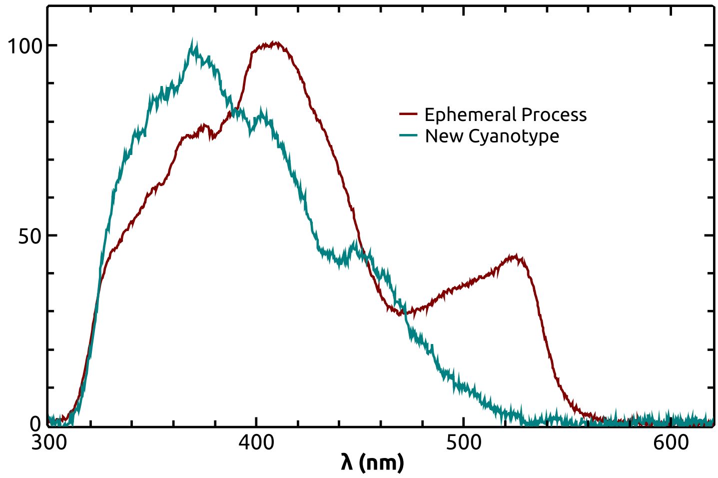

Figure 7.4. The calibrated detected spectra for new cyanotype and a particular EP paper, exposed by sunlight. The cyanotype paper responds to a smaller range of wavelength than the EP paper. Blueprint paper (a variation on the cyanotype process) has an even narrower wavelength response. Both detected spectra cut off at about 300 nm because Earth's atmosphere blocks sunlight at shorter wavelengths.

spectrometer in question. But even for our crude version, we can think of some examples.

For example, the paper was processed with water, and then dried. But doesn't wetting paper and then drying it alter its size? And so are the dimensions on the processed paper really the same as they were when it was exposed in the spectrometer? And what if the paper wrinkled during exposure? And what if there are imprecisions in the construction of my spectrometer, so that (for example) the supposedly 90° angles are slightly off? And most importantly of all, *what about other factors I have failed to think of?* This is just a small sample of the many reasons that wavelength calibration is complex.

But there is also a reason that wavelength calibration is, in practice, simple. A *gas discharge tube* emits light by exciting atoms in a gas with a high voltage. The key is that each gas emits its own set of precisely known wavelengths. And so such a tube can be used to make a *calibration source spectrum*. Iron, for example, emits many well-known wavelengths.

So simply record the spectrum of iron with the uncalibrated spectrometer. Create a table that lists, for each known wavelength of iron, the particular uncalibrated response recorded by the spectrometer. Then find a mathematical function that relates the two, and allows one to determine wavelengths in-between the iron wavelengths. Thus, the iron spectrum provides a reality check. It allows one to interpolate between those known wavelengths, to accurately determine unknown wavelengths.

The crude calibration I described for my spectrometer used, in effect, only *one* known wavelength—the 403 nm of the laser pointer. And so I had to assume the relation was simply proportional, and did not change from one part of the spectrum to another. But a scientist performing a precise calibration of a real spectrometer would use many known wavelengths for comparison, thus ensuring that the wavelength calibration is accurate at every part of the spectrum.

7.1.2 EP versus cyanotype

It is clear from figure 7.4 that the EP paper responds to a much broader range of wavelengths than does even the new cyanotype formula (the toy blueprint paper has an even more narrow range; see figure 7.2). Most of the response of cyanotype is in the near UV part of the spectrum, while the sensitivity of EP paper extends well into the green part of the visible spectrum.

It appears that both spectra cut off sharply at the same wavelength of roughly 300 nm, which might seem odd since cyanotype and EP are very different physical processes. But this is not a coincidence, and it is due to the one factor that was shared by both in making these spectra—the source of light. The spectrum of the Sun extends to wavelengths much shorter than 300 nm, but Earth's atmosphere (ozone in particular) absorbs much of that UV light. And so the steep left-hand side of the these spectra is well known to astronomers; they call it *atmospheric cutoff*.

The spectra in figure 7.4 tell us something about the wavelength response of cyanotype and EP paper. And they also tell us something about the spectrum of the sunlight that entered the spectrometer. But the curves in figure 7.4 depict neither the response curves of cyanotype and EP paper, nor do they show the solar spectrum. Rather, they are the *detected spectra*—a combination of both, as described in chapter 8, section 8.6. But for what follows, *that is exactly what we want*. For we will use cyanotype and EP paper to take pictures, and it is sunlight that we will use for our source of light. And the spectra in figure 7.4 show the overall effect of using those detectors to take pictures with that light.

7.2 Cyanonegative photography

Cyanonegative photography (Beaver 2017; Piper 2018; Ware 2016) uses cyanotype as a light detector. Because of its low sensitivity to (only ultraviolet) light, cyanotype has traditionally been used only as a printing process. But it is possible to use it as a detector in a camera for taking a picture with a long time exposure on a bright sunny day. I first began making cyanonegative photographs around 2000, and have since taught the process to students and other photographers; it is easy and inexpensive. There are, however, a few tricks:

1. Use the modern and much-faster 'new cyanotype' process, developed by contemporary photographer and chemist Mike Ware (Ware 2016, p 212; James 2016). Ready-made blueprint paper is also available from toy stores, and has a high enough sensitivity, although the tonal range is often poor.
2. Use a simple magnifying glass instead of an ordinary camera lens to focus the image onto the sensitized cyanotype paper. A simple magnifier typically has a very small focal ratio ($\approx f/2.5$), and absorbs much less ultraviolet light than an ordinary camera lens.

3. Offset the focus slightly, to correct for the fact that while we focus with visible light, the cyanotype ‘sees’ at a significantly shorter wavelength.
4. Given that cyanotype only responds to short wavelengths, the daytime sky is likely to be the brightest part of the picture by far. And thus the easiest scene to photograph is a silhouette against the sky.

Exposure times in bright sunlight at $f/2.5$, using Mike Ware’s cyanotype formula, are about 10 min to record a silhouette against the sky. Up to several hours may be required to adequately record bright sunlight reflected from a subject.

Cyanotype is only sensitive in the violet and near-ultraviolet part of the spectrum, and typical camera lenses block a significant fraction of the light at these wavelengths. Real lenses are affected by *chromatic aberration*; different wavelengths of light focus at different distances, because the refractive property of glass is different for different wavelengths. This effect is corrected by using many lenses in combination, all correcting for the chromatic aberrations of each other. But the lens elements must be made of different types of glass, some of which are denser, high-dispersion glasses (often called *flint glass*). These higher-dispersion flint glasses typically block more ultraviolet light than lower-dispersion *crown glass*.

In figure 7.5 (data from SCHOTT 2017, pp 13, 93) I plot the transmission curves of two typical glasses used for lenses, one a high-dispersion flint glass and the other a low-dispersion crown glass. For comparison I superimpose the detected spectra from figure 7.4 of both new cyanotype and EP paper. It is clear that a flint glass lens would block a significant fraction of the light to which cyanotype is sensitive. Using the ready-made blueprint paper from figure 7.2 would be even worse.

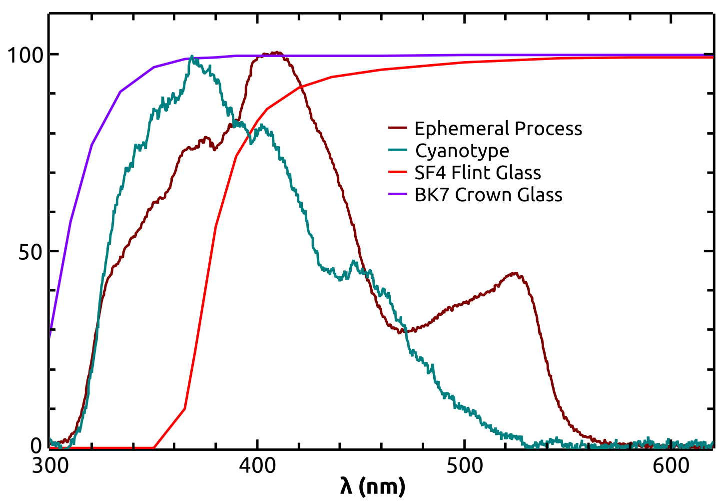

Figure 7.5. The transmission curves (relative transmission versus wavelength) for two common types of glass are superimposed on the cyanotype and EP response curves. Both types of glass transmit visible light ($\lambda > 400$ nm) well. But the flint glass is opaque over much of the wavelength range that cyanotype is sensitive to (glass data from SCHOTT 2017, pp 13, 93).

The simplest way to solve this problem is to use, instead of a normal camera lens, a simple, uncoated, single-element, inexpensive crown or soda-lime glass magnifier as a lens. Such a lens can often be found in the sale bin at the hardware store as a hand magnifier. They come in many sizes, and usually have a focal ratio of about $f/2.5$. Many of my early cyanonegative pictures were made with simple lenses of this type, mounted in a fixed manner onto a box of some sort. Getting the sensitized cyanotype paper in and out of the camera is easy; its very-low sensitivity means one must simply be quick about it and avoid exposure to direct sunlight.

Use of a simple lens would ordinarily result in a lot of chromatic aberration, but for cyanonegative photography this is not so bad as one might expect. Chromatic aberration means that different wavelengths focus at different distances, and so a blurry picture results when many different wavelengths are present at once in a picture. For an ordinary picture taken with visible light, there are wavelengths across the visible spectrum—from 400 nm to 700 nm—a range of 300 nm. But for cyanonegative photography, figure 7.4 shows that the lens only needs to focus wavelengths over a range about half that great (325–475 nm) for new cyanotype, even less for blueprint paper. And so the focus error is *less* for cyanonegative photography than it would be if the same lens were used with an ordinary light detector sensitive across the full visible range of wavelengths. The simple lens, of course, has other aberrations and distortions as well, and these still remain and give cyanonegative photography much of its look.

Figure 7.6 shows one of my earliest cyanonegative photographs, made with toy blueprint paper exposed in the 2 × 3 inch box camera shown at the right. The box has a translucent lid that acts like the focus screen in a view camera. And so it is simply placed on a tripod, and the picture is composed on the lid of the box. Then the lid is removed and the sensitized cyanotype paper is placed on the inside of the lid, facing the lens. I attach the paper to a pre-cut piece of sheet metal, which can then be held fast to the inside of the lid by placing a strong magnet on the outside.

Figure 7.6. Left: *Ash Tree No. 1*. John Beaver, 2000. When cyanotype-sensitized paper is exposed in a special camera, it produces a blue-tone negative. When digitally scanned and reversed to a positive, pleasant amber and sepia tones result. Right: a simple box camera for cyanotype negatives, using a 50 mm double-convex lens. The original negative for the image at left can be seen to the right of the camera. Also shown are the negatives for two of the cyanonegative photographs in figure 7.8, taken with cameras of larger format.

The lid, with cyanotype paper, is then placed back on the camera for the duration of the exposure.

The cyanonegative picture in figure 7.6 is mostly just a silhouette of the tree against the sky. The ground *would* be completely dark with this 10 min exposure—but it was covered with snow (except for the bare spot under the tree), which reflected the UV light from the sky. The simple lens shows pronounced barrel distortion, spherical aberration, coma and vignetting, and the texture of the grain of the blueprint paper is evident.

Before the advent of digital scanning and printing, a cyanotype negative would have been essentially useless. It will not contact print very well (Ware 2016, p 104) and since it is on paper it cannot be used in an ordinary darkroom enlarger. But a scanned and digitally reversed cyanotype paper *negative* has many intriguing features:

1. The use of a fast, simple magnifier for a lens gives an unusual image that naturally blurs toward the edge—one way in which this process is decidedly *unlike* pinhole photography.
2. The grain of the paper adds interesting texture to the image. The overall effect is reminiscent of a lithograph. If the cyanotype sensitizer is hand-brushed onto the paper negative, the brush marks can become an interesting element of the composition.
3. The cyanotype Prussian blue color of the negative, upon reversal to a positive, becomes a pleasing sepia.
4. A camera for cyanotype negatives is easy to make, and the lens need not cost more than several dollars. The cost of the cyanotype 'film' can be as low as a few pennies per exposure.

Since a long exposure is needed for a cyanonegative photograph, the relation between motion and time for the subject is an important factor. For the left-hand image in figure 7.7, I used a system of rods to keep my head in position during the hour-long exposure. For the last several minutes, I got up out of the chair and stood behind it, making a faint ghost in the picture. I was able to take a self portrait unassisted, composing the picture before I was in it. After loading the camera with cyanotype and beginning the exposure, I simply walked to the chair and took my place; the several seconds required to do so were insignificant compared to the hour's time needed to make the picture.

The cyanonegative progress can also be used—quite easily—to make photograms (shadow exposures). As a blue-tone negative *print*, this has been one of the most common historical uses of cyanotype. But we can accomplish something very different by scanning the blue-tone negative and inverting it to a positive; see the right-hand image in figure 7.7.

The top two examples in figure 7.8 reveal how the process of hand brushing the sensitizer can be incorporated into the composition. Since this is a negative process, to brush the sensitizer onto white paper is, ultimately, to brush light onto darkness—exactly the opposite of the effect achieved with a hand-sensitized *print*.

Figure 7.7. Left: *Self Portrait*. John Beaver, 2001. Right: *Dandelion*. John Beaver 2004.

Figure 7.8. Clockwise from upper left: *Trees and Ice No. 2*. John Beaver, 2004; *Bicycle*. John Beaver, 2003; *Take Me With You*, John Beaver 2002; *San Xavier Mission #2*, John Beaver 2004. The upper two images demonstrate how the brush strokes from brushing the sensitizer onto the paper can be incorporated into the composition of a cyanonegative photograph. Since this is a negative process, it is as if light has been brushed onto a dark background.

Since everything looks very different with this lens/detector combination, one often finds that the most ordinary of photographic subjects can be interesting. But it also means that a scene of great power to the eye in visible light may be rendered confusing—perhaps in an uninteresting way—by the cyanonegative process. The bottom two images in figure 7.8 are not at all representative of those scenes in real life—and that is likely much of their power.

7.2.1 Cyanonegative focus offset

Even if a simple lens does pass enough ultraviolet light to be useful for cyanonegative photography, it will inevitably have a different focal length for ultraviolet light (which the cyanotype paper sees) than visible light (which you see as you focus the camera). It is possible to correct for this by first focusing with the eye—which responds to light centered at about 550 nm. We then apply a focus correction, bringing the lens slightly closer to the light-sensitive paper, since the same lens will focus the shorter wavelength light to a slightly smaller distance.

We can easily estimate this focus offset as follows. The focal length of a simple lens is given by the *lensmaker's equation*. If we assume the lens is thin compared to its focal length, the lensmaker's equation reduces to this simple form:

$$\frac{1}{F} = (n - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{7.7}$$

where F is the focal length of the lens, n is the index of refraction of the glass, and R_1 and R_2 are the radii of the curvature of the two sides of the lens. For any particular lens then, we simply have:

$$\frac{1}{F} \propto (n - 1) \tag{7.8}$$

$$d_i \propto \frac{1}{n - 1} \tag{7.9}$$

where $\propto$ means 'is proportional to,' and I have noted that the image distance, d_i, is approximately equal to the focal length of the lens, so long as the subject is not too close. And so from equation (7.9) we can see that a higher index of refraction results in a *shorter* image distance.

We can use equation (7.9) to compare the image distance for one index of refraction as compared to the image distance with the same lens, but for a different index of refraction. That is the point of dispersion; the same piece of glass has a *different* index of refraction when used to image a different wavelength of light. And so let us use n to represent the index of refraction of the glass in the middle of the visible spectrum (550 nm), and d_i to represent the image distance for that particular n. And let us use n' and d_i' to represent the index of refraction and associated image distance for the shorter wavelength to which cyanotype is sensitive (350 nm).

When we use our eye to focus, we set the image distance according to the index of refraction at 550 nm. But the index of refraction, and thus the correct image distance, is different for the 350 nm that cyanotype is sensitive to. And so let us calculate the approximate percent error in our setting of the focus:

$$\%\ \text{shift} \approx 100\left(\frac{d_i - d_i'}{d_i'}\right) \tag{7.10}$$

$$= 100\left(\frac{\frac{1}{n-1} - \frac{1}{n'-1}}{\frac{1}{n'-1}}\right) \tag{7.11}$$

$$= \frac{n' - n}{n - 1} \tag{7.12}$$

As an example, let us assume we use a simple thin lens made of BK7 crown glass. The dispersion curve for this glass can be seen in figure 7.9 (data from SCHOTT 2017, p 13). For such glass the index of refraction at 550 nm (the middle of the visible spectrum) is 1.519, but for 350 nm, the peak sensitivity wavelength for cyanotype, the index of refraction is 1.539. And so if we assign $n = 1.519$ and $n' = 1.539$, we have from equation (7.12):

$$\%\ \text{shift} \approx \frac{1.539 - 1.519}{1.519 - 1} = 3.85 \tag{7.13}$$

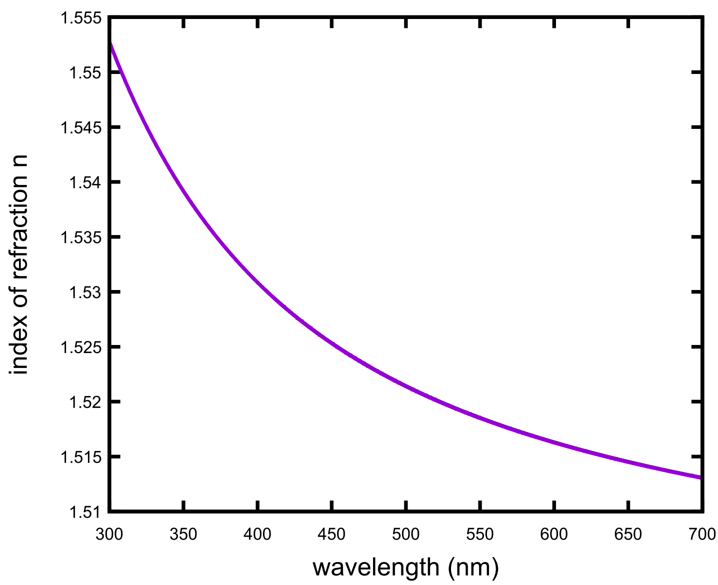

Figure 7.9. The dispersion curve (index of refraction versus wavelength) for BK7 glass, a common glass used for simple lenses. If a cyanonegative exposure is focused by eye (550 nm), the lens will not be in focus for the cyanotype detector (350 nm) (data from SCHOTT 2017, p 13).

The positive answer means that d_i', the correct image distance for cyanotype, is *less* than d_i, the image distance we set by focusing with our eyes. And so if our lens is made of this type of glass, after focusing with our eye we need to reposition the lens about 3.9% closer in order for the image recorded by the cyanotype to be in focus. Using numbers for simple soda-lime glass (the type used for bottles and windows, and sometimes for inexpensive magnifying lenses) gives a similar result. In practice, trial and error is important.

It should be unsurprising that the image distance would be smaller at shorter wavelengths. Figure 7.9 shows that the index of refraction is larger at shorter wavelengths. A larger index of refraction means that the lens bends the incoming light at a greater angle inward, and so the incoming light rays need less distance to converge to a focus.

7.3 Ephemeral process (EP) photography

Since the earliest days of photography, one of the central problems was how to make some material sensitive to light for the desired exposure, but for the final image to be *not* sensitive to light. A picture is not much good if the act of looking at it makes it disappear. It was John Herschel, the inventor of cyanotype, who came up with the first really-successful *fixer* for silver-based photography.

But the technology of digital image capture, scanning in particular, now makes it possible to make use of photographic processes that produce only a non-permanent image, processes that would have been useless in the pre-digital era. For now a permanent image can be produced from a negative even if it is impermanent and still sensitive to light—it need only have a sensitivity low enough such that it can be scanned without too much degradation from the light of the scanner.

It is well known that ordinary silver gelatin paper—intended for use in the darkroom to produce a black and white print from a film negative—will, with no developer, print out like cyanotype if exposed to enough light. The sensitivity is usually about the same as cyanotype, sometimes more, sometimes less depending on the paper and other factors. And so one *could* use it to take photographs in the same way as cyanonegative photography. The sensitivity is so low, that the image from the long exposure in the camera can be scanned to capture a digital negative.

This is a rather obvious idea, and so I quite naturally tried this back in the early 2000s, placing enlarging paper in the same homemade cameras I was using then for cyanonegative photography. I found the simplicity and chemical-free nature of the process intriguing, but I was not impressed with the results. Much of the appeal (for me) of the cyanonegative process is that the sensitizer is uneven and applied by hand, and thus introduces non-photographic detail that can result in happy accidents, a topic we explore in Volume 3 of *The Physics and Art of Photography*. But unlike cyanotype sensitizer brushed by hand onto watercolor paper, stock silver gelatin papers are manufactured with as much uniformity as possible. And so we are simply back to the traditional silver gelatin process—

Figure 7.10. *Front swing*. John Beaver 2016. Wet enlarging paper was placed in an 8 × 10 inch format camera and exposed for many minutes with a simple lens. After exposure, the paper was dried, scanned and the digital image inverted from a negative to a positive.

albeit without the chemicals—but with none of the high-sensitivity advantages of developing out[1].

This all changed in late 2015 when I realized that it is easy to *accelerate* the printing-out of silver gelatin enlarging paper by applying a simple liquid *accelerator* immediately before exposure. When printing out, the accelerated sensitivity can be many hundreds of times greater than the dry paper out of the box; see figure 7.11. After exposure, the paper is then washed and dried, and it returns to its very low sensitivity state, *making it easy to scan in order to digitally capture the negative image*. One of my first *ephemeral process* (EP) photos can be seen in figure 7.10.

The accelerator turns out to be very simple—water alone has a large effect. An oxygen scavenger such as ascorbic acid (vitamin C), and a binder (such as xanthan gum) to make the solution brush more readily are all the rest that is needed. More information on the choice of paper and recipes for the accelerator can be found in appendix C of this volume. The physical mechanism behind the action of the accelerator is discussed in Volume 3 of *The Physics and Art of Photography*.

And so we have the best of both worlds. We brush the accelerator onto the paper immediately before exposure in order to render it sensitive, far more sensitive than cyanotype. In fact the wet sensitivity brings us up to near our benchmark of the maximum sensitivity that is practical if one wants to handle the negative in a dimroom instead of a darkroom and to transfer the detector from dark bag to

[1] In more recent years, others have discovered the same process, and it commonly goes by the name *lumenbox*. Since lumenbox practitioners also typically use simple lenses, it often has a similar look to cyanonegative photography, but without the hand-brushed framing and interesting paper texture.

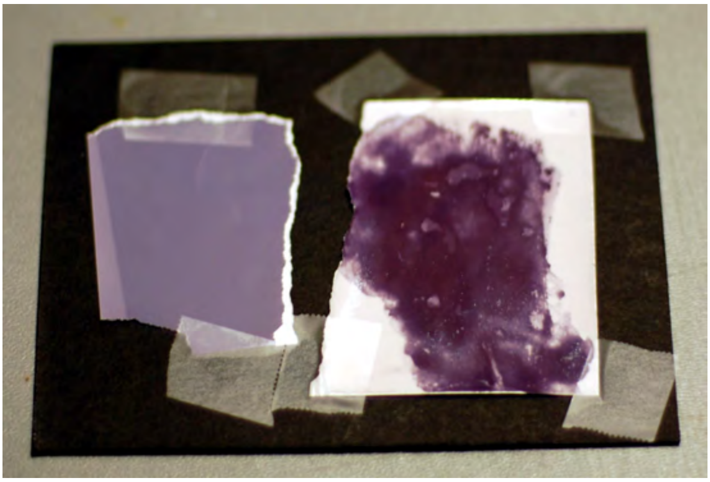

Figure 7.11. Left: a piece of dry enlarging paper after a 500 s exposure to sunlight. Right: the same paper, but with an accelerator brushed onto it, after exposure to the same sunlight for only 5 s.

camera (and back again) in shaded ambient light. But after exposure we can desensitize the negative simply by washing and drying it, thus rendering it back to a sensitivity low enough for permanent image capture by scanning. But even more importantly, since we effectively sensitize the paper by brushing water onto it, we are back to what is essentially a hand-brushed negative. Only the parts that receive the accelerator are sensitive enough for image making in the camera.

With the right choice of paper and accelerator, EP photography can achieve a usable negative of a standard scene on a sunny day at $f/4$ in only 5 min. *And the wavelength response of EP negatives is broader than that of cyanotype*, extending well into the green region of the visible spectrum. Thus, in contrast to cyanonegative photography, ordinary camera lenses can be used, there is little need for a focus offset, and reflected light from the subject is not nearly so drowned out by the sky.

Cyanonegative photography is barely worth the bother unless it is a bright sunny day, and even then a very long exposure with a fast simple lens is required to capture more than a silhouette against the sky. EP photography, on the other hand, is still possible with a (relatively inexpensive) $f/4.7$ standard lens from an old Speed Graphic camera, even on an overcast day. Even room interiors can be captured, if illuminated by daylight through a large window.

See figure 7.12 for two examples of the kind of look that is easily achieved with EP photography. In the example on the left, the patterns in the sky resulted from wrinkles in the paper negative and air bubbles in the accelerator. And of course, as is the case with all VLS photography, the exposures are long enough that motion is spread out in time or erased. In the example on the right, taken on a windy day, the ends of the tree branches moved but the trunk was motionless during the 15 min exposure. In the two examples on the right in figure 7.13, detail has been added to the sky by the physical process itself. In real life, the sky was hazy and uninteresting

Figure 7.12. A selection of EP photographs. Fiber-based paper was used, sensitized with water thickened with xanthan gum and with a small amount of ascorbic acid (vitamin C) added. Exposure times were from 10–15 min, at focal ratios from $f/4.7$–$f/5.6$. The paper negatives of 3×4 inch format were scanned and digitally inverted to positives, with adjustments only to overall levels and contrast and minor adjustments of tone.

Figure 7.13. Examples of ephemeral process photography. All of the exposures were made with a 3×4 inch format camera, with an old but inexpensive $f/4.7$ lens. Exposure times ranged from 10 to 15 min. For the two images on the right, the accelerator was applied to the paper on-site, just before exposure, under a dark cloth.

for both of those exposures; an ordinary photograph would not likely have been as successful.

As is the case with cyanonegative photography, EP photography can be accomplished with a simple lens attached to a box. But since an ordinary camera lens can be used, it is practical to use a simple large-format camera with adjustable focus. Also, the much-higher sensitivity to the EP paper, once it has been accelerated, requires more care in handling. Still, this can be accomplished in a dimly-lit room or even in the field under a dark cloth. In appendix C, I describe some ways to build or adapt inexpensive cameras for EP (and cyanonegative) photography, and some further details on the process.

7.4 Using EP photography to test the Jones–Condit equation

EP photography provides a simple method to test the validity of the Jones–Condit equation from chapter 2, section 2.3 and appendix A. The simplified version given by equation (2.52) relates the illuminance of light on the subject to the illuminance of the image on the light detector:

$$I_{\mathrm{SUB}} = \frac{4f^2}{R} I_{\mathrm{D}} \tag{7.14}$$

It is easy to see (Appendix A) that this can be related to exposure *time* as follows:

$$t_D = \frac{4f^2}{R} t_{\mathrm{SUB}} \tag{7.15}$$

where f is the focal ratio of the lens, R is the reflectivity of the subject, t_{SUB} is the time we expose the detector directly to the light incident on the subject, in order to bring the detector up to some particular level of density, and t_D is the time we expose the detector in the camera, in order to bring the image of the subject up to that same level of density.

And so, if equation (7.15) is correct, we should be able to place a subject of known reflectivity in sunlight, and expose our detector directly to that same sunlight, for time t_{SUB}. We then set up a camera, point it at the subject and take an exposure with the same type of detector, using equation (7.15) to calculate the exposure time, t_D in the camera. In the end, both the direct and in-camera exposures should achieve the same density.

For figure 7.14, I set up a gray card—a piece of cardboard designed to diffusely reflect 18% of the light ($R = 0.18$) that falls on it. I exposed EP paper for 2 s to the same light (and incident at the same angle) that was falling on the gray card. I then photographed the gray card at $f/5.6$, using the same type of EP paper in the camera. I calculated the in-camera exposure time from equation (7.15):

$$t_D = \frac{4f^2}{R} t_{\mathrm{SUB}} \tag{7.16}$$

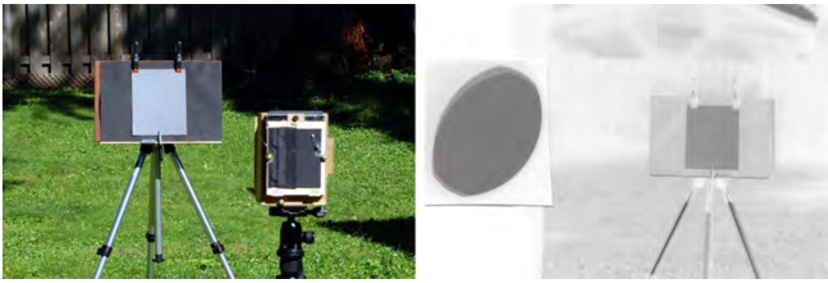

Figure 7.14. Left: a gray card of 18% reflectivity was photographed with EP photography. Right: a scan comparing the EP negative with a direct exposure of the same EP paper to the same light. The direct exposure was 2 s, while the in-camera exposure was 1394 s, as calculated by equation (7.15), given the 18% reflectivity of the gray card and the $f/5.6$ used for the exposure. The density of both is similar, although the direct exposure shows a slightly greater density. Equation (7.15) ignores vignetting and absorption of light by the glass of the camera lens.

$$t_D = \left(\frac{4 \times 5.6^2}{0.18}\right) 2\text{ s} \tag{7.17}$$

$$t_D = 1394\text{ s} \tag{7.18}$$

$$t_D = 23\text{ min }14\text{ s} \tag{7.19}$$

The image on the left shows the setup, while that on the right shows both the 2 s direct exposure and the in-camera exposure. The densities of both should be the same, and they are—almost. But it is evident that the in-camera exposure shows slightly *less* density than the direct exposure. The reason for this should be clear; my simplified version of the Jones–Condit equation ignores both vignetting and, crucially for EP or especially cyanonegative photography, absorption of light by the glass lenses.

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)

Piper J E 2018 Making cyanotypes in-camera URL http://www.alternativephotography.com/making-cyanotypes-in-camera/ (Retrieved September 2018)

SCHOTT 2017 Optical glass data sheets URL https://refractiveindex.info/download/data/2017/schott_2017-01-20.pdf

Ware M 2016 *Cyanomicon II–History, Science and Art of Cyanotype: Photographic Printing in Prussian Blue* http://www.mikeware.co.uk/mikeware/downloads.html

Part II

The art and science of color

Chapter 8

The physical basis of color

Most sources of light produce many wavelengths at once; the *monochromatic* (one wavelength only) light of a laser is the exception. And so unless one intentionally redirects different wavelengths of light from a source off in different directions, as with a prism or diffraction grating, one usually does not encounter light of only a single wavelength. These single-wavelength 'colors of the spectrum' are our perceptions of *individual* wavelengths of light. Clearly, we are able to perceive much more than the pure and simple red, orange, yellow, green, blue and violet present in such a spectrum. It is the infinite variety of *combinations* of these wavelengths that is at the heart of the physical basis of color.

As an example, the pure spectral colors are quite different even from the colors of the rainbow. It is true that a rainbow shows colors that are similar to, and in the same order, as the colors of the spectrum. But the similarity ends there, as can be seen if one compares them side-by-side (see figure 8.1). Each position in a rainbow is, in fact, a mixing of many different wavelengths at once.

But even a single wavelength of light is 'perceived' differently by different detectors. Figure 8.1 also shows the pure spectrum of the same light source, as photographed by two digital cameras with different light detectors. These digital detectors are not designed for accurate reproduction of single wavelengths of light. And so it should be unsurprising that when used for this purpose, the results are very different from the perception of the human eye, and from each other. Rather they are designed to mimic human color perception as best they can *under typical circumstances encountered in photography*. Even this simple example demonstrates that we have much work to do in order to untangle the many meanings of 'color' in photography. Color is about much more than simply the wavelength of the light; in nearly every situation many wavelengths are involved all at once.

doi:10.1088/2053-2571/aae504ch8

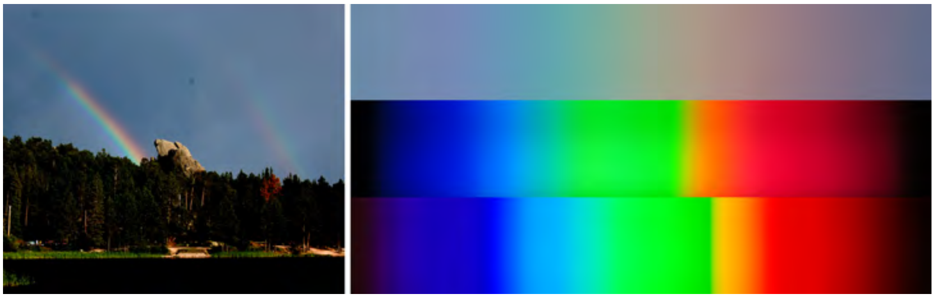

Figure 8.1. Left: a photograph, with contrast greatly enhanced, of a rainbow taken with a Nikon D40 digital camera. Right: a close-up cross section of the rainbow from the picture on the left can be seen at the top right. The center and bottom images are both of the pure spectrum of the same light bulb, but taken with two different digital cameras (Canon Powershot G15 in the center; Nikon D40 on the bottom). Clearly a rainbow is very unlike a pure spectrum of individual wavelengths. And the same pure spectrum will appear very different depending on the 'eye' of the beholder. The visual appearance to the human eye of the pure spectrum is quite unlike these two different digital captures, which also differ significantly from each other.

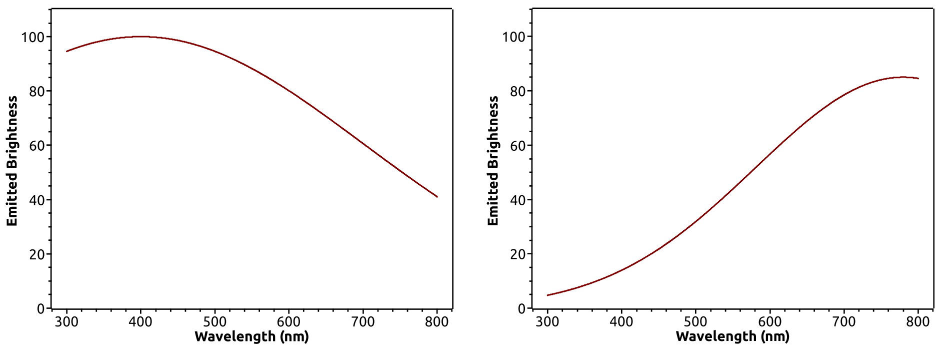

Figure 8.2. Two spectra, of two different sources of light. Both of these sources of light emit at all wavelengths in the visible portion of the spectrum. But the light source on the left emits more short-wavelength light than long, and so it would appear a pale bluish-white in color. The light source on the right, on the other hand, emits more long-wavelength light than short. And so it would appear a pale reddish-white.

8.1 Spectra and sources of light

To understand the complexities of 'color,' we need some tools for analyzing light in such a way that different wavelengths can be considered both separately and in combination with each other. Recall that the brightness of a source of light is not the only issue; there is also wavelength. And so we use a *spectrum* as a method for representing the wavelength information of light. But more to the point, since nearly every real source of light contains many wavelengths all at once, the spectrum of the source of light tells us *separately for each wavelength* how much light there is.

The most informative way to represent a spectrum is with a graph, such as those in figure 8.2. In this example the brightness of the light is represented by the vertical axis while wavelength is on the horizontal axis, with short wavelength on the left side and long wavelength on the right. The red line thus represents the brightness of the light, separately for whatever wavelength we are interested in.

The spectrum on the left in figure 8.2 is of a source of light that emits much more short-wavelength light than long-wavelength light; the example on the right is the opposite. These would then appear quite different colors to the human eye. The source of light on the left, judging by its spectrum, would appear much more bluish in color than the one on the right, which would appear reddish. Remember that the human eye is sensitive only from about 400 nm (violet) to 700 nm (red). But it is a complex question as to *precisely* what color sensation would arise when this spectrum of light enters the human eye, beyond the vague terms 'reddish' or 'bluish.' Untangling some of these complexities is the concern of much of this chapter.

8.1.1 Combinations of multiple light sources

When more than one light shines on a subject, the combined light has a spectrum equal to the sum of the two individual spectra. This means one would add the brightness from each *at each wavelength*; the result gives the brightness at that particular wavelength for the combined spectrum. Figure 8.3 shows the combined spectrum for the two light sources in figure 8.2.

When interpreting a spectrum, it is important to look carefully at the numbers. Notice that the two spectra in figure 8.2 show, respectively, a brightness of 90 and 18 (in arbitrary flux units) at a wavelength of 400 nm. Thus the combined spectrum shows, *at this same wavelength*, a brightness of 90 + 18 = 108. The combined spectrum is brighter at all wavelengths, but more importantly, the *shape* of the combined

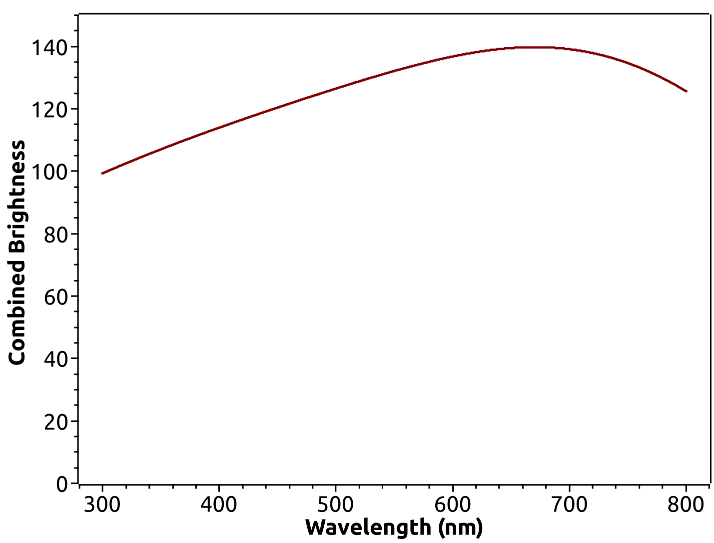

Figure 8.3. The two spectra from figure 8.2 added together. Adding the brightness wavelength by wavelength gives the spectrum for both sources of light shining on the same subject. Even though this source of light emits more light at around 600 nm (the yellow part of the spectrum) than any other wavelength, the color would appear nearly white, given the large amount of light emitted at other wavelengths as well.

spectrum clearly differs from either of the two spectra that were combined to make it. And so it would appear a different color, neither obviously bluish nor reddish. Also, notice that the percentage range of brightnesses in figure 8.3 is much smaller than for either of the original spectra. This means that the combined spectrum is more egalitarian; different wavelengths are represented more nearly equally. In fact, a source of light with the spectrum shown in figure 8.3 would appear nearly white.

8.2 Color, light sources and light detectors

The color of a photograph arises from the interaction between the spectrum of the color of the light falling on the light detector and how that detector responds to different wavelengths. Furthermore, if the camera is pointing at an object that is not self-luminous, then the light entering the camera depends on a combination of the spectrum of the light shining on the object and how that object reflects different wavelengths. To understand these interactions, we need to go beyond the general term 'color' and think of light in terms of its spectrum. And so we have three factors that are important for the color of a particular part of a photographic image[1]:

1. The spectrum of the source of light shining on the subject.
2. The percentage of light that is reflected by the subject *for each individual wavelength*. This too can be represented as a graph that is very similar to a spectrum. But instead of representing brightness of light versus wavelength, it is the *percentage of light reflected* versus wavelength. We call this the *reflection curve* of the subject.
3. The sensitivity of the light detector *for each individual wavelength*. If one shines light onto the detector, what percentage of the light is actually detected? For every real detector, the answer to this question depends on the wavelength, and it can be represented by a graph of percentage of light detected versus wavelength, called a *response curve* for the detector.

It is important to note that the second item in the above list is not relevant in all cases. For example, if one takes a picture of a sunset, the sky is both the object being photographed *and* its own source of light; it is not an object reflecting light from some other source. In this case it is just the interaction between the spectrum of the source of light (the sky) and the response curve of the detector that is important.

There is another important factor that I will mention here, but then put aside until later. A given detector may be very sensitive to a particular wavelength, 530 nm for example. This means that a large percentage of light of that wavelength would be detected by the detector. But there is also the question as to how the detector *interprets* this detection. Can it, for example, distinguish between this wavelength and, say, an equal amount of light at 560 nm? Or would it instead detect both wavelengths, but lump them together, without being able to tell how much of the total was from 530 nm and how much was from 560 nm? These last questions go to

[1] We set aside for now the important issue of the relationship between the color of, say, a digital image and the color of a *print* that is made from it.

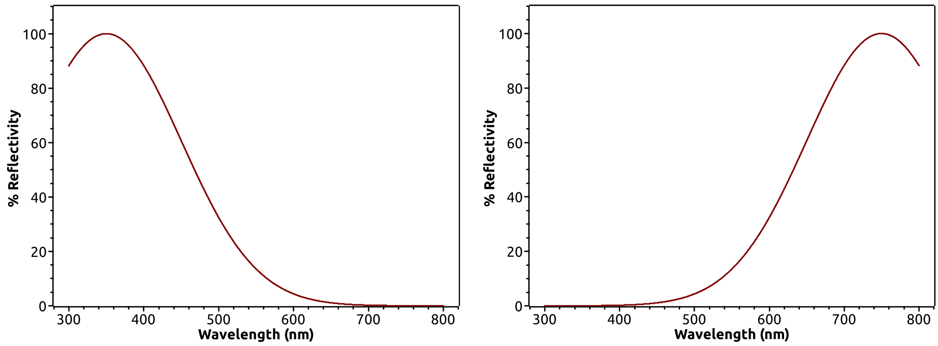

Figure 8.4. Two reflection curves. The one on the left is for blue paint while that on the right is for red paint. The blue paint reflects a large percentage of 400 nm light, but only a very small percentage of 700 nm light. The opposite is true for the red paint.

the heart of the distinction between black and white detectors on the one hand and color detectors on the other. We take up that issue in section 8.6.1 of this chapter, and in even more detail in Volume 3 of *The Physics and Art of Photography*.

8.3 The reflection curve and the reflected-light spectrum

Figure 8.4 shows two hypothetical reflection curves, one for red paint and the other for blue paint. These curves tell the percentage of light that is reflected off the paint for each wavelength. Notice that the blue paint reflects a lot of blue light (short wavelengths) and not much red (long wavelengths); the red paint does the opposite. And so a reflection curve for a surface describes what happens to light that reflects off it. How much reflects? The answer is that *it depends on wavelength*; the reflection curve is a graph that shows the percentage of light reflected versus wavelength.

Notice that a reflection curve, as in figure 8.4, looks superficially like a spectrum. But the reflection curve is hypothetical; by itself it doesn't say anything about the light reflecting off the object. It simply says, '*whatever light falls on the object* of such and such a wavelength, this is the percentage of that light that would reflect.'

And so to describe the spectrum of the *reflected* light that actually makes it to the camera, we must combine the spectrum of the light source with the reflection curve of the object the light is bouncing off of. Since a reflection curve is a percentage, we must *multiply* the two curves together. How does one take 25% of 24? 'Twenty five percent of 24' means, literally, '25 divided by 100 times 24'. In this mathematical context the word 'per' means 'divide by,' the word 'cent' means 100, and the word 'of' means 'multiply by[2].'

But neither the spectrum nor the reflection curve are simple numbers; they are graphs. Since each is telling what happens for each wavelength, we must multiply the two curves wavelength by wavelength, thus producing a third curve. This third curve then represents the *spectrum of the light reflected by the object when illuminated with*

[2] Instead of percentages between 0% and 100%, a reflection curve can also be portrayed in terms of decimal *fractions*, between 0 and 1; 0.57 instead of 57%, for example. Then one simply multiplies the reflection curve by the spectrum, without dividing by 100.

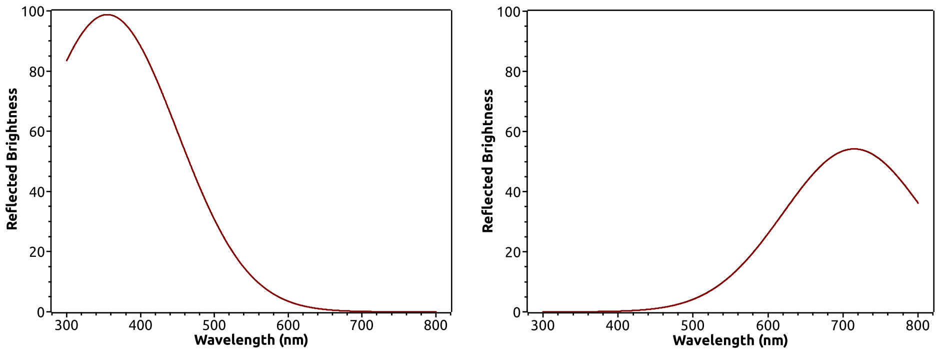

Figure 8.5. The reflected spectra for the two reflection curves of figure 8.4, when illuminated by the bluish light with the spectrum on the left in figure 8.2. The figure on the left, then, shows the spectrum of bluish light reflecting off of blue paint, while the figure on the right shows the spectrum of bluish light reflected off of red paint. Since the red paint reflects very little blue light, its reflected spectrum is much dimmer.

that particular spectrum of light, and I will call this the *reflected spectrum* of the light reaching the camera.

Let us take the bluish light with the spectrum shown on the left in figure 8.2 and shine it on both the blue paint and the red paint, with the reflection curves shown in figure 8.4. Figure 8.5 shows the resulting reflected spectra. The example on the left is of the bluish light reflecting off of blue paint; unsurprisingly, the reflected-light spectrum also shows mostly short wavelengths. The bluish light reflecting off of the red paint (the right side of figure 8.5) shows mostly longer wavelengths. This is because the light source still had a significant amount of red light; it just had much more blue. But it is clear from comparing the two graphs that when illuminated by bluish light, the blue paint reflects much more light overall than the red paint.

Figure 8.6 shows the opposite case. The longer-wavelength reddish light, with the spectrum shown on the right side of figure 8.2, illuminates the two colors of paint. And the opposite happens—the reflected light from the red paint is brighter than that from the blue light.

Notice that, for either case, the reflected-light spectrum is zero at any wavelength for which *either* the source spectrum *or* the reflection curve were zero. It matters not whether, at a particular wavelength, the source had zero light or the object reflected zero percent. One hundred percent of zero is still zero, and so is zero percent of anything.

Let us consider another example. Consider the spectrum shown in figure 8.7, of a low-pressure sodium vapor lamp. This type of lamp, sometimes used as a home security light, appears an odd yellow-orange color. A glance at the spectrum shows that this source of light is *monochromatic*; it emits essentially all of its light at only the single wavelength of 589 nm[3]. So what happens if we shine this light on both our red and blue paint? What reflected-light spectrum results?

[3] The sodium spectrum is not truly monochromatic; it consists of *two* very closely-spaced wavelengths, one at 589.0 nm and the other, weaker line at 589.6 nm. The spectrum shown here does not have enough detail to distinguish between these two closely-spaced spikes. Furthermore, sodium does emit *very* faint light at other wavelengths in the visible spectrum, as well as bright light at many wavelengths outside of the visible part of the spectrum.

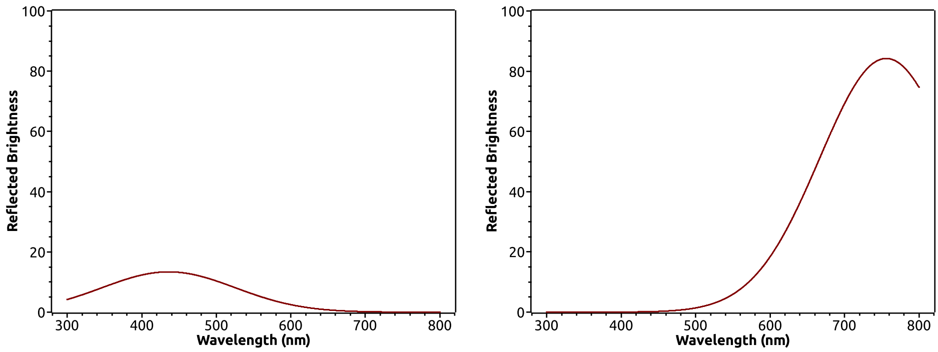

Figure 8.6. The reflected-light spectra for the two reflection curves of figure 8.4, when illuminated by the reddish light with the spectrum on the right in figure 8.2. The figure on the left, then, shows the spectrum of reddish light reflecting off of blue paint, while the figure on the right shows the spectrum of reddish light reflected off of red paint. Since the blue paint reflects very little red light, its reflected-light spectrum is much dimmer.

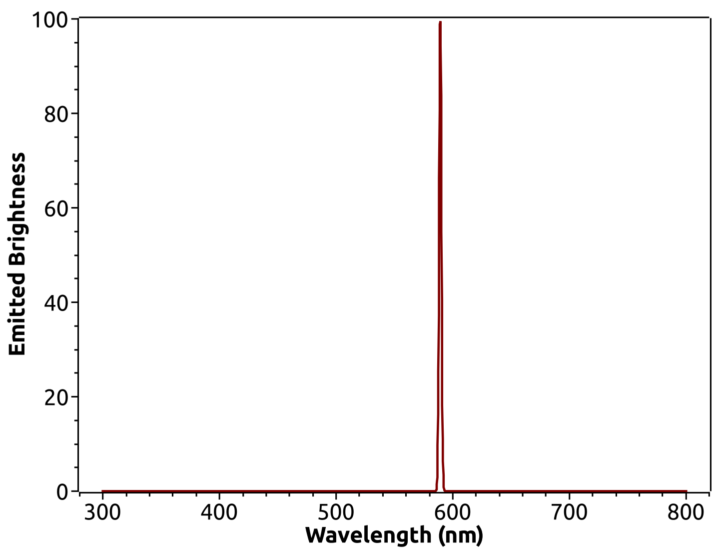

Figure 8.7. The spectrum of a low-pressure sodium lamp. In the visible range, only one wavelength (at 589 nm) is emitted. Thus the graph of the spectrum appears as a single spike at that wavelength.

To figure this out, we do what we just did—we multiply the source spectrum, wavelength by wavelength, by the reflection curve. This is easy to do because the source spectrum is so simple. For the single wavelength of 589 nm, the brightness is 100 (in some arbitrary set of physical units). For all other wavelengths, the brightness is zero. Since zero times anything is zero, the reflected-light spectrum would also have a brightness of zero at all wavelengths except for 589 nm, *no matter what the reflection curve*.

Figure 8.8. A set of colored pencils illuminated by white light (left) and the monochromatic light of a low-pressure sodium lamp (right). The individual pencils reflect either more or less of the single wavelength of the sodium light, but they cannot alter its wavelength. But when illuminated by the many wavelengths at once of white light, each pencil reflects different amounts of *different* wavelengths of light, while selectively absorbing other wavelengths.

And so in this case the reflection curves would look very similar for both red and blue paint. They would both show zero light, with a single spike of brightness at the single wavelength of 589 nm. If the spectrum of the light source is monochromatic, so too is the reflected-light spectrum, *no matter what the reflection curve*[4]. *Thus, both the red and blue paint would appear the same color (yellow-orange) when illuminated with a monochromatic sodium lamp.* The difference would be *how much* light is reflected from the two kinds of paint. A glance at the two reflection curves of figure 8.4 shows that the red paint would reflect a greater percentage of the 590 nm incident light, and thus would appear brighter.

Figure 8.8 shows two views of a set of colored pencils, illuminated by white light (with a spectrum somewhat like that in figure 8.3) on the left and by a sodium vapor lamp on the right. Under monochromatic light, the colors disappear; each reflects the only wavelength that is illuminating it. But some appear lighter and some darker, depending on what percentage of 595 nm light they reflect.

Given the image on the right in figure 8.8, we can't 'go backwards' and figure out the colors of the pencils. The brightness or darkness only tells us how much light is reflected at the single wavelength of 589 nm, and to know the color of the pencil one would need information about the other wavelengths as well. And so a pencil may appear dark under the sodium lamp because it reflects little light at any wavelength (a black pencil, for example). Or it may appear just as dark under the same sodium lamp because it is painted with a bright violet paint that happens to reflect very little light at 589 nm (but a lot of light at other wavelengths). Without more information, there is no way to tell.

8.4 Physical causes of the reflection curve

When light encounters an object, some wavelengths reflect while others do not. The reflection curve details the end result of this process, and this is why objects appear

[4] I assume here that *fluorescence* (section 8.4.3) is not at issue.

Figure 8.9. A selection of artists paints applied to both black and white backgrounds. The colors are often named after the chemical nature of the pigments used (e.g. cobalt blue, cadmium red, phthalocyanine green). The combination of the pigment and its medium (e.g. oil, acrylic, water) also affects the *transparency* of the color. Some of the paints shown here have a high transparency, and nearly disappear when placed against a dark background.

to have color when illuminated by white light. But what happens to the wavelengths that do not reflect, and what are the physical processes that cause some wavelengths to reflect strongly while others reflect poorly?

The answer is that there are several different answers, depending upon the type of interaction between light and matter. Sometimes there is a complex interplay between these different processes, but much of the time we can break it down to one of these: selective absorption, structural color, and fluorescence.

8.4.1 Pigments and dyes: color from selective absorption

Pigments are microscopic particles that selectively absorb light to produce color upon reflection. In practice, pigments may be suspended in some type of liquid or semi-liquid *medium* to allow them to be applied to a surface. In some cases (watercolor for example), the pigment particles are simply suspended in water. The pigment may remain suspended in the dried medium as a thin layer on the surface of some object, as in paint (see figure 8.9). Or, as in the case of a dye, it may be carried by its medium into the surface of a porous material such as paper, wood or fabric.

Pigments and dyes produce color by way of *selective absorption*. Some wavelengths are absorbed, while other wavelengths are not—and what is *not* absorbed is instead reflected. And so for example, if a pigment absorbs blue and only blue, then it reflects both red and green. The selective absorption properties can be described according to a reflection curve, as described in section 8.3, but one could instead say

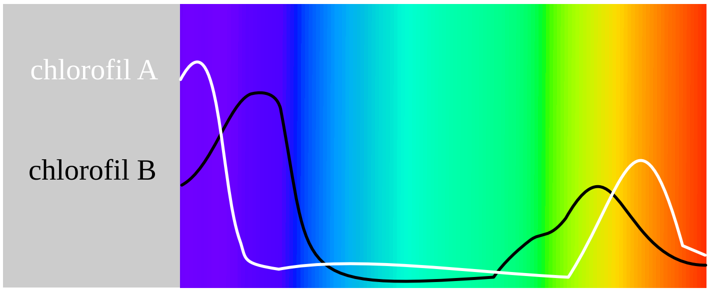

Figure 8.10. The absorption curves for two types of chlorophyll (graphic by Mix321 - Own work, Lewiński, Holak, Biologia dla liceum, wyd. Operon, GFDL).

the same thing with its inverse—an *absorption curve*. If, at a particular wavelength, the pigment reflects 80% of the light that falls on it, then it must absorb 20%. And so wherever the reflection curve is high, the absorption curve is low, and vice versa.

This selective-absorption usually occurs as part of the interaction between an individual photon of light with an individual molecule or atom. Individual photons are absorbed, raising electrons to higher energy states. But since there are only certain energy states allowed according to the rules of quantum mechanics, only certain energies of photons can be absorbed. But for light, photon energy is proportional to the frequency of the light, and so also its wavelength; shorter wavelength light is made of photons that, individually, have more energy. The precise energies of these electron energy states—and so the particular wavelengths of light that can be absorbed—are complex functions of the atomic and molecular structure of the pigment. And so a slight chemical change can lead to a dramatic difference in the absorption/reflection curve.

Most often, this absorbed photon energy is transferred ultimately to thermal energy, and the electrons return to their low energy states, ready to absorb more photons. But not always; chlorophyll—what makes leaves green—is a good example. Some of the photon energy absorbed by chlorophyll, especially that by high-energy, short-wavelength ultraviolet photons, is used to power chemical reactions that ultimately store this photon energy as chemical energy, in the form of sugars. This is the ever-so-important process of photosynthesis. See figure 8.10 for a graph of the absorption curves of chlorophyll (there are two types).

Figure 8.10 shows only the visible-light part of the absorption curve, but notice that the curve is lowest in the blue–green part of the spectrum. Since it absorbs less at these wavelengths, it reflects more—and that is the origin of the *green* in the Green, Green Grass of Home.

In order for light to do something—to cause a physical change in a material—it must be absorbed. Figure 8.11 shows a mixture of potassium ferricyanide and ferric ammonium citrate applied to a piece of watercolor paper. The properties of this are such that it reflects yellow light, and so absorbs short-wavelength blue light. But some of the energy from these short-wavelength photons is not converted to thermal energy. Rather it enables a chemical reaction that produces a new chemical, ferric ferrocyanide—the pigment Prussian blue. This is the famous *cyanotype* process

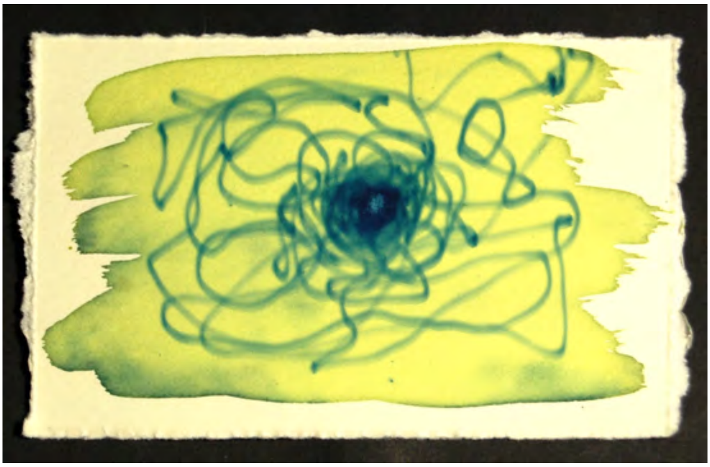

Figure 8.11. Cyanotype sensitizer, made from potassium ferricyanide and ferric ammonium citrate, was brushed onto paper. It appears yellow under white light because it reflects most wavelengths apart from the violet–blue part of the spectrum, which it absorbs. But upon exposure to short-wavelength light (in this case, from a 405 nm laser), chemical changes occur, and ferric ferrocyanide is formed—the pigment Prussian blue, which does nearly the opposite—it absorbs most visible wavelengths except for blue–green.

invented by John Herschel in 1842, and that we used as a detector in a camera for the cyanonegative photography described in chapter 7. The new pigment that is produced by this exposure to light absorbs the very wavelengths that were reflected by the original sensitizer.

8.4.2 Structural colors: interference and scattering

Structural colors may arise from materials that are completely transparent, with no significant absorption of light. Instead, the phenomena of interference, diffraction and scattering[5]—hallmarks of the wave nature of light—cause different wavelengths to be deflected in different directions. Structural colors depend not on individual molecules absorbing some wavelengths while reflecting others. Rather, they result from regular patterns in the microscopic structure of the material—in many cases structures that are too small to be seen with the naked eye but still far larger than individual molecules.

Interference colors are one example. In *thin-film interference*, light reflects off the top and bottom surfaces of a thin transparent film. These reflections may or may not cancel out, depending on the details of the thin film and the wavelength of the light. Figure 8.12 shows an interference filter. It is just clear glass, with multiple layers of microscopically-thin layers of other transparent materials applied to the surface, each with its own carefully-chosen thickness and index of refraction. The result is that some wavelengths interfere destructively—they cancel out; others do the opposite, reinforcing each other (called constructive interference).

[5] Some authors do not include scattering as an example of a structural color. I include it here because it shares many of the properties of structural colors—in particular that color is produced without absorption.

Figure 8.12. Interference colors arise not from selective absorption, but rather through wave interference. This wheel of interference filters is seen with: light shining through it (left); against a black background with light reflecting off of it (center); and with light shining through it but as seen from a steep angle (right). For light arriving at a given angle to the surface, whatever wavelengths pass through are exactly those wavelengths that are *not* reflected, and vice versa. But this interference effect depends upon the angle, and so different wavelengths reflect depending upon the angle of view.

For thin-film interference, if a particular reflected wavelength interferes destructively, then that same wavelength will interfere *constructively* when passing through. And the opposite is true. Any wavelength that cancels out when trying to pass through, will reinforce, interfering constructively, upon reflection. *And so whatever wavelengths pass through do not reflect, and wavelengths that reflect do not pass through.* The different parts of the interference filter in figure 8.12 are designed such that red, blue or green interferes constructively when passing through, while not-red, not-blue and not-green interfere constructively upon reflection. We will see in chapter 9 that there is an important relation between these transmitted and reflected colors—they are, respectively the additive primary and secondary colors.

One hallmark of interference is that the details depend upon the geometry of the incident light. And so wavelengths that interfere constructively at one particular angle may interfere destructively at other angles. The right-hand image in figure 8.12 is the same filter and back-lighting as the first image. But as seen from a different angle, the wavelengths that interfere constructively are different. And so interference colors are often shimmery—they shift when seen from different angles.

Certain feathers from some birds show interference colors as well. Figure 8.13 shows a feather of a Mallard as seen from two different angles. Micro-structures in the feather split light waves apart, reflecting them from slightly different positions. When they arrive at the viewer, some wavelengths interfere constructively while most cancel out. In this case, short wavelength blue light interferes constructively. But the effect depends critically on the relation between the angle of view and the geometry of the micro-structures in the feather. And so with a slight twist, the blue color disappears. This effect is known as *iridescence*, and it is a common way that blues and greens (and sometimes reds) are produced in the animal and insect world.

Wavelength-dependent scattering is another way to make color with materials that do not selectively absorb light. Unlike interference, which depends on a carefully-organized layering of materials at the microscopic level, scattering is a random process. As such, it is often considered its own category, rather than an example of a structural color.

Figure 8.13. This duck feather shows *iridescence*, colors that disappear or change when viewed at different angles. In this case, microscopic patterned structures in the feather cause reflected blue wavelengths to interfere constructively while other wavelengths interfere destructively. But the constructive interference requires a specific range of angles between the viewer and the light source.

Rayleigh scattering is responsible for the blue color of the daytime sky. Individual photons of light are deflected by air molecules, and this scattering effect is greater for shorter wavelengths than for longer wavelengths. The blue light you see in the daytime sky is made of photons from the Sun that would have missed you—but they were instead deflected toward you by scattering. Since the shortest wavelengths are more likely to do this, the sky looks blue.

The very similar *Tyndall effect* occurs when transparent materials contain suspended particles that are of the right microscopic or just sub-microscopic size—but still *much* larger than the individual molecules that cause Rayleigh scattering. Opalescent glass (see figure 8.14) and the blue smoke from the exhaust of a faulty car engine are familiar examples. The printed-out silver gelatin enlarging paper used for EP photography (see chapter 7) is another example. The transparent gelatin emulsion contains suspended crystals of silver halide, and these are often of the right size to cause Tyndall scattering. The precise colors that appear vary widely from one brand of paper to the next. See figure 8.14. We explore some of the photographic possibilities of this phenomenon in chapter 13.

8.4.3 Fluorescent colors

Fluorescent materials absorb individual photons of light at the atomic or molecular level. The energy from the absorbed photon raises an electron to a higher energy level, and in this way the process is similar to selective absorption (section 8.4.1). But the similarity ends there, because the excited energy of the electron is used to create *a new photon of a different wavelength*. Usually a higher energy photon—of short wavelength—is absorbed, and a lower-energy photon—longer wavelength—is emitted. Thus a fluorescent material can, for example, absorb ultraviolet light, while using some of that absorbed energy to emit its own visible-wavelength light.

Figure 8.15 shows fluorescent materials illuminated by an LED of roughly 400 nm wavelength, at the violet edge of the visible spectrum. The fluorescent materials in the picture mostly absorb the short wavelength light of the LED, and then emit

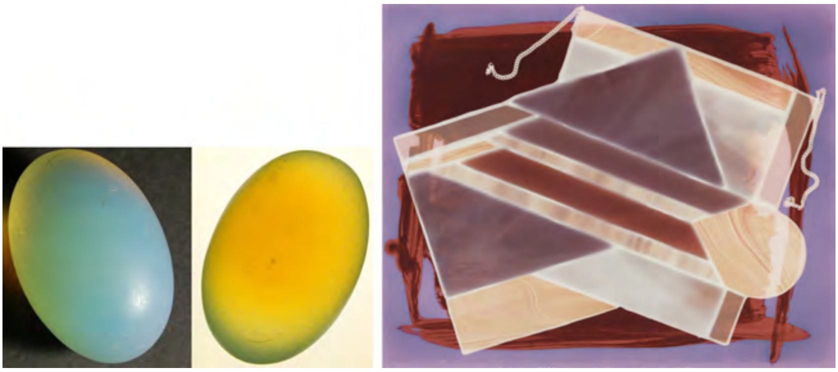

Figure 8.14. Left: Tyndall effect scattering in this transparent piece of opalescent glass causes mostly blue light to scatter backwards when illuminated from above, but (center) leaving only longer wavelengths to pass through when illuminated from behind. Right: the subtle colors in this ephemeral-process photogram arise from scattering by the regularly-sized microscopic crystals of silver halide suspended in the gelatin coating on the surface of the paper. Exposure to light of different colors, in the presence or absence of a liquid accelerator, altered the chemical and physical properties of the crystals, causing different colors to scatter back toward the viewer.

Figure 8.15. Fluorescent colors arise from molecules that absorb high-energy photons (short wavelength), and use this energy to emit their own light at longer wavelengths. The image on the left is illuminated with only short wavelength light at the very violet end of the visible spectrum. Yet some of the objects emit light at wavelengths in the middle of the visible spectrum, even though the source of light had none of those wavelengths in its spectrum. The image on the right shows the same objects illuminated with white light that contained very little of the short wavelengths needed for fluorescence to occur, and so the colors seen are due instead to selective absorption by pigments.

their own light in the much-longer-wavelength yellow–green part of the visible spectrum. Some of the materials in the picture are *not* fluorescent, and so they simply reflect the deep violet light they are illuminated with. The visible-light picture on the right, of the same objects, shows instead their ordinary pigment properties of selective absorption.

The phenomenon of fluorescence is distinct in a very important way from the other processes described in this section—the phenomenon of fluorescence cannot be described by a reflection curve. The reflection curve tells—for each individual wavelength—the percentage of the light incident on the subject that reflects. This concept is useful whatever the physical mechanism for that reflection—whether it is due to selective absorption, interference or scattering. But the whole point of fluorescence is that the subject can *emit* light at wavelengths for which no light at all is incident. As such, we will not consider fluorescence further in *The Physics and Art of Photography*.

8.5 The detector response curve

The reflection curve combines with the spectrum of the light falling on the subject to produce an altered, reflected spectrum, and *this* is the light that actually makes it to the camera. But there is another factor, as no real-world light detector measures all wavelengths equally.

Figure 8.16 shows the *response curve* for two hypothetical detectors. The response curve tells us, for each wavelength, what percentage of the incoming light is actually detected by the detector. It doesn't take a detective to discover that the detector on the left detects blue light better than red, while the opposite is true for the detector on the right.

We call the example on the left a *blue-sensitive* detector, while that on the right is a *red-sensitive* detector. When light, with whatever spectrum it has, enters the camera, it will interact with the detector according to the response curve. We can thus also describe a *detected spectrum*—the spectrum of the light *as seen by the detector*.

To determine the detected spectrum we perform the same basic operation we used to determine the spectrum of light reflected by an object. The response curve tells, for each wavelength, what percentage of light striking the detector is actually detected. So we must multiply the spectrum of incoming light by the response curve, and we must perform this multiplication separately for each wavelength. The result is another graph—the detected spectrum. We will do this in detail in the next section, but let us note here that the detected spectrum depends on *both* the spectrum of the

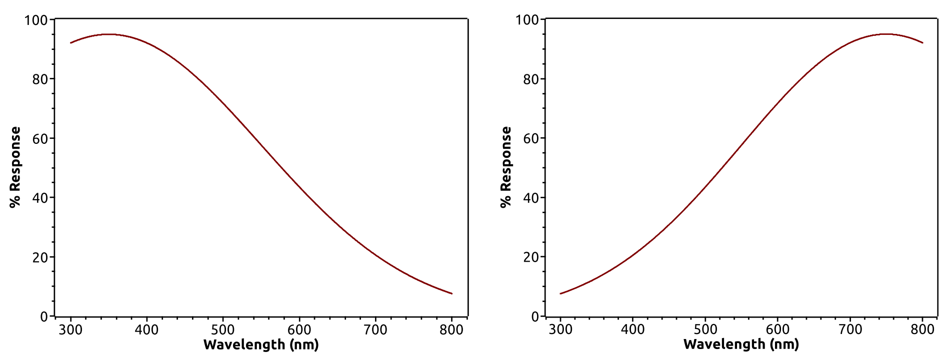

Figure 8.16. Response curves for two different detectors. The detector on the left is most sensitive to short-wavelength light, while the detector on the right is more sensitive to long-wavelength light. The graphs only show the range of wavelengths between 300–800 nm, but that does not mean the detectors are unresponsive to wavelengths outside this range.

light entering the camera *and* the response curve. Thus either can greatly affect how the detector 'sees' the world.

8.6 Color and integration

A perfect system for recording a color picture would preserve a full detected spectrum of the light *at each point* in the picture. But this is not at all practical. Instead one typically compresses the information present in a spectrum into a much simpler form that (hopefully) still captures enough information so as to allow one to reproduce color, at least as perceived by humans.

We discuss more fully in chapters 9 and 10 some of the practical methods for encoding color information. But the basic idea is that the full spectrum of wavelengths is replaced by only three numbers, each of which is a single measure of the overall brightness of the spectrum *over a particular range of wavelengths only*.

Thus, instead of measuring the brightness of the light at each wavelength between 400 nm and 700 nm, we have only three measurements. For example, we could have one number that represents only the *total* light between 400–500 nm, while another number represents the total light with wavelengths between 500–600 nm and the third between 600–700 nm.

This trio of numbers contains much less information than the original spectrum. And this means many different spectra could end up with the same set of three measurements. But if done strategically, one can record enough information about the spectrum to at least represent our *perception* of color with reasonable accuracy. We shall see in chapter 9 that this procedure has much in common with how the human eye records color information.

But just as importantly, any *black and white* detector of light does essentially the same thing. It is just that the black and white detector, at each position in the image, turns the full spectrum of light into only *one* number instead of three.

And so any given detector has a response curve, and the spectrum of light entering the camera interacts with this response curve to produce a detected spectrum. But although the interaction occurs at every wavelength in the spectrum, the detected spectrum itself is not recorded. Most detectors cannot tell, in the end, one detected wavelength from another; they are lumped together into one number if it is a black and white detector, or three numbers if it is a color detector.

We can, in fact, describe a color detector simply as three separate black and white detectors, each of which has a *different* response curve. We will describe how a color detector accomplishes this feat later, in Volume 3 of *The Physics and Art of Photography*. But whether it is a black and white detector or one of the three detectors in a color detector, the same thing happens—all of the light from the detected-light spectrum is added up to form a total measurement of the brightness of the light at any given position in the image.

The mathematical process by which a detected spectrum is compressed into one number is called *integration*; it works like this. We multiply, at every wavelength, the reflected spectrum by the response curve, as described in section 8.5, to determine the detected spectrum. We then take the detected spectrum and *add up all of the area*

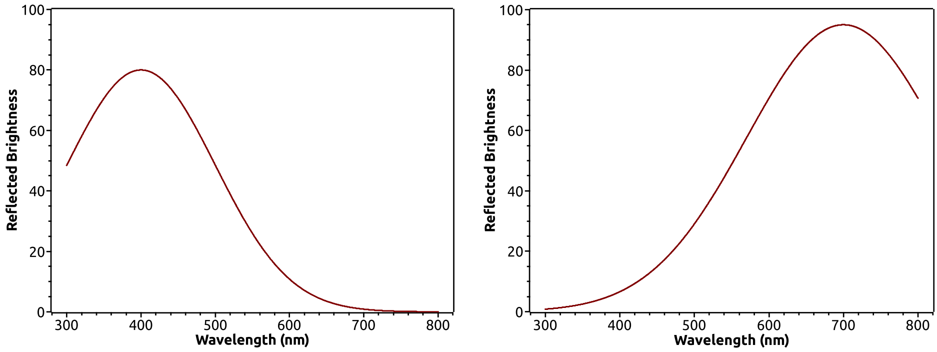

Figure 8.17. Two hypothetical reflected-light spectra entering a camera. When the light of all wavelengths is added up, the light source on the right is about 43% brighter than the source on the left. But even more importantly, the energy of the two sources is distributed differently over the range of visible wavelengths.

under the curve of the graph. The greater this area, the more total light has been measured by the detector.

Let us do an example, first with the simplest case—a black and white detector. Let us take two very different reflected spectra, as seen in figure 8.17. The spectrum on the left clearly would appear more bluish, while that on the right would appear more reddish in color. But notice that although the two spectra are obviously different in shape, we can also see that the curve on the right 'looks a little bigger' than the curve on the left. What we really mean by this is that, when all wavelengths are added together, the spectrum on the right emits a bit more light than the spectrum on the left. This quantity, the total brightness for all wavelengths added together, is represented graphically by the *area under the curve of the spectrum*. In this particular case, the spectrum on the right has 43% more area under its curve than does the spectrum on the left.

And so how will our detector respond to these two spectra? To determine this, we must consider the response curve of the detector. Let us say that we are using the detector with the blue-sensitive response curve shown on the left in figure 8.16. To determine the spectrum of the detected light, we multiply the response curve by the spectrum of the incoming light. The resulting detected spectra can be seen, for both examples, in figure 8.18.

But the detector does not actually record this detected spectrum. Instead, it simply adds all of the light up, equivalent to the area under the curve in each side of figure 8.18. Clearly, there is more area under the curve on the left, and so this detector would record a larger value for the blue light than for the red light. And so this detector would 'see' the blue light as brighter than the red light, even though 43% more light from the red source entered the camera.

In practice, the detector performs this integration process automatically; in fact, it has no choice in the matter. Although it responds differently to different wavelengths, the typical photographic detector has no way to sort the wavelengths into a spectrum. It simply adds them.

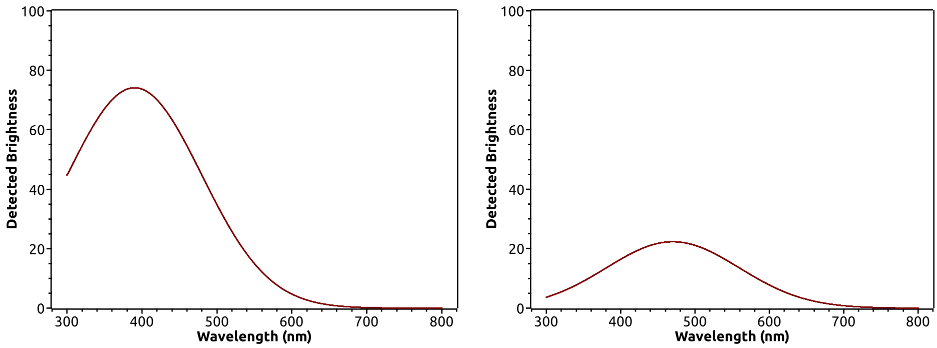

Figure 8.18. Left: the detected spectrum for blue light, with a blue-sensitive detector. Right: the detected spectrum for red light, with the same blue-sensitive detector. When the light of all wavelengths is added up, the blue-sensitive detector recorded the blue light as 71% brighter than the red light, even though the red light source was actually the brighter of the two.

If we know the detected spectrum, as in the left side of figure 8.18 for example, we can use a computer to numerically add up the area under the curve. Or we can do a crude approximation graphically, by superimposing a grid on the graph, as in figure 8.19, and counting up the squares. Clearly, we would get a different answer if we counted the squares of the same size arranged instead under the graph on the right side of figure 8.18.

Let us now consider the same two spectra of incoming light, but detected instead by the red-sensitive detector, the response curve of which is on the right side of figure 8.16. The resulting detected spectra are shown in figure 8.20. Clearly, this different detector would 'see' the red light as brighter than the blue light. And so we see that even if the detector ultimately measures only the overall brightness of the light, with no color information, color is still crucial.

8.6.1 Color detectors

We can easily extend these ideas to color detectors. A color detector is simply a detector that simultaneously records the picture with three different response curves, thus recording three separate measurements.

Our examples in the previous section showed that a red-sensitive detector records red light as brighter than an equal-intensity blue light. And we saw that a blue-sensitive detector does the opposite. And so it is clear that if we measure the light with *both* detectors, the result tells us something about the overall color of the light, even though we only have two numbers. If the red measurement is much greater than the blue measurement, then we are probably looking at red light. If the opposite is true, well then maybe we are looking at blue light.

If both give the same measurement, well maybe it is green light (with a peak wavelength in between red and blue). Or maybe it is white light, with all wavelengths equal. In this case, we wouldn't be able to tell whether it is one or the other. Our two detectors alone do allow us to make some educated guesses about the spectrum of

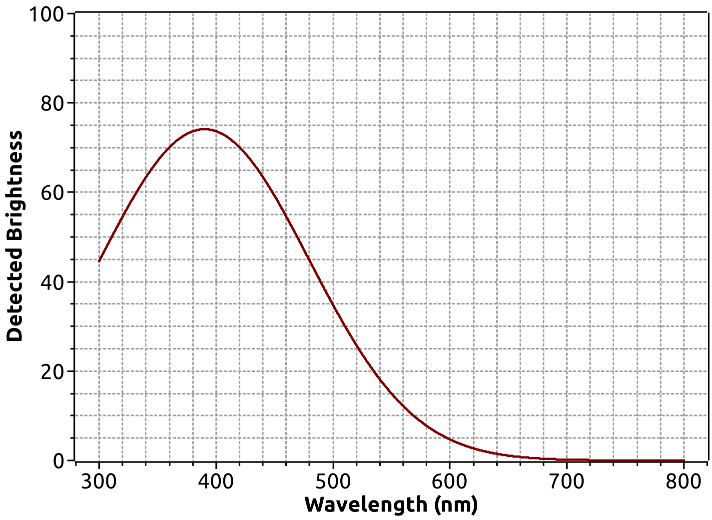

Figure 8.19. *Integration* adds up the area under the curve of the detected-light spectrum; it turns a full detected-light spectrum into a single number. If one superimposes a grid on the graph, one can then simply count up the squares (using fractions for partial squares) under the curve. It is clear that fewer squares of the same size would fit under the curve on the right side of figure 8.18. For most real light detectors, this integration process is performed automatically as part of the physical process of detecting light.

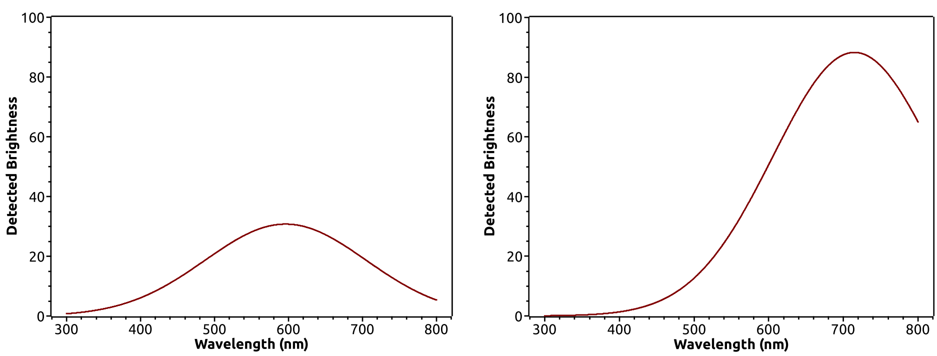

Figure 8.20. Left: the detected-light spectrum for blue light, with a red-sensitive detector. Right: the detected-light spectrum for red light, with the same red-sensitive detector. The detector only records the area under the curve of these spectra. Thus, this detector records a larger measurement (nearly four times greater) for the red light than for the blue light.

the light entering the camera, but we can do much better by adding a third detector, with a response curve in between the other two.

And so most color detectors use some clever scheme for combining *three* separate detectors, each with its own response curve, onto the same detector surface. Usually the response curves are somewhat like those shown in figure 8.21, with peaks in the red, green and blue parts of the spectrum. The light interacts with all three simultaneously and produces three separate measurements. The combination of these three measurement tells us both how bright the light is and also enough crude

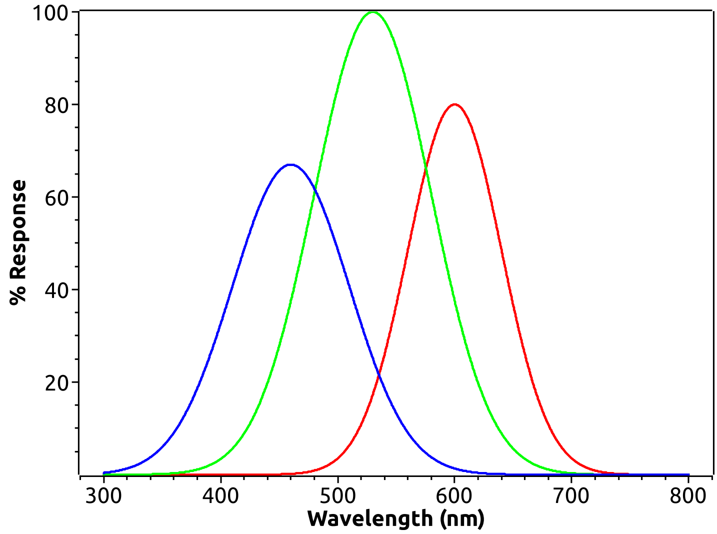

Figure 8.21. Idealized response curves for the separate red, green and blue-sensitive detectors in a typical digital camera. In a real digital camera the response curves have somewhat more complex shapes than what is shown here.

information about the spectrum to, hopefully, reproduce a realistic color perception in a final print or digital display.

From those three numbers alone, we cannot 'go backwards' and figure out what the spectrum was. With additional information we may be able to make an educated guess, but that is all it would be. We will see in chapter 10, however, that our eye/brain combination does something very similar. And thus if we choose our system similar to the way the eye/brain does it, then this seemingly-crude scheme can do a pretty good job at reproducing our impressions of color.

8.7 The relation of color to black and white photography

A black-and-white detector is simply a single detector that has only one response curve. But it should be clear that all of the issues discussed in the previous sections still apply. In the final picture, what is dark and what is light depends on a combination of the spectrum of the light, the reflection curves of the objects the light shines upon, and the response curve of the camera's detector.

This becomes particularly obvious if we take pictures of the same subjects with detectors that have very *different* response curves. In figure 8.22 I have taken photographs of the same scene with the following detectors:

1. An ordinary color digital camera, designed to show the scene much as it would appear to the eye.
2. An ephemeral-process photograph that is sensitive from the near-ultraviolet through the blue part of the visible spectrum. So the response curve extends from roughly 325–450 nm.

Figure 8.22. The scene on the left photographed with three different detectors, sensitive over different ranges of wavelength. Second from left: ephemeral process photograph (325–450 nm). Third from left: ordinary digital with yellow filter (450–700 nm). Right: infrared-converted digital camera (800–1000 nm). Notice that the green foliage reflects more light at very long wavelengths, but the blue sky is brighter at short wavelengths.

3. An ordinary color digital camera, but set to a black and white mode and using a yellow filter that blocks very short wavelengths. This basically adds together the responses of the separate R, G and B detectors. And so the result is a response curve that spans the visible spectrum but excludes deep blue and violet, giving a wavelength range from roughly 450–700 nm, and peaks roughly in its middle.
4. The same type of digital camera, but modified so that it also responds to near-infrared wavelengths from 700–1000 nm. A filter was placed over the camera lens to absorb light in the visible part of the spectrum. The overall effect, then, is of a camera that only responds from 800–1000 nm.

The most obvious effect is that the green foliage, made so by the pigment chlorophyll, appears very dark at short wavelengths, and very bright at long wavelengths. The daytime blue sky, on the other hand, does the opposite. Chlorophyll absorbs light in both the ultraviolet–blue part of the spectrum and also the red part of the spectrum (see section 8.4.1). As a consequence, it reflects well at the other wavelengths—the blue–green and infrared parts of the spectrum. Thus we see the foliage appears brightest with the infrared detector and darkest with the ultraviolet–blue detector. The blue sky does the opposite because the color comes from Rayleigh scattering (see section 8.4.2), and so it is brightest at the shortest wavelengths and darkest at the longest wavelengths.

Chapter 9

The physiological basis of color

The light-sensitive part of the human eye is called the retina, and it is lined with many thousands of individual microscopic light-sensitive cells. It is known that there are two basic types of these cells, named after their shapes when seen through a microscope. The *rods* are, apparently, only sensitive to the intensity of light in general, and do not seem to take a direct role in the perception of color. They are, however, necessary for perceiving very dim light, and this is why it is difficult to sense color in extremely dim light.

The *cones*, on the other hand, are connected to color perception. It is not that each cone cell senses color directly. Rather, some of the cones are most sensitive to a range of mostly longer-wavelength light, some to a range of mostly short-wavelength light, and some with a wavelength response in between these two. A given part of the retina (and thus a given part of an image) makes use of many of these cells, which are especially close-packed in the central part of the visual field. These rod cells are denoted S, M and L, for their short, middle and long wavelength responses. A graph of their approximate response curves can be seen in figure 9.1.

And so, when the eye's lens focuses an image onto the retina, any given spot excites many cone cells, some of which are S, some of which are M and some of which are L. Thinking of the three types of cells as individual detectors with their own response curves, it is not difficult to see how these responses carry information about the color of the light focused on the retina. As an example, the S cone is not affected at all by light of 600 nm wavelength, but the M and especially the L cones are. On the other hand, nearly the opposite is true for light of 450 nm.

One could take any spectrum of light, and use the procedure outlined in chapter 8, section 8.6 to determine the individual responses of the S, M and L cones. Try a different spectrum, and one is likely to get a different answer. And so the idea is that a different *combination* of S, M and L results in a different color perception.

Thus, the retina has a physiology that seems similar in many ways to the example of the color detector of a digital camera described in chapter 8. There too, three

doi:10.1088/2053-2571/aae504ch9

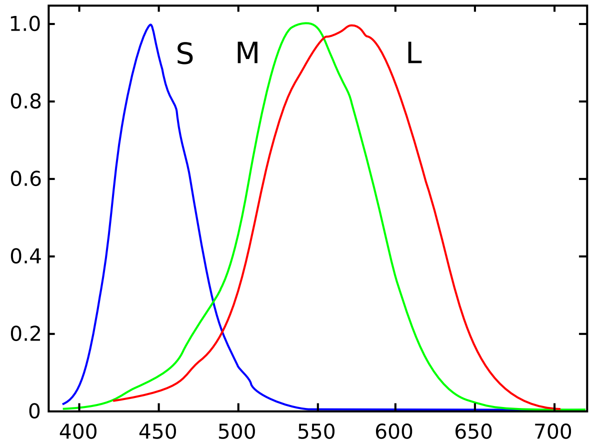

Figure 9.1. The response curves for the three types of cone cells, labeled with the S (for short), M (for middle), and L (for long). Note the similarity to figure 8.21 (graphic by Vanessaezekowitz at en.wikipedia CC BY-SA 3.0).

different types of detectors, each with its own broad response curve centered at different parts of the visible spectrum, are all exposed to the light simultaneously. And it is some combination of those three different images that produces color.

And so it is tempting to assume that human color perception is essentially the same as the way a digital camera or color film records a color picture. The two certainly do have some things in common, and this is the basis of the three-color model we describe in section 9.1. But the similarity ends there, and the differences are profound, for most of color perception is in the brain, not the eye. And so we will have much else to say.

And this brings up an important point: what is the distinction between the 'physiological' aspect of color and the 'psychological' aspect of color? The answer is that it depends a lot upon whom one asks, since the neuro-physiology of the brain is clearly related to psychology—so much so that there is also the study of 'neuro-psychology.' In *The Physics and Art of Photography* I will not belabor this point, and it is unlikely that I would argue strongly with any reader who takes issue with how I have arranged this material to distinguish between things 'physiological' and 'psychological.'

9.1 The three-color model of color perception

Figure 9.2 shows a photograph of a portion of the LCD screen of my ordinary computer monitor, and a magnified close up of one tiny piece. Looking closely, it is clear that the screen emits only three colors: red, blue and green. The screen has separate light emitters for these three colors, all arranged in a grid-like pattern. Notice, for example, that the yellow star has the red and green spots lit up, while the blue ones are dark, and the white areas have all three colors lit evenly. At every point on the screen these little red, green and blue dots combine by lighting up in different

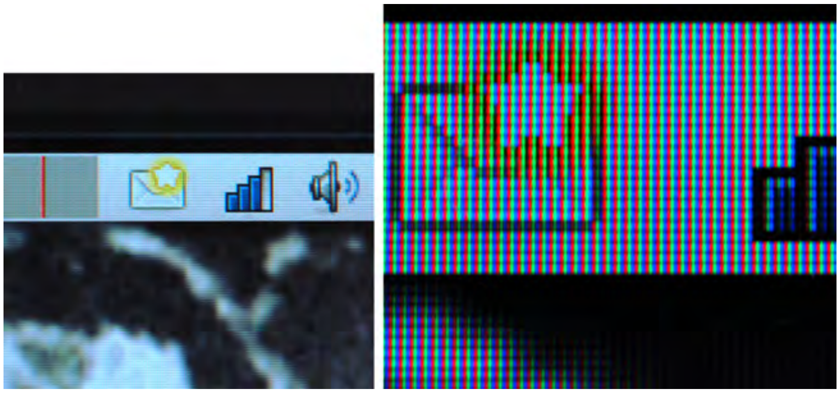

Figure 9.2. A picture of part of an LCD computer screen (left) and a close-up of the same part of the screen (right). All of the colors are synthesized by different combinations of red, green and blue alone. As an example, notice that the yellow star is made of red and green, with no blue.

combinations of brightness. And yet, it produces what I perceive to be a wide (almost full) range of colors, and it is able to simulate subtle color differences.

Combining only red (R), green (G) and blue (B) light in various combinations to make color is called *color synthesis*, and the fact that it works is part of the basis of the *three-color model of color perception*. Another basis is the seemingly-obvious connection to the three-response L, M, S cone cell physiology of the eye just described. In fact it is tempting to move immediately to a simple hypothesis:

> The separate R, G and B lights emitted by the computer monitor separately stimulate the L, M and S cones in the retina. The R parts of the monitor stimulate the L cones, while the G parts of the monitor stimulate the M cones and the B parts of the monitor stimulate the S cones. And so RGB light directly simulates LMS cones, and the brain turns those three separate stimuli into color. Thus the brain has some kind of mental 'chart;' for any given combination of LMS responses, the brain simply looks on the chart and gives us one of the many hundreds of thousands of possible color perceptions.

Regarding the computer monitor itself, there *is* such a chart that relates the separate brightnesses of R, G and B to different colors. It is called a *color map*, and we will consider some of its properties in what follows. But these simple assumptions about the brain are very much off-the-mark, as has been demonstrated time and again by neuroscientists, topics we touch upon in chapter 9. $RGB \neq LMS$.

Furthermore, there is another problem with the above hypothesis: it is physically impossible to make three different colors of light, each of which separately stimulates *only one* of the three types of cone cells. This is obvious as soon as one looks carefully at the actual cone response curves in figure 9.1—all three of the responses overlap each other significantly. This is especially true for M and L, which are more similar than they are different. Thus *any* source of light will trigger responses of more than one type of cone (except for either a single extremely-short or extremely-long wavelength).

The fact remains, however, that RGB color synthesis works pretty well at consistently reproducing a wide range of color perceptions, and so it is the foundation for the technology of color photography, color display and color printing. Exactly *how it works in the brain* is a different question. We take up that question further in chapter 10, but the short answer is—the answer is unknown, at least in detail. That said, let us proceed with some details of RGB color synthesis, as it has great practical value.

The spectra of the lights typically used for successful RGB color synthesis bears an only superficial resemblance to the LMS cone response curves. The RGB emitted colors usually have a somewhat narrower range of wavelengths emitted than the separate LMS responses. And they are usually spaced more evenly across the visible spectrum, with much less overlap between G and R than between M and L.

We can denote the separate brightnesses of R, G and B for a particular color with a trio of numbers. Most commonly, each is assigned a number between 0 and 255, with 0 meaning no light at all and 255 the maximum amount of light. And so RGB = 0–255–0, for example, would be pure green, while RGB = 50–125–255 would be a little bit of red light mixed with much more green and even more blue. A *color map* of these possible combinations is a graphical representation of all of the possibilities, and what RGB numbers produce them. R, G and B give us three dimensions, and so only some of it can be depicted at once in any two-dimensional graph.

If we encode our RGB colors with 256 possible values separately for R, G and B, this gives us $256^3 = 1.6788 \times 10^7$ possibilities. This is called *24 bit color*, because a range of 256 possibilities can be represented with 8 binary computer bits. There are three of these 8 bit numbers, and so there are 24 binary bits needed to code this many possible colors.

9.2 Additive and subtractive colors

We would like to have a scheme that not only accurately describes color perception in an intuitive yet reproducible way, but that also allows us to predict what happens when, say different colors are combined. We have already described a way to represent colors in terms of combinations of red, blue and green light (RGB), and we describe yet another, more intuitive scheme in chapter 10, section 10.5.

But we must make an important distinction between two basic ways of combining colors. If we are talking about combing different sources of light to make an overall color, that is one thing; we are *adding* light of different wavelengths in order to make a particular spectrum, which gives a particular color sensation. But if we are talking about light reflected off of surfaces, it is really about what wavelengths of light have been *subtracted* by absorption. Thus mixing light is *additive*, while mixing pigments (such as paint) is *subtractive*. Not surprisingly, a particular scheme for describing one will give incorrect results for the other.

The computer screen in figure 9.2 is an example of addition of light. Mixing paint, on the other hand, is an example of subtracting light. When one mixes red and green

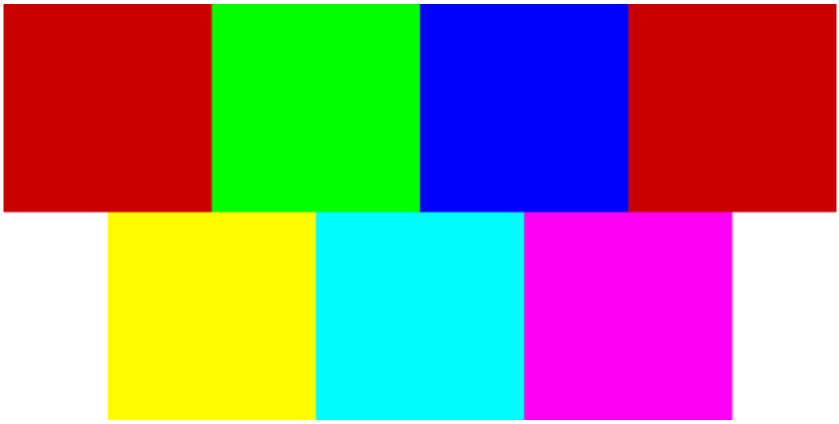

Figure 9.3. The additive primaries—red, green and blue—are in the top row. The additive secondaries—yellow, cyan and magenta—are in the bottom row. Adding any two colors of light in the top row produces the intermediate additive secondary color in the bottom row.

light, one ends up with more light than either the red or green light alone, and yellow results. But if one instead mixes red and green *paint*, the red paint pigment *removes* everything *except for* red, while the green removes everything except for green. Thus in this case there is really not much left, and a very dark and muddy color of paint results (the exact color depends on the precise absorption curves of the two pigments).

Since we can synthesize colors by mixing only red, green and blue *light*, we refer to red, blue and green as the *additive primaries*. When two of the additive primaries are combined in equal amounts, we produce cyan, magenta and yellow; these are known as the *additive secondaries*. See figures 9.3 and 9.4.

The rules for subtractive mixing of colors are quite different. Clearly we cannot use red, green and blue as a set of primary colors for mixing paint. As we have seen, mixing red and green paint takes away pretty much all color. But if we think about it a bit, there are three colors that will work as *subtractive primaries*. That is, there *is* a set of three color pigments that can be mixed in different combinations in order to produce pretty much whatever hue we desire. It is not too difficult to see that the *additive* secondaries can serve as our set of *subtractive* primaries. Thus, although red, blue and green will not work as primaries for mixing paint, cyan, magenta and yellow will. See figure 9.5.

The best way to illustrate this is by example. Let us mix yellow paint and cyan paint. If yellow paint reflects yellow, it means it reflects both red and green light, while absorbing blue. For after all, yellow light is simply a combination of red and green light. And so for light, yellow means 'add red and green,' while for paint, yellow means 'subtract blue.' Cyan light is a combination of blue and green light. And so cyan paint must reflect blue and green, while absorbing red. And so cyan *light* means 'add blue and green' while cyan *pigment* means 'subtract red.'

Thus if we mix yellow and cyan paint, the yellow will absorb blue, while the cyan absorbs red. Now shine white light on our mix of cyan and yellow paint—white light

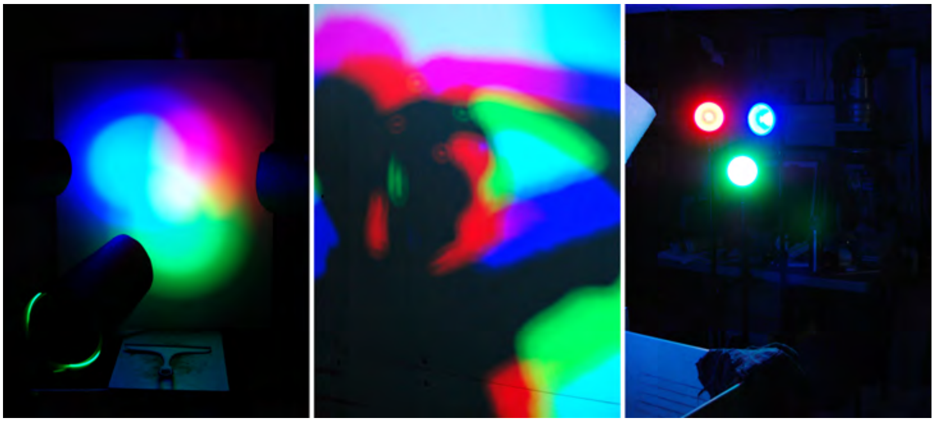

Figure 9.4. Left: Red, green and blue light adds to make white, but in pairs they make the additive secondaries cyan, magenta and yellow. Center: I photographed my shadow on a wall lit by red, green and blue flood lights (seen on the right). Where my shadow blocked all three lights it is dark; where I blocked, none it is white. The additive primaries can be seen in places where I blocked two of the lights but not the third. The additive secondaries can be seen where I blocked one light, leaving the other two to combine.

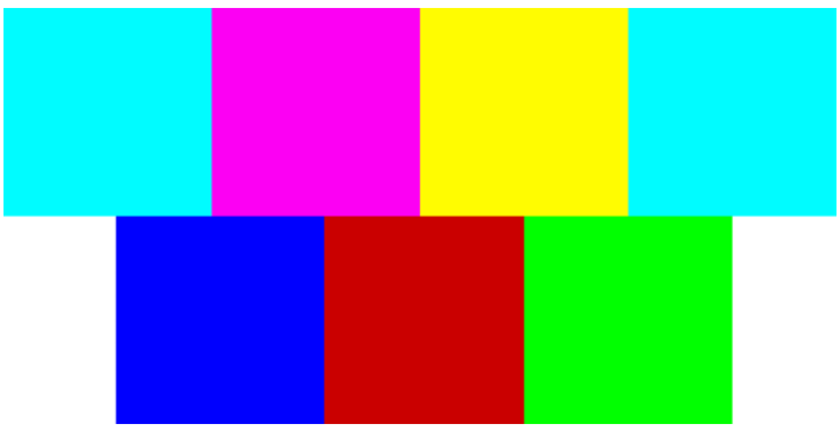

Figure 9.5. The subtractive primaries—cyan, magenta and yellow—are in the top row. The subtractive secondaries—blue, red and green—are in the bottom row. Combining any two pigment colors in the top row produces the intermediate subtractive secondary color in the bottom row.

that combines red, blue and green light. The yellow pigment will absorb the blue, while the cyan pigment will absorb the red, *leaving only green to be reflected*. Thus yellow and cyan pigments mix to make green pigment. It is not difficult to see that magenta and yellow pigment combined will reflect red, since the yellow absorbs blue while the magenta absorbs green, and thus only red remains. Similarly, cyan and magenta combined will absorb red and green and thus reflect blue.

And so we have an interesting reciprocity. The additive primaries are red, blue and green light. When combined in pairs, they produce the additive secondaries cyan, magenta and yellow light. A combination of all three produces white light.

The subtractive primaries are cyan, magenta and yellow. When combined in pairs, they produce the subtractive secondaries red, blue and green. Combining all three of the subtractive primaries takes away everything, and produces essentially black. Of course, we often use pigments (in oil paints, for example) that are much more complex than the fundamental subtractive primaries. And there is the added complication of to what degree a given pigment allows light to pass through it and get to other pigments at deeper layers. So the mixing of paints, for example, is far more complex than what I have laid out here.

The additive view of color will be more useful to us for most of our discussion of color in photography. Subtractive colors, on the other hand, are more important for consideration of printing. A good example is the ordinary inkjet print which synthesizes colors by printing many pigmented ink dots on white paper. If one looks at an inkjet print under a microscope, it is clearly made of cyan, magenta and yellow dots (some more-expensive printers use additional colors as well). There are also often black dots, and these allow one to lower the *value* of the color without changing its *hue*. The *saturation* of a color can be decreased by simply leaving more white space on the paper, between the dots. In chapter 10, section 10.5 we discussion the concepts of hue, value and saturation in some detail, but in terms of additive colors.

9.3 RGB color arithmetic

We can easily explore the relations between additive and subtractive primaries and secondaries with a simple arithmetic. We use a plus sign (+) to represent the addition of light, and a minus sign (−) to represent the subtraction of light by a pigment. We then attach these mathematical signs to the letters R, G, B, C, M, Y, W, K[1], to represent red, green, blue, cyan, magenta, yellow, white and black. And so, for example, we represent the addition of red light as $+R$, and the subtraction of green light as $-G$.

We then recognize the following relations for the addition of light:

$$R + G = Y \tag{9.1}$$

$$G + B = C \tag{9.2}$$

$$R + B = M \tag{9.3}$$

$$R + B + G = W \tag{9.4}$$

This is our statement that the additive primaries, red, green and blue, add together in pairs to make the additive secondaries, cyan, magenta and yellow.

Given these relations, we can see with simple algebra that the following must be true for the subtraction of light:

[1] It is traditional to use 'K' to represent black in this context, presumably because 'B' is already taken by blue.

$$W - R = (R + G + B) - R \tag{9.5}$$

$$W - R = G + B \tag{9.6}$$

$$W - R = C \tag{9.7}$$

And so we see that subtractive cyan *pigment* is given by $C = -R$. When we say a pigment is 'cyan,' we really meant that *white* light reflecting off of it will appear cyan. And this means it is a pigment that subtracts red light and only red light. Non-white light reflecting off cyan pigment may *not* appear cyan in color. But the pigment is still the same; it is a pigment that subtracts red light while reflecting blue and green. We can thus recognize the following relations for the subtractive primary *pigments*:

$$C = -R \tag{9.8}$$

$$M = -G \tag{9.9}$$

$$Y = -B \tag{9.10}$$

But then, what do we mean by a red pigment? It must be the following:

$$R = W - G - B \tag{9.11}$$

$$R = W + M + Y \tag{9.12}$$

Thus a red pigment can be made by combining magenta and yellow pigments. The magenta absorbs green, while the yellow absorbs blue, leaving only red to reflect when white light shines on it. It then follows that we must have:

$$W + M + Y = R + G + B - G - B = R \tag{9.13}$$

$$W + Y + C = R + G + B - B - R = G \tag{9.14}$$

$$W + C + M = R + G + B - R - G = B \tag{9.15}$$

$$W + C + M + Y = R + G + B - R - G - B = 0 = K \tag{9.16}$$

This is our statement that the subtractive primaries, cyan, magenta and yellow, combine in pairs to make our additive primaries, red, green and blue, if we assume the pigments are illuminated by white light. Combining all three subtracts everything, leaving an arithmetic zero, or black.

We can also then see the following results for the secondary *pigments*:

$$R = -G - B \tag{9.17}$$

$$G = -R - B \tag{9.18}$$

$$B = -R - G \tag{9.19}$$

To use these relations, we have to take care whether we are talking about light or pigments. Red light is $+R$, while red pigment means $-G - B$. Cyan light is $+G + B$, while cyan pigment is −R. But keeping those distinctions in mind, we can use these relations to answer specific questions about light (even colored light) shining on subtractive pigments. For example, what happens when white light ($R + G + B$) shines on a mixture of the pigments cyan ($-R$) and red ($-G - B$)?

$$(R + G + B) - R + (-G - B) = R + G + B - R - G - B \tag{9.20}$$

$$= 0 = K \tag{9.21}$$

Thus, the answer is no light would be reflected, and it would appear black.

Here is a more complicated example. What happens when cyan light ($+G + B$) and blue light ($+B$) shines on a mixture of green pigment ($-R$ − B) and yellow pigment ($-B$)? The answer is easy. Simply add up all the letters keeping track of the signs, and see what is left:

$$(+G + B) + B + (-R - B) - B = G + B + B - R - B - B \tag{9.22}$$

$$= G - R \tag{9.23}$$

So clearly green light would be reflected, but what would be the meaning of the remaining '$-R$?' This simply means that although the pigment combination absorbs red light, no red light was shining on it in the first place. And so that fact would make no difference, and the answer is that if one shines a mixture of cyan and blue light on a mixture of green and yellow pigment, it would appear green.

We must note another feature of the previous example. Notice that the $+B$s were canceled by an equal number of $-Bs$. What if instead we had had more $+B$s than $-Bs$? This is a place where our color arithmetic needs a different rule than ordinary numerical arithmetic. When we say '−B' we are implying that the pigment absorbs *all* blue. Thus it should not matter how much blue light is shining on such a pigment; no blue light should reflect. And so we must modify our color arithmetic so that a single (−) of a given color cancels *all* (+) values of that same color. And so, for example, in our color arithmetic: $B + B - B = 0$, not B. Table 9.1 gives a summary of the different color arithmetic translations for both light and pigments. To determine what color of light reflects when a source (or sources) of colored light shines on colored pigments, simply follow these steps:

1. Use the table to translate the light source colors and the pigments. For example, translate magenta light to '$+ R + G$' and translate red pigment to '− G–B.'
2. Add up the translations, but with the rule that a single (−) of a given color cancels *all* (+) values of that color. And so add the translations as if, for example, $B + B -$ B $= 0$, rather than B.

   And so if we were to shine blue and magenta light on red pigment, we would add the following: $B + (R + B) + (-$ G–$B) = B + R + B -$ G–$B = R -$ G.

Table 9.1. The relations between the primary and secondaries for both adding light and (left) and for subtracting light with pigments (right). To find the color of the light that is reflected in a given case, simply add the combination of light from the left side to the combination of pigments from the right side, and see what is left over.

Name	Light	Pigment
Red	R ⟺ + R	R ⟺ −G−B
Green	G ⟺ + G	G ⟺ −R−B
Blue	B ⟺ + B	B ⟺ −R−G
Cyan	C ⟺ + G + B	C ⟺ −R
Magenta	M ⟺ + R + B	M ⟺ −G
Yellow	Y ⟺ + R + G	Y ⟺ −B
White	W ⟺ + R + G + B	W ⟺ −0
Black	K ⟺ + 0	K ⟺ −R−G−B

Notice that the single '$-B$' canceled out both $+B$s. This is because if the pigment absorbs all blue, then it matters not how much blue light shines on it; no blue will reflect.

3. Interpret the result by using the 'light' side of the table to translate backwards; it is, after all, the reflected *light* that we are trying to determine. *Ignore negative numbers*, since it simply means that that color of light *would* have been absorbed, but no light of that color was shining on the pigment. Also, interpret (for example) '$2R + B$' as $R + (R + B) = R + M$, or 'magenta with extra red added.'

   And so, for example, shining a combination of green and cyan light onto cyan pigment would give the result $G + C - R$. And this would translate as reflected light with the color 'cyan with extra green.'

I repeat here the same caveats expressed in the previous section. This three-color arithmetic of light and pigments assumes *idealized* light sources and *idealized* pigments. Real light sources and real pigments are far more complex, and for accurate results must be analyzed with the more complex procedure laid out in chapter 8. One must compare the precise spectrum of the light source to the reflection curve of the pigment, in order to determine the actual spectrum of the reflected light. The three-color model can then be applied to that spectrum of reflected light, in order to determine an approximate color perception. So think of this color arithmetic as only a useful starting point for a far more complex reality.

Chapter 10

The psychological basis of color

10.1 The opponent-process model of color perception

We can easily talk of colors such as 'blue-green' or 'yellow-orange,' but what to do with a color name such as 'orange-blue' or 'reddish-turquoise?' In recognition of this basic fact, it has long been proposed that our color perception is based in part upon fundamental opposites, or *opponents*. The most likely candidates are a contrast between something like red and cyan, on the one hand, and something like yellow and blue on the other. Furthermore, although we have only three different cone types—S, L and M—we seem to have four psychologically fundamental hues: red, green, blue and *yellow*.

And so it seems that a particular combination of signals from the cone cells is not translated directly into some simple map of color perception stored in the brain. Instead the signals from the three types of cone cells are *pre-processed*. That is, they are combined with each other in various ways, to form the foundation of our color perception.

A simplified version is as follows (Conway 2003, Conway *et al* 2010). The direct responses of the three kinds of cone cells—S, L and M—are pre-processed in the retina by cells called *cone-opponent retinal ganglion cells*. There are four types, and they provide the basic computational *comparisons* of the direct responses of the cone cells:

1. L-on/M-off.
2. L-off/M-on.
3. S-on/(L+M)-off.
4. (L+M)-on/S-off.

And so we can think of two pairs of ganglion cells. One pair responds to L versus M, while the other pair responds to S versus (L+M). This would seem to have some direct connection to the red–green and blue–yellow opponents that seem to be part of our color psychology. The first two represent a comparison between the signals of the L and M cones, which have responses most similar to what we call red and green. The last two represent a comparison between the S cone and the combination of L and M together. This would seem to be related to the opponents of blue and yellow, since S has a response close to blue, and the sum of the responses of the middle (M) and long (L) responses are similar to the sum of red and green, which makes yellow. There is also a similar neural opponent process for black versus white.

Clearly, there is something at the fundamental neurological level that acts in a way analogous to our color psychology that sees blue and yellow on the one hand, and red and green on the other, as opposites, rather than ends of a continuum, and that also sees yellow as a fundamental hue on a par with red, green and blue. But such a simple and direct relationship has not been born out by the research, and much mystery still surrounds the neural origin of our color psychology (Conway *et al* 2010, Schmidt *et al* 2014).

RGB synthesis of color is a convenient way to design a computer monitor or a digital color light detector. And it makes sense to encode the color information this way in a digital image. But it has little relation to our intuition regarding color. In section 10.4, I describe a common numerical way to represent colors that fits better with our intuitive understanding.

10.2 Yellow without yellow

When synthesizing colors with R, G and B alone, there is mostly a one-to-one correspondence between combinations of RGB and the color perception that results. This is not *literally* true—if a variation of RGB values is too minute, then we cannot distinguish it as a different color, especially for colors of low saturation. But it is not as though a combination such as RGB = 10–125–37 on the one hand and RGB = 230–87–145 on the other look the same as each other.

But this general point is not true for the *spectrum* of the light. For there are many possible spectra that, even when they are qualitatively very different from each other, can nonetheless appear as the same color. The perception of yellow is a good example. We see yellow when the spectrum of light occupies a narrow range of wavelengths between the orange and green parts of the spectrum—what we call the yellow part of the spectrum. But we may also see light as essentially the same color *even if it has no light at all of those wavelengths*. It only needs to have the right combination of red and green.

Violet is another case. Monochromatic light with a wavelength of 400 nm appears violet. But a color perception very similar can be synthesized with the right combination of red and blue, even if neither has any light at wavelengths as short as 400 nm. And this is doubly strange when one thinks about it; the red light in this color synthesis is on the *opposite* end of the spectrum from the short-wavelength blue light.

That yellow light can be made from red and green can be explained to some degree by simply looking at the SML response curves. Each responds to a broad range of wavelengths, and so there are many different combinations of wavelengths that, when added up in relation to the SML response curves, will provide the same

Figure 10.1. Our perceptions of single wavelengths—the colors of the spectrum—form a circle in our color psychology, not a line. But our psychological perception of a color between red and blue—magenta—does not appear in the visible spectrum of single wavelengths.

combination of SML triggering. That violet can be synthesized with a combination of red and blue, however, is not at all evident from the SML response curves. And it speaks to one of the key features of our color psychology: we perceive color hues not as a spectrum, with violet on one end and red on the other—but rather as a circle, where blue merges to violet, which merges back to red. Figure 10.1 shows an example of such a *color wheel*, taken from the color dialog used in the popular image processing software, GIMP.

The color wheel of hues in figure 10.1 is not the whole story, as evidenced by the triangle inside of it that represents a two-dimensional range of colors, all of which occupy only *one* place on the color wheel (red). We take up these issues in sections 10.4 and 10.5. At first glance, the color wheel of hues looks like the visible spectrum of monochromatic colors, but with the ends connected to make a circle. There is indeed a resemblance, but with an important exception. There is no single wavelength of light that appears *magenta*, halfway between blue and red on the circle of our color perception.

10.3 Seeing and context

Since we have only three types of cone cells in our retinas, it makes sense that—even if the signals are pre-processed in some as-yet-unknown complex way—any particular combination of stimulation of these three cells always results in the same particular color sensation. Right? That is, we should be able to make some kind of chart such that, for a particular combination of stimulus of the S, M and L cone cells (represented by numbers, for example), the chart would show us the particular color perception that results. Correct?

It turns out that this seemingly-obvious idea is incorrect (Conway *et al* 2010). There is overwhelming evidence, both psychological and neuro-physiological, that the color perception of our human eye/brain combination very much depends also on the *context*. And so there is far more to color perception than simply the stimulation of our cone cells by different wavelengths of light. *The exact same spectrum of light can cause very different color perceptions, depending upon the context*. And this basic insight—perception depends not only on the stimulus but also the context—applies to more than just color.

A rather obvious example, clear from simple introspection, is that we unconsciously ‘correct’ our vision for the amount of light falling on the subject. We see a surface as painted black, rather than white, based on much more than simply how much light reflects off that surface and makes it to the retina of the eye. We also take into account, mostly without conscious thought, how much light is falling on that surface. And a person viewing a scene in the world may have many more cues regarding that light than what is available in the direct image of the particular scene being photographed. And so the act of looking at a picture can be very different from the act of looking at that same scene in the world.

A photographer must take this into account, because the viewer of the picture has only the picture. That is why we often use incident-light metering in order to adjust for the lighting, whatever it might be in the world, so it appears ‘typical’ in the photograph. And if we want the picture to represent a scene illuminated with either very dim or very bright light, *the picture itself must provide the context* necessary for our eye/brain combination to make the right assumptions.

As an example, see the two images in figure 10.2; you may recognize them as close-up views of rolls of duct tape. They both look gray, but the right-hand example is clearly darker. Maybe it is gray duct tape on the left and black duct tape on the right? Now see, in figure 10.3 the original image from which these details were directly cropped, with no additional adjustments of levels. From the context we see a roll of black duct tape on the left, in direct sunlight, with a roll of white duct tape in the shade on the right. We unconsciously use the pattern of light and shadow to make assumptions about the light on the subject, and thus to interpret the meaning of the different parts of the image. Seen out of context, we attach very different meanings to those details.

These rather familiar observations about how we perceive levels of brightness also apply to color. The color we perceive depends not only on the spectrum of the light that arrives at the retina—it also depends critically on unconscious assumptions we make regarding the light that is illuminating the subject. This basic observation about the contextual nature of our color perception reached a worldwide audience in early 2015. A photograph of a dress was posted on social media, and it quickly went

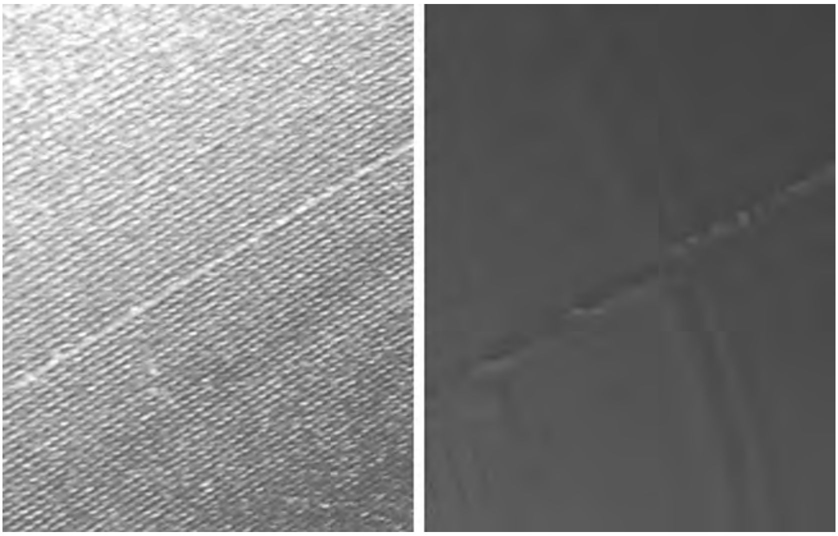

Figure 10.2. Photographs of two rolls of duct tape. Seen out of context they could be pictures of gray duct tape on the left and black duct tape on the right. The two images are cropped, with no adjustments of levels or contrast, from the image in figure 10.3.

Figure 10.3. The two pictures of duct tape from figure 10.2, seen in the context of the original scene. Our perceptions of light and dark are dependent upon the context of the light on the subject.

viral. Some people saw it as a black and blue dress, while others saw the same picture as a white and gold dress.

An illustration simulating the same effect can be seen in figure 10.4. On the left is a black and blue dress while on the right the dress is yellow and white. But when the dress on the left is illuminated by yellowish light the colors are spectrally identical to the those of the dress on the right when it is illuminated with blue light. When the

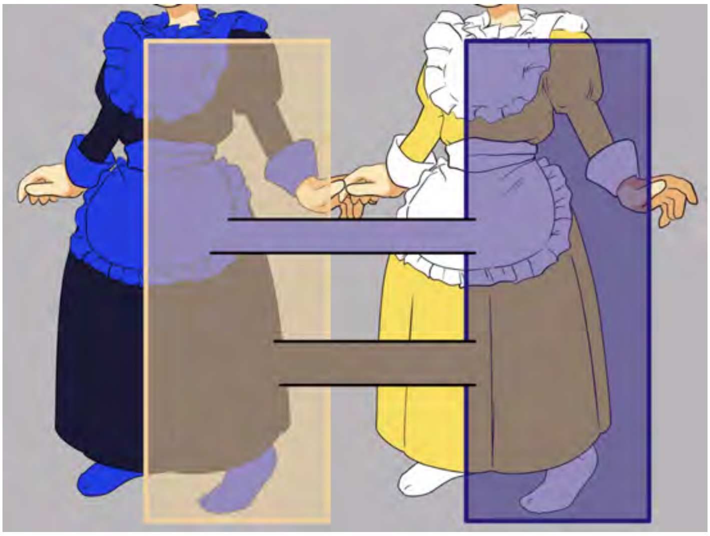

Figure 10.4. The boxed-off parts of the two dresses have identical colors. But in the context of the picture—where it looks as though the dress on the left is illuminated by yellow light while the dress on the right is illuminated by blue light—we see those colors as very different (graphic: figure design by Kasuga jawiki; vectorization by Editor at Large; 'The dress' modification by Jahobr, CC BY-SA 3.0,).

context of the lighting is provided, we don't see it that way—and the boxed-off parts of the dresses look for all the world like very different colors (cover all but the joining strips if you don't believe that the colors are actually identical).

And so we interpret colors in terms of the context of the lighting. If the color of the lighting changes, we still see the colors as the same, even when their literal spectral properties have changed substantially. This basic fact about our color psychology is called *color constancy* (Conway 2003), and it is a very important point for photographers. In particular it greatly affects what we perceive as white or gray, with no particular hue at all. We consider these issues further in chapter 11, section 11.2.1; it goes to the heart of the important task of controlling the *white balance* of a picture.

The famous viral picture of the dress was unique in that the context provided by the picture was ambiguous in just the right way. And so some individuals could see it only as the black and blue dress on the left of figure 10.4, while others could see it only as the gold and white dress on the right.

10.4 HSV and HSL

To understand how we perceive color, it is necessary to understand not only the physical basis of the spectrum of a given source of light, but also how our brains translate that physical sensation into a color perception. The three-color model is probably the simplest way to give a practical *description* of our color perception, a description that works pretty well most of the time. But it is not a very intuitive way to describe color. It is, for example, non-obvious to most that an RGB value of 255–255–0 (full red, full green, no blue) produces yellow.

There are thus other more intuitive ways to convey the same basic information as RGB. One of the most popular is called HSV (sometimes called HSB), which stands for *hue, saturation and value (or brightness when called HSB)*. Like RGB, HSV separates the vast array of possible color perceptions into three categories. But the similarity ends there:

Hue: This is the first thing one would typically ask about a 'color.' It answers that most basic question, 'Is it red, green, blue, orange, violet or what…?'

Saturation: How 'colorful' is it? Is it a pure color, or is it nearly gray? One can *lower the saturation* of a paint by mixing gray paint with it.

Value (or Brightness): How dark or light is it?

For a system such as this to be useful, there must be a way to separate the three issues from each other, and consider them individually. That is, for a given hue and saturation, we should be able to show precisely what we mean by different values, and ditto for the other combinations. Furthermore, since practical color systems often use RGB as their foundation, there should be a straightforward way to translate between the two schemes.

It turns out, rather unfortunately, that there is more than one way to set up such a scheme, when we look closely at the details. And so there are several systems that differ subtly one from another, the most important alternative to HSV being a system called HSL, or hue, saturation and *lightness*.

Although both HSV and HSL have essentially the same definition of hue, they differ in their meanings of saturation and value (or lightness). And this results in some inevitable trade-offs in 'intuitiveness' between HSV and HSL. In my opinion, HSL has a more-intuitive definition of saturation than HSV, but a less-intuitive definition of lightness (called value in HSV). It would seem at first glance that the solution would then be to combine both—use HSL's definition of saturation in combination with HSV's definition of value. But that turns out to be logically impossible, when one looks carefully into the details; the two competing systems represent a real trade-off, not just a style choice.

In what follows I will use the particular definition of HSV assumed by the popular (and free) image-processing program GIMP, and HSL will not be considered further in *The Physics and Art of Photography*. But once HSV is learned in detail, it is straightforward to transfer that knowledge to HSL (or another of the competing color systems) if practical considerations dictate so.

10.5 HSV and RGB

To begin our discussion of HSV, consider figure 10.5, made by using the color dialogue in GIMP. It shows three rectangles. Each rectangle holds one of these three variables constant, while the other two change from top to bottom on the one hand and left to right on the other:

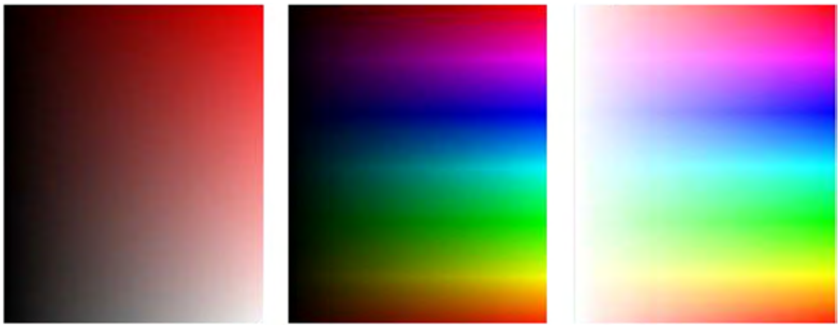

Figure 10.5. **Left**: Hue = 360°, with saturation increasing from bottom to top, and value increasing from left to right. **Center:** Saturation = 100%, with hue increasing from bottom to top, and value increasing from left to right. **Right:** Value = 100%, with hue increasing from bottom to top, and saturation increasing from left to right.

Left: This rectangle is all of the same *hue*, H = 360°.

This is also the same as H = 0°, because hue makes a circle that begins and ends at the same point. It is apparent that the starting and ending point for hue is chosen (by convention) to be red.

Up and down on this rectangle represents different levels of *saturation*. And so the top has a saturation of S = 100%, while the bottom has a saturation of 0%. Notice that the bottom is made of simple shades of gray, with no color; that is the meaning of 0% saturation.

Left and right on this rectangle represents different *values*. And so the left side is a value of 0%, while the right side is a value of 100%. Notice that a value of 0% is black, regardless of the saturation.

Center: This rectangle is all of the same *saturation*, S = 100%.

Up and down on this rectangle represents different *hues*, with 0° at the bottom and 360° at the top. We can see that the hues go around a circle (here stretched out into a line) that begins and ends at the same hue, and that goes approximately (but not quite) through the spectral colors of ROYGBIV in order, and then through magenta and back to red.

Left and right on this rectangle represents different *values*. And so the left side is a value of 0%, while the right side is a value of 100%. Notice that a value of 0% is black, regardless of the hue.

Right: This rectangle is all of the same *value*, V = 100%.

Up and down on this rectangle represents different *hues*, with 0° at the bottom and 360° at the top.

Left and right on this rectangle represents different levels of *saturation*. And so the left side is a saturation of 0%, while the right side is a saturation of 100%. Notice that a saturation of 0% is white, regardless of the hue—but only if (as in this example) V = 100%. A range of *different* V, with S = 0%, appears as different shades of gray (see the bottom of the left-most rectangle).

If we look closely at these examples we can see two interesting facts about HSV that may not be obvious at first:

1. If S = 0% then V matters, but H does not. To say this a different way, if S = 0% then all hues look the same, but different values produce different shades of gray.
2. If V = 0%, then neither S nor H matters. V = 0% is black, whatever the values of S and H.

And so now let us relate these values of HSV to RGB. In figure 10.6 I show a GIMP dialog window that allows one to set values for RGB and HSV, along with an image of the resulting color (labeled 'Current:'). The example on the left shows a selection of pure red, represented by RGB = 255–0–0, representing the maximum amount of red, but with no green or blue. Note that the values for hue, saturation and value are HSV = 0–100–100. This means a hue of 0°, and saturation and value both set to 100%.

Hue is represented in degrees because it refers to position on a circular color wheel. The numbers go around the circle with increasing angle from pure red (0°), through pure green (120°), through pure blue (240°) and back to pure red at 360°. Exactly halfway between each, at 60°, 180°and 300°, represent equal mixes of these primaries. These are the additive secondary colors yellow (red and green with no blue), cyan (blue and green with no red) and magenta (red and blue with no green). The example on the right in figure 10.6 shows hue set for magenta, at H = 300°, with both saturation and value set to 100%. This corresponds to an RGB = 255–0–255, or full red with full blue, but with no green.

With both saturation and value at 100%, different hues represent different combinations of *at most two* of R, G and B. As we start at 0° and work around the color wheel, R goes down while G goes up (with B = 0) until there is no more R, and G is at its maximum. As we continue, B increases while G decreases (and R = 0),

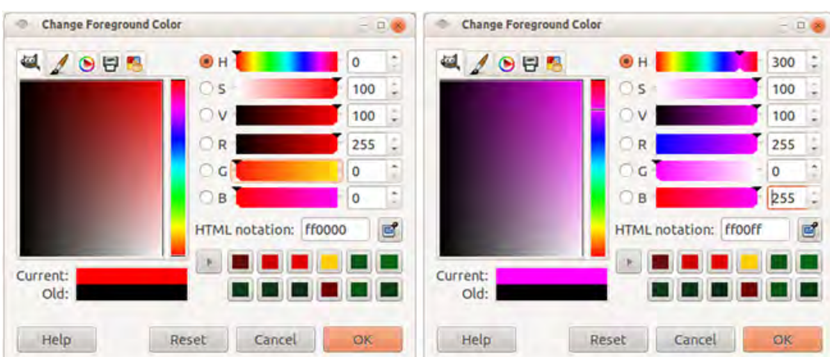

Figure 10.6. Dialog boxes from GIMP showing the effects on RGB of changing hue. The example on the left is pure red (hue = 0°), while that on the right is magenta (hue = 300°). Both have value and saturation of 100%. The chosen color is in the box labeled 'Current.'

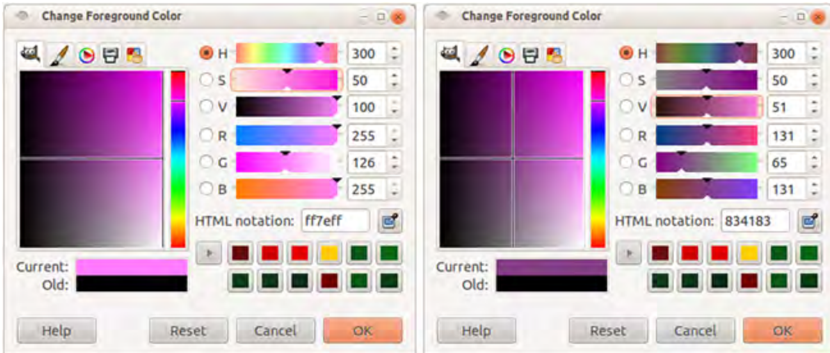

Figure 10.7. Dialog boxes from GIMP showing the effects on pure magenta (RGB = 255–0–255) when saturation is lowered (left) and then value is lowered (right) from 100%. The chosen color is in the box labeled 'Current.'

until there is only B and no G. Finally as we work around back to the beginning, B decreases while R increases (with G = 0), until we are back to only R.

Now let us see what happens when we adjust saturation and value. The left side of figure 10.7 shows the effect of lowering the *saturation* of pure magenta from 100% to 50%. Notice that the red and blue stayed the same, both at a full 255, but the green *increased* from zero. And so the effect of adding green, which was *not* originally present, leaves the hue unchanged, but less intense. It is clear that if we keep lowering the saturation, the green will increase until it is equal to red and blue, thus eventually making, for 0% saturation, RGB = 255–255–255, which is white.

On the right side of figure 10.7 we lower the *value* from the example on the left. So the hue is still 300°(magenta), and the saturation is still 50%, but we have lowered the value from 100% to 51%. In this case *all* of the RGB values decrease *in proportion to each other*. So a lower value keeps the same *percentages* of red, blue and green, but all of the numbers are smaller. And so clearly, a value of 0%, whatever the hue and saturation, will give black.

For these examples so far, we have started with an equal mix of full red and full blue (255–0–255). What if we had an uneven mix of red and blue? The left image in figure 10.8 shows RGB = 125–0–199. This is a hue of 278°; it is to the blue side of magenta, as it has more blue than red. Since one of the colors (G in this example) is zero, the saturation is 100%. But neither R nor B is at a full brightness (255), and so this example has a value less than 100%. It is not shown here, but increasing the value would increase both R and B so that they kept their same proportions; at a value of 100%, blue would reach its maximum of 255.

The right side of figure 10.8 shows the effect of decreasing the saturation, from 100% for the left-hand example to 50% on the right. Before lowering the saturation, we had RGB = 125–0–199, with B as brightest. With the saturation lowered to 50%, RGB changes from 125–0–199 to 162–99–199. Notice that by lowering the saturation, *both* R and G increased in brightness—but in such a way that B is still brighter than R, and R is kept greater than the G that has been added. The effect is as if we mixed gray with the color on the left side of figure 10.8. The resulting

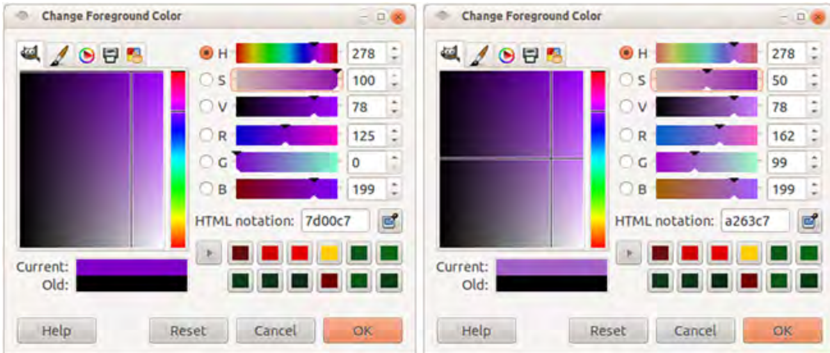

Figure 10.8. Dialog boxes from GIMP showing (left) the effect of lowering the value on an uneven mix of red and blue, and (right) the effect of then lowering the saturation. The chosen color is in the box labeled 'Current.'

color looks less intense (and also brighter overall), because the RGB values are more nearly equal to each other. If we were to decrease the saturation even more, red and green would get brighter still, in such a proportion that at 0% saturation we would have RGB = 199–199–199. Thus all trace of the original hue disappears, and we are left with a shade of gray.

The best way to understand the relation between HSV and RGB is to try these exercises yourself. The color dialogue in GIMP is an accessible option, but there are many other similar interactive tools, both online and as part of image-processing programs.

References

Conway B R 2003 Colour vision: A clue to hue in v2 *Curr. Biol.* **13** R308–10

Conway B R, Chatterjee S, Field G D, Horwitz G D, Johnson E N, Koida K and Mancuso K 2010 Advances in color science: From retina to behavior *J. Neurosci.* **30** 14955–63

Schmidt B P, Neitz M and Neitz J 2014 Neurobiological hypothesis of color appearance and hue perception *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **31** A195–207

IOP Concise Physics

The Physics and Art of Photography, Volume 2

Energy and color

John Beaver

Chapter 11

Filters

A filter is a piece of glass that fits over the front of the camera lens and selectively absorbs some wavelengths of light while allowing others to pass. Thus we can use a filter to alter the spectrum of light before it reaches the detector. There is a different but related meaning of the word 'filter' in photography—the use of digital processing to alter, according to a prearranged scheme, a digital photograph. In some cases such a *digital filter* might do very much the same thing that a physical filter would, only it does it *after* the picture has already been taken. But most digital filters are quite unlike anything one can do with a piece of glass placed over the camera lens.

The most common types of glass filters work by selective absorption. Since the filter absorbs some wavelengths of light better than others, it also transmits some wavelengths better than others. Whatever is *not* absorbed is transmitted. And so if the filter absorbs 90% of the light at a wavelength of 550 nm, then only 10% of the light *at that wavelength* enters the camera. Thus we can describe the properties of a filter with a *transmission curve*, a graph of the percentage of light transmitted versus wavelength. This concept is very similar to the reflection curve discussed in chapter 8, section 8.3—but the transmission curve describes the amount of light (at a given wavelength) that passes through the filter, while the reflection curve describes the same percentage for light reflected off the subject. There are other types of filter too, for which the light not transmitted is reflected (rather than absorbed), either through thin metal coatings or thin-film interference (chapter 8, section 8.3). But the use of such filters is less common in photography.

As is the case for a reflection curve, the filter cannot create light that is not already there; it can only subtract light, not add to it. Furthermore, it can only do this *separately for each wavelength*. One cannot, for example, take away some of the blue light and make it red. One can, at best, only take away some of the blue light while leaving the red light alone.

Thus a filter always decreases the total amount of light entering the camera, often at least somewhat for all wavelengths. And so the use of a filter usually requires one

doi:10.1088/2053-2571/aae504ch11

to alter the exposure to compensate. If one is using a TTL meter built in to a camera, then this is not much of a problem; the meter is looking through the filter too. But if a separate light meter is used, then one must apply an exposure correction. This is called the *filter factor*, and it is often printed on the side of the filter. Unfortunately, filter factors are usually written as the *factor*—rather than the number of exposure *steps* by which the exposure must be increased. Thus a filter factor of 2 means that only half the light overall gets through the filter, and this corresponds to a change of *one* exposure step, not two.

There is a special kind of filter called a *neutral density filter* that decreases all wavelengths equally. Such a filter, then, would decrease the total light but not alter its spectrum. A neutral density filter is useful, for example, in a circumstance where one wants to use a large aperture with a slow shutter speed in bright light. But most other filters do alter the spectrum of the light entering the camera, and that is the point. And so we can use a filter to accomplish an effect similar to changing the color of light falling on the subject. Or, to look at it a different way, the use of a filter can have an effect similar to using a detector with a different response curve. In either case, it alters not only the color of the light, but also what subjects appear darker or lighter than others.

11.1 Filters and black and white photography

Filters are widely used in black and white photography. Although no color information is recorded by a black and white detector, it still has its own response curve that interacts with the spectrum of the light entering the camera. By using a filter to alter the spectrum of the light entering the camera, one can exercise some control over what parts of the picture are rendered dark or light. See figure 11.1 for some examples of filters designed for black and white photography.

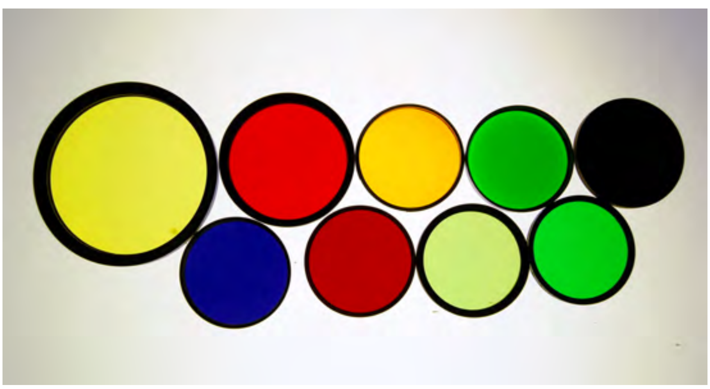

Figure 11.1. Colored filters are used in black and white photography *more* often than in color photography. Strongly colored filters can greatly restrict some colors (rendering those objects darker) while passing other colors (rendering those objects relatively brighter). This shows a selection of filters for black and white photography. The seemingly-black filter on the upper right is for technical applications; it passes only non-visible ultraviolet light. The deep blue filter on the lower left is unusual; most filters for black and white photography are reds, oranges, yellows, or greens.

Figure 11.2. The same black and white picture taken through three different color filters. The scene in color is portrayed on the left. The next three pictures are the same scene taken with black and white film through blue, green and red filters, respectively. Notice especially the brightness of the sky compared to the trees and grass.

Figure 11.2 shows an example of the same scene photographed with color film on the left, and then with black and white film through (from left to right) blue, green and red color filters. Unsurprisingly, the blue sky appears bright through the blue filter, but dark through the red filter. In this particular case, the green filter probably provides the most 'natural' rendering of the scene.

Even the more 'natural-looking' image through the green filter is still not the same as just a black and white version of the color picture. Notice in particular the larger plants in the foreground (mullein, I believe) compared to the more yellowish grasses. In the color picture they are of nearly the same value, and so they blend in with each other. The black and white picture through the green filter, on the other hand, better separates them from each other, giving the picture four layers (mullein, grasses, trees, sky) instead of just three (mullein/grasses, trees, sky).

This illustrates why filters are probably *more* commonly used for black and white photography than for color photography. A deep red filter makes the blue sky much darker, causing the white clouds to stand out in stark contrast. But if one were to use this same red filter with a color detector, the entire picture would be red! But with black and white photography one can use strong filters to drastically alter the spectrum of the incoming light, thus increasing or decreasing the contrast between objects of different colors.

The most commonly used filter for landscape black and white photography is a light yellow-green. This darkens the blue sky somewhat, making the clouds stand out more, and it also lightens up green foliage. A typical yellow-green filter of this type has a filter factor of less than 2, and so only a small exposure compensation is needed.

Another filter commonly used for outdoor black and white photography is deep red, often with a filter factor as high as 4 (and so two full steps of exposure compensation are required). Apart from making the blue sky appear very dark in comparison to clouds, a deep red filter greatly diminishes the effect of atmospheric haze, which can greatly lower the contrast in a distant landscape.

11.2 Filters and color photography

In color photography, filters are typically used simply to change the overall color balance of the picture. Remember that the spectrum of light entering the camera depends not only on the way different objects in the picture reflect light (because of their different reflection curves), but just as critically it depends on the spectrum of light illuminating those objects.

In table 11.1 I have listed some common examples of typical lighting sources. Each has its own characteristic spectrum, and in section 11.2.1 I describe some useful ways to categorize these light sources. But the reflected-light spectrum from a given object will be different, depending on the source of light that illuminates it. And thus, depending on the source of light illuminating the subjects in our picture, all of the colors will be altered from how they would appear with a different source of light.

Common sources of 'white' light vary significantly in their spectra—far more than it seems from our everyday use of those light sources. The basic reason for this should be clear from the discussion in chapter 10, section 10.3. The *color constancy* property of our eye/brain automatically adjusts for different light spectra illuminating things in the world. But the light detector in a camera is not a brain, and so it needs some help from us in order to make a photograph look natural. For example, what if the colors produced by our camera's particular detector look natural whenever our subject is illuminated by daylight? The colors of these same subjects will be rendered very *un*naturally by our detector, if they are illuminated instead by the incandescent tungsten light of an ordinary household light bulb. And the resulting difference between the two *photographs* will be far greater than our own

Table 11.1. Examples of different lighting sources. Each of these has its own characteristic spectrum and associated color temperature. Data adapted from Stroebel *et al* 2000, p 10, and http://www.en.wikipedia.org/wiki/Color_temperature.

Lighting source	Typical color temperature (K)
Candle	1800
Household tungsten bulb	2800
Tungsten photo flood	3200
moonlight	4100
Compact fluorescent	3000–5000
Tube fluorescent	5000
Horizontal daylight	5000
Vertical daylight	5500–6000
Electronic camera flash	6000
Daylight (overcast)	6500
Clear blue sky	20 000
Neon sign	NA

Figure 11.3. Filters for color photography are mostly used to change the overall white balance of the picture. More strongly colored filters are used to change the color of the lighting to match that which is assumed to be white by the light detector. The upper-left filter is to turn much-more-bluish daylight into the color of light emitted by a tungsten light bulb. The two blue filters on the top are for doing the opposite. The bottom row of filters is for *warming* the color (the three on the lower left) or *cooling* the color (the two on the lower right). The filter on the upper right is neutral density (gray), but it polarizes the light as well as dims it.

subjective experience of seeing the subject in the world when illuminated by these two different types of light.

Since a filter selectively absorbs some wavelengths more than others, we can use a filter to 'correct' for a particular source of light. In the above example, we could use a filter that alters the spectrum of the light from a tungsten bulb so that it looks like daylight. Then our detector—which in this example gives the right result for daylight—will record the image under tungsten light as if the subject had instead been illuminated by daylight. This example describes a very common use of filters in color photography; we will consider other examples as we go along. See figure 11.3 for some examples of filters for color photography. Notice that they are, for the most part, less 'colorful' than the filters for black and white photography shown in figure 11.1. A filter for color photography is typically used to make one variety of 'white' light look like another, and so the colors are of very low saturation.

11.2.1 Color temperature and white balance

Heat a solid to a high enough temperature and it will emit its own light. This *thermal radiation* is emitted by the objects all around us. But the thermal spectrum of a room-temperature object is dominated by wavelengths at about 10 000 nm—roughly twenty times the wavelength of visible light. To emit visible light, the temperature must be proportionally higher, as measured on the absolute temperature Kelvin scale. Room temperature on that scale is about 300 K, and twenty times that is 6000 K, roughly the temperature of the surface of the Sun—which is, needless to say, a very good emitter of visible light.

An emitter of purely thermal radiation (often called a *blackbody*) produces a particular broad, hump-shaped spectrum of light that peaks at a particular wavelength. And the higher the temperature, the *shorter* the wavelength at which most of the light is emitted. For a blackbody to peak in the middle of the visible spectrum, it must be roughly the temperature of the Sun. The hot filament of an incandescent light bulb emits mostly thermal radiation, but the temperature is significantly lower than that—only about 3000 K. And so such a light source most efficiently emits wavelengths longer than visible light, in the near-infrared part of the spectrum; what we see is the relatively smaller percentage of visible light it emits too.

The color of direct sunlight on a clear day is a complicated case; it has been altered by Rayleigh scattering by the air, for example. And so the direct sunlight is lacking in the blue that has been scattered away—the same wavelength that makes the sky blue. On a cloudy day, this is all mixed back together, and it usually seems to have no color at all—as white as anything can be. And so, although we must always be mindful of the fact that we see color *in context*, the 5770 K blackbody of the Sun is about as white as it gets.

A blackbody hotter than the Sun (the star Vega is a good approximation) emits the peak of its spectrum at wavelengths *shorter* than that of the Sun. And so, compared to the solar spectrum, it emits more short-wavelength light and less long-wavelength light; it appears more bluish in comparison to sunlight. The cooler filament of an incandescent light bulb on the other hand (or the cooler star Betelgeuse), emits more long-wavelength and less short-wavelength light, compared to the Sun. And so the incandescent light would appear more reddish in color. Seen in isolation, all three of these examples appear pretty-much *white* to our eye/brain, which seems to automatically see an almost-white source of light as white, unless there is a different color context to compare it with. The light of a quartz-halogen incandescent flashlight may look brilliant white at night. But in the daytime, when compared to direct sunlight, it seems distinctly yellow.

Many sources of light do *not* emit like a blackbody—fluorescent lights, LEDs, the blue sky and neon signs are all good examples. For these *non-thermal* sources of light, there is no simple connection between temperature and their spectra; indeed the spectra of these non-thermal sources of light are often far more complex than the simple hump shape of a blackbody spectrum. But for any source of 'white-ish' light, we can still make an overall comparison of this complex non-thermal spectrum to the overall effect of an ideal blackbody spectrum. We can ask the question, 'If we were to compare the color of this (for example) fluorescent light to various blackbody thermal sources of different temperatures, what blackbody temperature would appear the same color?' This is the basis of the concept of *color temperature*.

> The color temperature of a source of light is the temperature of the perfect thermal blackbody light source that would appear the same color.

And so, roughly speaking, a more bluish source of light has a higher color temperature than a more reddish source of light. If the source of light is *too* different from that of a blackbody, then the concept of color temperature is ill-defined and not at all useful. And so although a neon sign appears an odd sort of reddish-orange, the concept of color temperature is inappropriate to describe its color; color temperature is most useful for differentiating between sources of light that are nearly white. If two almost-white sources of light are placed next to each other, the one of lower color temperature will look redder while the one of higher color temperature will look bluer. See table 11.1 for examples of sources of light and their approximate color temperatures.

So what color temperature, exactly, *is* pure white? There is no simple answer! Recall the discussion in chapter 10, section 10.3 of 'The Dress.' Part of the reason some see the dress as black and blue while others see it as white and gold is because different individuals, when they look at that same picture, make different unconscious assumptions about the whiteness of the light illuminating the dress. The point is, what we see as 'white' depends as much upon the context—in complicated ways that are only partially understood—as it does upon the particular spectrum of the light. Most of the time, we humans mostly agree what 'white' is when we look at the same scene or the same picture. 'The Dress,' because of the particularly ambiguous color context of the picture, is an odd case for which we split down the middle and disagree.

And so, regarding the spectrum of white—we know it when we see it. But it is *always* context dependent. In natural scenes, with a lot of information about the source of light, we seem to make similar assumptions when we look at a given picture. And so it is possible to match—at least roughly—the response curve of the detector to the spectrum of the light source in such a way that the colors in the picture are likely to appear similar to how we perceive them in the real world.

But if we use the same detector with a light source of a different color temperature, then the results will likely look unnatural. Figure 11.4 shows a picture taken in 2700 K 'tungsten light.' In the example on the left, a detector optimized for daylight was used. Since tungsten light is of a much lower color temperature, the picture appears too yellow. The picture on the right was taken in the same light, but with the detector response curve altered so as to properly render tungsten light.

Figure 11.5 shows the opposite effect. Both pictures were taken in daylight, but for the photo on the left the detector response curve was mistakenly optimized for tungsten light. Expecting a more yellow light source, the detector rendered the picture too blue.

For photochemical detectors such as traditional film, the response curve is a fixed property of the detector. And so there are versions of film that have a response curve designed for the 3000 K color temperature of an incandescent light (called tungsten

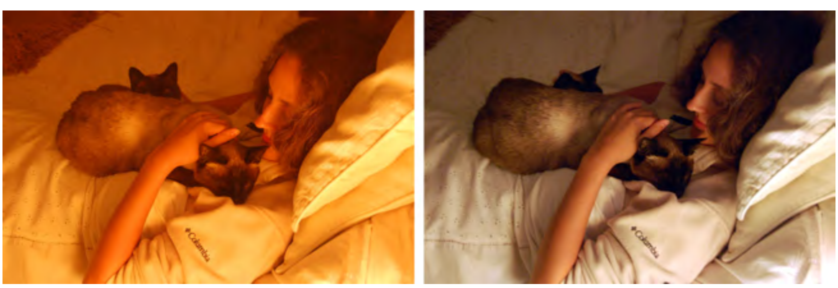

Figure 11.4. Left: an image taken in tungsten light (2700 K) with a detector optimized for daylight (5000 K). Right: the picture retaken in the same light, but with the detector optimized for tungsten light.

Figure 11.5. Left: an image taken in daylight (5500 K) with a detector optimized for tungsten light (2700 K). Right: the picture retaken in the same light, but with the detector optimized for daylight.

film), and films designed instead for 5770 K daylight. Use one with the light intended for the other, and the examples in figures 11.4 and 11.5 result.

With the photoelectric detectors in a digital camera, on the other hand, it is possible to alter the response curve internally (within limits). And so the same detector can properly record an image according to the illuminating spectrum of either figure 11.5 or 11.4. Choosing a particular source of light—and thus the appropriate detector response curve to properly render colors of objects illuminated by that light—is called setting the *white balance*. Most digital cameras have pre-set response curves for common sources of light, and offer these as choices. Typical examples are daylight, cloudy daylight, tungsten, fluorescent and electronic flash. Most also have an option for *auto white balance.* This feature makes reasonable assumptions about the typical colors in a scene, and so the algorithm's success depends upon how typical is the scene. Cameras with more elaborate features allow for the white balance to be defined manually by the photographer from picture to picture. The photographer takes a picture of a white card illuminated by the same light that illuminates the subject to be photographed. The camera then uses that photograph to internally define the meaning of 'white' for the subsequent picture of the subject.

11.2.2 Filters and color temperature

The transmission curve of a filter alters the spectrum of the light that passes through it. And so one can use a carefully-designed filter to change the spectrum of the light to match the expectations of the light detector. The most common example is the distinction between the 5700 K color temperature of daylight and the ~3000 K color temperature of the tungsten light from an incandescent, indoor photo floodlight. There are special filters designed specifically to alter the spectrum of the light entering the camera from one of these to the other. The #80 filter is blue in color, and it is made to alter the spectrum of tungsten light so it looks like daylight. Subtle variants—80A, 80B and 80C—are designed for specific types of tungsten flood lights. The #85 filter (also with slightly-different 'B' and 'C' options) does the opposite; it makes daylight look like tungsten light. As such, it looks orange in color. The first filter in figure 11.3 is an 85B filter, while the next two are different #80 filters.

Sometimes one wants to alter the picture as if the subject were illuminated by a light source of only a slightly-different color temperature. There are many special filters for this. If a given filter is designed to make the subject appear as if illuminated with light of a slightly-higher color temperature than 'white,' it is called a *cooling filter*. And if the filter makes the subject appear as if illuminated by a slightly-lower color temperature source, it is called a *warming filter*.

Yes, it is the *warming* filter that makes the subject look as though it were illuminated with light of *lower* color temperature. That sounds crazy, but the terms 'warming' and 'cooling' come from our emotional response to color, which turns out to be opposite the *physical* relation of color to temperature. A 'warm' light, in an emotional sense, looks more reddish, while an emotionally 'cool' light looks more bluish. Color temperature on the other hand, works in exactly the opposite sense. The three filters on the bottom left of figure 11.3 are warming filters, while the two on the lower right are cooling filters.

The color temperature of direct sunlight varies greatly with the angle of the Sun. Low-angled sunlight must pass through more air on its way to you, and so more short-wavelength light is scattered out of your line of sight by Rayleigh scattering. This loss of short wavelengths lowers the color temperature of the sunlight—dramatically so when the Sun is just rising or setting, or if there is haze in the atmosphere. There is a particular angle of sunlight, most typically from one to two hours before sunset or after sunrise, for which we humans seem to find the color temperature of sunlight to be particularly appealing. And so a warming filter can be used to simulate this *magic hour* or *golden hour* at other times of the day.

As a further example of the context dependency of color—and especially of what we see as white—notice from table 11.1 that the color temperature of moonlight is *lower* than that of sunlight. This means that moonlight should appear 'warmer' than sunlight, but we mostly experience the opposite effect. moonlight is 'cold' and 'silvery,' while every child uses a yellow crayon to portray the Sun. The color temperature of moonlight (even when the Moon is high overhead) really is slightly lower than sunlight because the reflection curve of the lunar surface is, on average, non-gray. There are even some lunar soils that are noticeably orange in color when seen up close. But the color temperature of moonlight is still much higher than even the hottest quartz incandescent bulb—which seems brilliantly white when it is the sole source of light. One obvious reason that direct sunlight seems yellow is that we see it in the context of a blue sky.

11.3 Polarizing filters

The filter on the top right in figure 11.3 seems to have no particular color at all, and it is in fact designed that way. It is a *neutral-density filter*—a filter with a transmission curve that is constant over the visible spectrum. Such a filter is sometimes useful for reasons having to do with exposure, as discussed in chapter 4. But this particular example, although it does act as a neutral-density filter, has a different purpose; it is a *polarizing filter*.

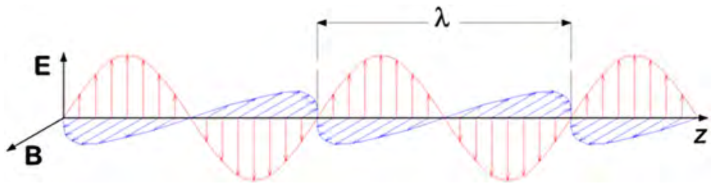

Figure 11.6. A polarized electromagnetic wave traveling from left to right. The electric and magnetic fields (marked **E** and **B**) in a polarized wave maintain a particular orientation. In an unpolarized wave the fields have random orientations perpendicular to the direction of travel, and to each other (graphic: P.wormer - Own work, CC BY 3.0).

Brightness and wavelength are not the only ways in which two sources of light can differ from each other. Light is an electromagnetic wave, a changing pattern of electric and magnetic fields that travels through space. And these fields point perpendicularly to each other in space, and *the light 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.

For most sources of light, countless individual atoms and molecules are each producing light in their own fashion, with little coordination between them. Thus the 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 11.6. 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. Rayleigh scattering is a good example. 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; 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 that is everywhere 90° from the Sun is strongly polarized.

Reflections provide another important example. Whenever light reflects off a non-metallic surface, some polarizations reflect more than others. At a particular angle, called the Brewster angle, the light reflected off a smooth non-metallic surface is completely polarized. For most surfaces in air, reflections are most strongly polarized when the light glances off the surface at a shallow angle.

We do not see the effects of polarization directly because, to the human eye or to the light detector in a camera, one polarization looks just the same as any other. A *polarizing filter* cleverly employs long-chain molecules to effectively block light polarized in one direction, while allowing polarizations at right angles to pass. For unpolarized light, which has all polarizations at once, this means the polarizing filter blocks half the light.

But if the light entering the filter is *already* polarized, then it may block all or none of it, depending on how the filter is oriented in space. And so by simply rotating such a filter, it can be used to alternately pass or block polarized light. A common use is to darken the blue sky relative to the clouds in a color picture. See figure 11.7. For black and white photography one can simply use a deep red filter to darken the sky. But this option is unavailable for color photography, unless one wants *everything* to be deep red. Since the blue sky is strongly polarized, one can simply rotate the polarizing filter until the sky is at its darkest. Objects in the picture that are not strongly polarized will remain unaffected, and so contrast is increased.

Figure 11.7. A polarizing filter was used to greatly darken the blue sky and reduce the glare off of the rock surfaces. The polarizing filter was oriented so as to reduce the glare off the (mostly) vertical surfaces, but this *increased* the glare on the horizontal surfaces.

A polarizing filter may also be used to reduce glare in order to make colors more vibrant. Glare essentially mixes skylight (since it is a mirror-like reflection) with the diffusely-reflected light from the subject, and this added white light reduces the saturation of the colors. A polarizing filter can be used to reduce this glare, because light glancing off a surface is often strongly polarized. See figure 11.7 for an example.

Reference

Stroebel L, Compton J, Current I and Zakia R 2000 *Basic Photographic Materials and Processes* 2nd edn (Waltham, MA: Focal Press)

Chapter 12

‘Color’ in astronomy

In chapter 10, section 10.1 we considered the opponent-process model of color perception. The brain does not directly process the individual red, green and blue sensitive stimuli. Rather, these signals are pre-processed, and combinations are sent instead. There may be a physical ‘logic’ to this; it is sometimes a good way to get the most out of the information available. Astronomers, with no conscious awareness of advances in the neuro-physiology of the human retina, have long taken a similar approach.

It is very important for an astronomer to observe the *spectrum* of an astronomical body; the spectrum gives us the biggest clues as to the source’s physical nature. And so an astronomer will often use a telescope to gather the light from a star or galaxy, and then disperse that light with a spectrometer. The problem, however, is that taking a spectrum is very difficult and time consuming. And what if one wants to analyze the light of, for example, the thousands of stars that make up even one single star cluster? Often it is simply impractical to gather a full spectrum of so many astronomical objects, especially if they are very faint.

And so astronomers often settle for the next best thing, what is called *filter photometry*. Instead of sending the light through a spectrometer, it is sent through a color filter that allows only a particular range of wavelengths to pass. This process is then repeated through *different* colored filters. And thus, the astronomer takes multiple pictures of the same object, with each picture representing the overall amount of light for a different part of the spectrum.

If an astronomer were to do this through red, green and blue filters, she would be doing something very much like what is done by the cone cells in the human retina. Instead, the filters are carefully chosen based on the particular kinds of astrophysical questions the astronomer is trying to answer—but the basic idea is the same. In fact the most common system traditionally used by astronomers involves three filters that, coincidentally, have much in common with the S, M and L cone cells of the human eye. But the responses are at shorter wavelengths overall; they are called U

doi:10.1088/2053-2571/aae504ch12

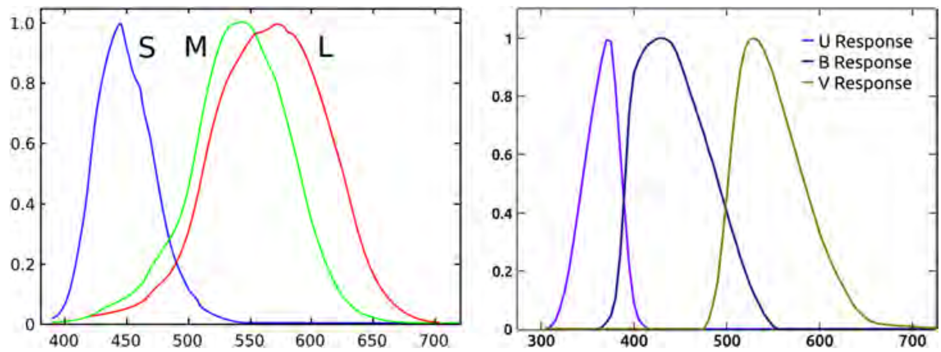

Figure 12.1. Left: the wavelength responses of the short (S), middle (M) and long (L) wavelength cone cells in the human retina. Right: the wavelength responses for the U, B and V filter system used by astronomers (data from Ažusienis and Straižys 1969).

(for ultraviolet), B (for blue) and V (for 'visual'—i.e. the middle of the human visual part of the electromagnetic spectrum). If more filters are included—there is no reason there has to be only three—then even more information about the spectrum of the object can be inferred. And so this system is often extended to include R (for red) and I (for infrared). See figure 12.1 for a graph that shows the wavelength responses of these filters, next to the responses of the cone cells of the human eye.

But as is the case for the human eye, astronomers do not analyze these colors in terms of the individual measurements from the filters. Instead, they use combinations of these measurements, and such a combinations is called, somewhat metaphorically, a *color index*, or just simply, a 'color.' And so, instead of publishing the measurements obtained through the filters U, B, and V, the astronomer publishes V, U−B and B−V.

Compare this to the pre-processing of the signals from the three types of cone cells, S, M and L. The brain seems to interpret not S, M and L separately, transforming various combinations into colors and brightnesses. Rather, the brain seems to act upon something more like M−L, S−(M+L) and S+M+L (Conway 2003; Conway *et al* 2010; Schmidt *et al* 2014). The latter combination gives the overall brightness, with no color information in and of itself, and so is analogous to the astronomer's V magnitude. The other two combinations contain color information, and they are analogous to the astronomer's color indexes U−B and B−V. The color indexes represent the excess of one filter measurement over another, and so represent how that part of the light spectrum is *tilted.*

We can see how the color index relates to what we normally think of as color. And so, for example, a reddish star would give a stronger signal through the yellowish V filter than through the bluish B filter. While a bluish star would do the opposite. Of course, when talking about an ultraviolet or infrared filter, these are outside the range of human perception altogether. And so a 'color' such as U−B would have no direct relation to any human-perceived color. But it is not difficult to see that the idea is the same, and that it tells *something* about the nature of the source's spectrum (its tilt).

Astronomers sometimes go one further and calculate even more complex combinations of the basic measurements, producing a 'color–color.' There is another filter system used by some astronomers—called the Strömgren system—that uses direct measurements through four filters, called u, v, b and y. But again, the individual measurements are never even published. Instead, astronomers work with the following: y, b−y, (u−v)−(v−b) and (v−b)−(b−y). The first of these (y alone) gives the overall brightness of the spectrum, the second (b−y) gives the *tilt* of the spectrum at a particular wavelength, while the remaining two give the *curvature* of the spectrum over two different wavelength ranges. For this particular example, information about the brightness, distance, size, temperature and age of the star can be determined, as well as the amount of intervening dust that happens to lie in our line of sight to the star.

See figure 12.2 (data from Beaver *et al* 2013) for an example with the star cluster NGC6705 in the constellation Scutum. It shows a plot of y versus b−y for all of the many hundreds of stars in the star cluster. This basic 'color-magnitude diagram' is one of the most important tools of any stellar astronomer. It shows the relation between the overall brightness of the star (top is bright; bottom is dim), and its overall color, and thus its *temperature* (bluer, and so hotter on the left; redder, and so cooler on the right). Indirectly, it tells us the size of the star. The little group of stars on the upper-right of the diagram are called *giants*. They are both cool and very luminous, and the only way to be both of those at once is to be very large.

And so we start out with four separate measurements, and we end up with four different numbers that are combinations of what was directly measured. But these odd mathematical combinations give more useful information about the spectrum of the observed star than do the individual measurements of u, v, b and y alone. There is another important reason that astronomers traditionally do not publish the direct

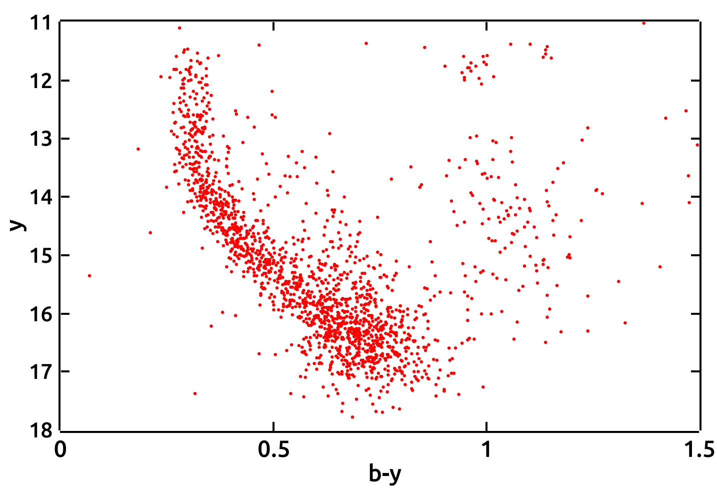

Figure 12.2. A 'color-magnitude' diagram for the star cluster NGC 6705. Each point represents an individual star in this cluster of many hundreds. The vertical axis is the overall brightness (through the y filter), while the horizontal axis is the *color index* (b−y).

measurements through their filters, instead only publishing color combinations. For technical reasons, a color such as B−V is sometimes easier to measure accurately than the individual B and V brightnesses. Accurate measurement of the individual filter brightnesses requires a complex process of calibration. But in many situations when measuring the *difference* between B and V, much of that calibration partially 'cancels out.' And this leaves us with a measurement of B−V that is, surprisingly, more accurate than the measurements of either B or V alone. And the eye/brain combination apparently includes specific physiological structures—cone-opponent retinal ganglion cells—that directly measure *differences* in the cone responses rather than measuring the individual responses directly (see chapter 10, section 10.1).

References

Ažusienis A and Straižys V 1969 Improved determination of the response curves and parameters of the UBV system. summary of results *Sov. Astron.* **13** 316

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

Conway B R 2003 Colour vision: A clue to hue in v2 *Curr. Biol.* **13** R308–10

Conway B R, Chatterjee S, Field G D, Horwitz G D, Johnson E N, Koida K and Mancuso K 2010 Advances in color science: From retina to behavior *J. Neurosci.* **30** 14955–63

Schmidt B P, Neitz M and Neitz J 2014 Neurobiological hypothesis of color appearance and hue perception *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **31** A195–207

Chapter 13

Color experiments with EP photography

In chapter 8, section 8.4.2, we saw that the unfixed, printed-out silver gelatin enlarging paper used in EP photography can result in subtle colors from wavelength-dependent scattering. The precise result depends on many factors, one of the most important being the particular brand and variety of enlarging paper. Over the century-long history of silver gelatin emulsions, there have been hundreds of varieties produced.

But there is an even more intriguing result: certain varieties of papers produce *different* scattering colors when exposed to *different* colors of light. And this means that it is possible to use this light-sensitive material to make a *color* picture, even though it was intended by the manufacturer to be used only for black and white prints. Figure 13.1 shows two examples. The colors in the images, although adjusted digitally, all derive directly from the original EP negatives that were exposed in the camera. And these negatives were made from enlarging paper intended only for black and white printing. This is especially surprising when we consider that the paper is not even sensitive to much of the visible part of the spectrum; it is only sensitive to ultraviolet, violet, blue, and green light.

Figure 13.2 shows another example, along side an image of the same scene made with an ordinary digital camera. The color of the EP photograph looks surprisingly realistic, but a close inspection shows that it not full color. Rather it is a *duotone* image, with only blue and orange, and various combinations in between. The duotone colors just happen to match up well with most of the actual colors in the scene, in particular the blue sky and the blond-colored wood fence in the background.

For figure 13.3[1], physics student Gabriella Cutié Rodriguez exposed several different types of accelerated enlarging paper to the same source of light, but

[1] First presented at the symposium *Pixels, Pallettes & Perception*, held March 2–3, 2018 in Madison, Wisconsin.

doi:10.1088/2053-2571/aae504ch13

Figure 13.1. Both of these pictures were made directly from EP negatives exposed in cameras. Although the silver–gelatin enlarging paper used for the exposures was intended only for making black and white prints, a color image can sometimes be produced. The color is duotone, rather than full color, but it sometimes correlates well with colors in the world.

Figure 13.2. The image on the left is a full-color picture taken with an ordinary digital camera that separately records the image in three different colors (red, green and blue). The picture on the right is the same scene imaged with ephemeral process photography. It is only a duotone image, consisting of various mixes of two tones only (blue and orange). But for this particular scene, the image looks (to humans anyway) almost as natural as if it were full color.

through two different colored filters. The circles on the left side of each exposure were made through a yellow 'K2' filter, of the sort used by photographers to darken the blue sky in a black and white picture. It blocks only wavelengths shorter than about 475 nm, in the blue–green part of the spectrum, while passing longer wavelengths. The circles on the right side of each exposure were through a Schott UG-11 ultraviolet filter. It blocks all light that has a *longer* wavelength than about 375 nm, in the near ultraviolet part of the spectrum. The wavelength responses of the two filters can be seen in figure 13.4, superimposed upon the approximate overall wavelength response of the enlarging paper itself.

Figure 13.3. Color filter tests of ephemeral-process exposures with twelve different enlarging papers. For each test, the paper was exposed through a yellow filter (Hoya K2) on the left and an ultraviolet filter (Schott UG11) on the right. Left: a photograph of the original test pieces. Right: the same, but with the contrast stretched and the saturation increased. Note that most of the papers respond similarly to the light through the two filters, but the first and last papers in each group show marked differences in hue when exposed through the two filters.

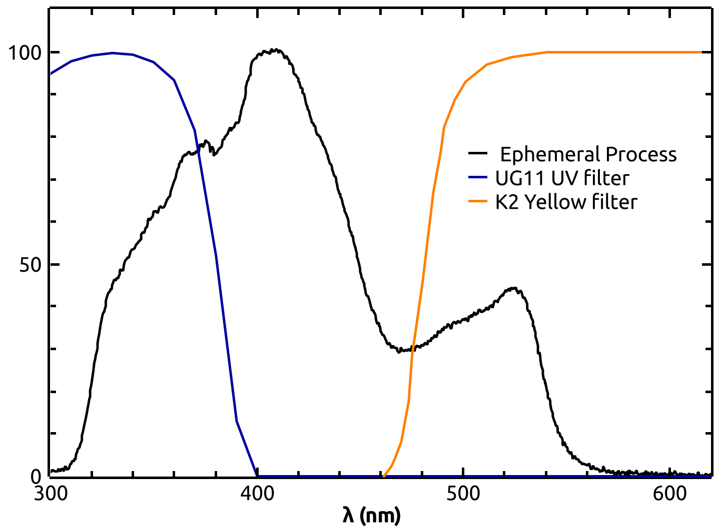

Figure 13.4. The detected spectrum of one type of EP paper as illuminated by sunlight, compared to the transmission curves of the two filters used for figure 13.3 (filter transmission data from http://www.schott.com and http://www.hoyafilterusa.com).

It is clear from figure 13.3 that the scattering colors produced are either blue or orange, depending on the variety of paper, and this has been born out with similar experiments with many other paper varieties as well. For most of the varieties of paper shown in the test, the same scattering color resulted whatever the color of the light. But there are two tests (the first and the last) in figure 13.3 that are different: the exposure through the yellow filter came out a pinkish-brown color, while that through the UV filter came out bluish. This is especially evident if the paper is scanned, and the contrast and color saturation is digitally increased.

But figure 13.4 shows that the EP paper is sensitive only to wavelengths shorter than about 540 nm, which is in the green part of the spectrum. It is not sensitive at all to the yellow, orange and red parts of the spectrum. So how is it that reds and

Figure 13.5. Color pictures made from in-camera EP negatives. The process of inverting the values from negative to positive makes the hues turn backwards; yellow turns to blue and vice versa. For the image at the upper-left I have left them that way, while for the others I have switched the hues back.

yellows are visible in these pictures, and they correspond to reds and yellows of objects in the world?

The answer seems to be partly luck. Recall from chapter 10, section 10.2 that just because something looks yellow, it does not mean that it reflects light in the yellow part of the spectrum. The yellow filter used for the exposures in figure 13.3 is a case in point. The transmission curve in figure 13.4 shows the filter transmits very little light at wavelengths shorter than a cutoff at about 480 nm, in the blue–green part of the visible spectrum. But it transmits light at essentially *all* visible wavelengths longer than that—green, yellow, orange and red light. The EP paper, however, knows nothing of those longer wavelengths; it only responds to that little bit between about 480–540 nm. And apparently this particular EP paper turns reddish-orange when exposed to only that long-wavelength tail end of the paper's response curve. Shorter wavelengths on the other hand, make the paper turn more blue in color.

And so it is luck that many things in the world have spectra that look something like these two filter responses—either blocking long or short wavelengths, and so either making the paper turn blue or reddish-orange or some mix of the two. Figure 13.5 shows four more examples of images made from scans of EP negatives made directly in cameras.

Note that it is the longer wavelengths—associated with the tail end of the spectra of yellow, orange and red objects in the world—that make the paper turn reddish-orange. And it is the short wavelengths—associated with blue objects in the world—

Figure 13.6. EP photograms may make especially intense colors, if the right combination of paper, objects, and light source are chosen. The example on the left was illuminated by a single-wavelength 405 nm laser. But the mineral crystals fluoresced, producing their own light at longer wavelengths. For both of these images, I have left the EP colors reversed from their real-life counterparts.

that make the paper turn bluish in color. And so the duotone hues correspond directly to those in the world. But regarding the *values*, the paper responds as a negative; brighter light makes the paper darker. And so to make a natural-looking image from the scanned negative, it must be digitally inverted to a positive. Ordinarily, this process would also reverse the hues to their complementary colors—and so blue would turn to yellow and vice versa. But since the hues are already in the correct direction, inverting the image from a negative to a positive makes the colors go backwards. For the upper-left image in figure 13.5, I have left them that way, but for the others I have digitally rotated them back around the color wheel by 180°[2].

Photograms are an especially good technique for exploring the color possibilities of ephemeral process. Objects can be chosen in part for their color filtering properties, in order to more strongly emphasize the often-subtle EP colors. The source of light can also be better controlled than when taking EP pictures directly with a camera. For the image *Blue Monarch*, on the right side of figure 13.6, it seemed fitting to leave the colors reversed, rather than switching the hues back to something more natural.

One can even make a positive *print* using this same technique. Simply expose the paper in contact with a *negative* instead. The hues will come out reversed. But if a digital negative is used, the hues can be reversed in the negative instead, so they come out correctly on the positive print. See the left side of figure 13.7. Such a print is partly where 'ephemeral process' gets its name. It is still sensitive to light, and so can only be viewed briefly in dim light. The rest of the time, it must be stored in the dark; it is damaged by the act of looking at it.

The right side of figure 13.7 shows another way to get very strong colors. The EP paper was exposed in contact with a positive print, made from an ordinary camera, thus making an EP negative. The negative is then scanned, digitally inverted and adjusted in the same ways as for in-camera negatives. But the reds, yellows and blues

[2] This is easy to do with, for example, the hue/saturation/value dialogue in GIMP.

Figure 13.7. The image on the left is a photograph of an ephemeral process positive print, made from a transparent color negative. The digital negative was printed on plastic with a laser printer. The image on the right was made from an EP negative exposed not in a camera, but in contact with a color positive print from an ordinary camera.

in these *chromogenic prints* correspond very well to the spectral responses of these colors in some EP papers, and very intense colors are possible. For more-detailed instructions on how to perform your own EP photography experiments, see appendix C.

Appendix A

Lambertian reflectors and in-camera image intensity

Let us imagine a flat, matte-finished reflective circular disk of radius, S_O (for 'object size'). And let us say that our subject is illuminated by light with an illuminance, $I_{\rm SUB}$, in W m^{-2}, and that the disk reflects a fraction, R, of the light that falls upon it. At a distance d directly in front of the disk, we place a camera. The light reflected from the disk will spread out over a hemisphere in the forward direction with area equal to $2\pi d^2$, and will have an intensity before entering the camera lens of B_C. The camera lens will then focus this light to a circular image of radius S_I, and the light in this image that falls on the detector will have an illuminance, I_D. Our question is then, what is the mathematical relation between $I_{\rm SUB}$ and I_D?

First, we must recognize that the total power, P_R reflected by the disk is the illuminance on the disk multiplied by its reflectivity and by its surface area:

$$P_R = I_{\rm SUB}R(\pi S_O{}^2) \tag{A.1}$$

This power then spreads out over our hemisphere of radius, d, before reaching the camera. And so over that hemisphere that includes the camera it has an *average* intensity, power per area, of simply $P_R/(2\pi d^2)$. But is that intensity the same over different parts of the hemisphere? In particular, is the intensity the same directly in front of the disk as it is at a large angle?

If we do assume that the disk reflects light in such a way that the intensity, at distance d from the disk, is the same in every direction, we could call the disk an *isotropic reflector*. The intensity at the camera, B_C, would then be:

$$B_C = \frac{P_R}{2\pi d^2} \tag{A.2}$$

$$B_C = \frac{I_{\rm SUB}R\pi S_O{}^2}{2\pi d^2} \tag{A.3}$$

doi:10.1088/2053-2571/aae504ch14

$$B_C = \frac{RS_O{}^2}{2d^2} I_{\text{SUB}} \tag{A.4}$$

In order for this to happen, the surface would have to appear brighter per surface area as seen from an angle than as seen from straight-on, in order to make up for the fact that less area is seen from an angle. A more realistic assumption is the *Lambertian reflector*, which emits the same *specific intensity* in every direction. But to do so, a Lambertian reflector must emit relatively *greater* energy flux normally than at an angle. And so the Lambertian reflector emits a specific intensity (W m^{-2} per solid angle) that is constant with direction, but an intensity (W m^{-2}) that varies with angle (at a given distance). Thus by assuming an isotropic emitter, we must have *under*estimated the energy flux emitted in the forward direction.

It requires calculus to calculate the face-on reflected energy flux from a Lambertian reflector. But the correct result turns out to be simply twice the equal-energy-flux value given by equation (A.4). To show this, imagine that a certain power, P, is reflected from our circular spot. The intensity then, at a given point distant from the spot, is that power per unit area crosswise to the ray. So let us imagine we are at a distance d from that spot, off in some direction making an angle of θ to the normal.

For the simple but unrealistic case of an isotropic reflector, the intensity would be the same for any θ. For a Lambertian reflector, on the other hand, the intensity is proportional to the projected area of our circular spot. If θ is large, then from that perspective one sees a spot of small area, and so the intensity of the light is proportionally less. We can summarize these results as follows:

$$B_I(\theta) = B_{I0} = \text{constant} \tag{A.5}$$

$$B_L(\theta) = B_{L0}\cos(\theta) \tag{A.6}$$

where the subscripts refer to isotropic and Lambertian reflectors, and the zero subscript refers to the intensity for $\theta = 0°$.

Since power equals intensity times area, the integral of the intensity over the area of a hemisphere of radius d must equal the total power reflected, P_R, for both the isotropic and Lambertian cases:

$$P_R = \int B_I(\theta)dA \tag{A.7}$$

$$P_R = \int B_L(\theta)dA \tag{A.8}$$

where dA is an infinitesimal element of surface area at distance d, and the integral is carried out over the forward facing hemisphere. We can think of the area element, dA, as a thin ring on that hemisphere, defined by the space between θ and $\theta + d\theta$. This ring would have an area given by the product of its thickness ($d \times d\theta$) and its circumference ($2\pi d \sin\theta$). And so we have:

$$dA = 2\pi d \sin\theta \times d \times d\theta \tag{A.9}$$

$$dA = 2\pi d^2 \sin\theta d\theta \tag{A.10}$$

To integrate over the forward hemisphere, the integral over θ would be carried out from $0 \longrightarrow \pi/2$, and thus equations (A.7) and (A.8) become:

$$P_R = 2\pi d^2 B_{I0} \int_0^{\pi/2} \sin(\theta)d\theta = 2\pi d^2 B_{I0} \tag{A.11}$$

$$P_R = 2\pi d^2 B_{L0} \int_0^{\pi/2} \cos(\theta)\sin(\theta)d\theta = \pi d^2 B_{L0} \tag{A.12}$$

Note that the two definite integrals evaluate to 1 and $\frac{1}{2}$, respectively. Thus we have $B_{L0} = 2B_{I0}$, and equation (A.4) should instead be:

$$B_C = \frac{RS_O^2}{d^2} I_{\text{SUB}} \tag{A.13}$$

Let us now assume that this reflected light is incident normally upon a lens of diameter D and focal length F. Thus, the power in watts, P_L, that makes it through the lens would be:

$$P_L = \pi \left(\frac{D}{2}\right)^2 B_C \tag{A.14}$$

If we assume the lens focuses this light to a circle of radius S_I, on the detector, the illuminance in W m^{-2}, I_D, on the detector would then be,

$$I_D = \frac{P_L}{\pi S_I^2} \tag{A.15}$$

$$I_D = \frac{\pi \left(\frac{D}{2}\right)^2 B_C}{\pi S_I^2} \tag{A.16}$$

$$I_D = \frac{\pi \left(\frac{D}{2}\right)^2 \frac{RS_O^2}{d^2} I_{\text{SUB}}}{\pi S_I^2} \tag{A.17}$$

$$I_D = \frac{D^2 RS_o^2}{4d^2 S_I^2} I_{\text{SUB}} \tag{A.18}$$

We can relate this to the image distance, d_I, the distance between the lens and the in-focus detector:

$$d_I = \frac{S_I}{S_O} d \tag{A.19}$$

And so we have:

$$I_D = \frac{D^2 R}{4 d_I^2} I_{\mathrm{SUB}} \tag{A.20}$$

or:

$$\boxed{I_{SUB} = \frac{4f^2}{R} I_D} \tag{A.21}$$

where f, the focal ratio of the lens, is defined by $f \equiv d_I/D \approx F/D$ when $d >> F$.

We can use this result to compare the time, t_{SUB}, that a given detector is illuminated with the same light as the subject to the time, t_D, required for the detector in the camera to receive the same exposure, E, from the image of the subject, noting that exposure is the product of illuminance and time.

$$E = It \tag{A.22}$$

$$E = I_D t_D \tag{A.23}$$

$$E = I_{\mathrm{SUB}} t_{\mathrm{SUB}} \tag{A.24}$$

$$I_D t_D = I_{\mathrm{SUB}} t_{\mathrm{SUB}} \tag{A.25}$$

$$\frac{t_D}{t_{\mathrm{SUB}}} = \frac{I_{\mathrm{SUB}}}{I_D} \tag{A.26}$$

And so we have:

$$\boxed{t_D = \frac{4f^2}{R} t_{\mathrm{SUB}}} \tag{A.27}$$

IOP Concise Physics

The Physics and Art of Photography, Volume 2

Energy and color

John Beaver

Appendix B

A practical way to denote VLS-detector 'speed'

Since exposure times for VLS photography are measured in minutes rather than fractions of a second, it makes sense to abandon the idea of detector 'speed,' and instead define a detector *slowness*, which I will denote by S. There is no reason for this to have any connection at all to ISO speed, and so we can define it in whatever manner is convenient.

The ordinary ISO speed uses numbers that make sense when applied to ordinary film or digital detectors, for which the typical exposure time is a small *fraction* of a second. And so it is natural for the ISO speed to be linked to typical *reciprocals* of exposure times. This is the foundation of the sunny-16 rule: the ISO speed is roughly equal to the reciprocal of the exposure time on a sunny day at $f/16$ (a typical aperture setting for sunny-day film photography). And so, for example, with ISO 200 speed film a proper exposure would be 1/200 s at $f/16$. And so the definition of ISO speed for film is linked in an intuitive way to its typical use. And furthermore, ISO speed, when applied to film, gives us easy-to-use numbers such as 100, 200, 400, etc.

But for VLS photography, typical sunny-day exposure times are measured in minutes, not *fractions* of a second. Corresponding ISO speeds are tiny fractions requiring scientific notation (or careful counting of zeros) to express. Furthermore, one is more likely than not to use small focal ratios in VLS photography, even on a sunny day. And so instead of defining a *speed* based on the reciprocal of the sunny-day exposure time in seconds at a large focal ratio, I choose to do the opposite. I define instead a *slowness*:

> The slowness, S, of a VLS detector is defined according to a 'Sunny 5.6' rule. Under sunny-bright, $EV_{100} = 15$ conditions, the slowness equals the proper exposure time *in minutes* at $f/5.6$. And so, under sunny-bright lighting, a detector with a slowness of 15 will require a 15 min exposure at $f/5.6$.

doi:10.1088/2053-2571/aae504ch15

I choose $f/5.6$ as the standard focal ratio because it is a typical largest aperture of a medium-fast large-format lens.

We can relate this definition of slowness to incident metering, via equation (A.27), simply by defining more clearly what we mean by a 'proper exposure.' In ISO-standard photography, a reflectivity of $R = 0.18$ is taken to be the standard. But for VLS photography that may not be the best choice.

Part of the point of VLS photography is that non-photographic content, related to the paper detector surface itself, can be intentionally incorporated into the image. This happens most with what would ordinarily be considered an *under*-exposure, which leads to shadow detail dominated by non-photographic content. Thus I argue that it is more often the case that we want a mid-range of detector density to correspond to a subject reflectivity significantly *greater* than 18%.

But to stay somewhat in line with traditional metering, I propose to use instead a standard of 20.9% reflectivity, or $R = 0.209$. I make this odd choice, instead of an even 20%, because it leads to a simple numerical connection between S and incident-light metering time, t_{SUB}. If we choose $R = 0.209$ and also incorporate our *f*/5.6 standard, while applying a conversion from seconds to minutes, equation (A.27) then becomes:

$$t_D\ (\mathrm{min}) = 10.00\ t_{\mathrm{SUB}}\ (\mathrm{s}) \tag{B.1}$$

This leads to the following connection between incident light metering and my definition of slowness:

> A detector that reaches a mid-range density in t_{SUB} seconds, when illuminated normally by sunny-bright, $EV_{100} = 15$ light, has a slowness given by the following:

$$S = 10\ t_{\mathrm{SUB}}\ (\mathrm{s}) \tag{B.2}$$

We can extend this definition to measurements at other levels of incident light with the following equation:

$$S = \frac{10\ t_{\mathrm{SUB}}\ (\mathrm{s})}{2^{(15-EV_{100})}} \tag{B.3}$$

And so for example, if one finds that under lighting conditions of EV $_{100} = 13$ (cloudy-bright) the VLS detector reaches a mid-range density in 8 s, then we have:

$$S = \frac{10 \times 8}{2^{(15-13)}} = \frac{80}{2^2} = 20 \tag{B.4}$$

Note that this same detector would presumably take only 2 s to reach that same density under sunny-bright lighting conditions of $\mathrm{EV}_{100} = 15$, and equation (B.3) would still give the same answer of $S = 20$. And so for this particular detector, we should expect a proper exposure to take 20 min at $f/5.6$ under sunny-bright lighting.

And so this simple test of our VLS detector to measure *S* then informs us what exposure time, t_D, we should expect, given the focal ratio and EV_{100} lighting conditions, if we simply count steps in the proper direction from both *f*/5.6 and $EV_{100} = 15$. Gathering all of this into one equation, we have:

$$t_D(\min) = S \times (f/5.6)^2 \times 2^{(15-EV_{100})} \tag{B.5}$$

Notice that equation (B.5) means that, at *f*/5.6 and sunny-bright lighting, the exposure time in minutes is simply equal to the detector slowness.

If you disagree with my choice of $R = 0.209$ as a standard, the procedure and equations described here can still be used. Simply make your own choice of what you mean by 'mid-range density' when you take your incident light reading. If your incident-light test density is *darker* than mine, then your application of these equations will result in exposures that are *longer* than mine. To associate a more-than-mid-range density to $R = 0.209$ is equivalent to associating a mid-range density to a reflectivity that is less than $R = 0.209$.

This is not a laboratory test standard, such as ISO. It is a rough standard that would in practice be interpreted by each photographer in their own subjective way. It gets us all in the same book, and in the same chapter—but probably not exactly on the same page.

Appendix C

A practical guide to EP photography

Ephemeral process (EP) photography uses silver–gelatin enlarging paper, intended for making black and white prints in the darkroom from black and white negatives. But instead of exposing the paper to faint light, and then treating it with a chemical developer, the paper is instead allowed to *print out*. And so it turns dark with the action of light itself, with no chemical amplification after the exposure.

This requires orders of magnitude more light than the normal developing-out process, but printing out can be made much more sensitive, temporarily, by brushing onto the paper an *accelerator* immediately before exposure. The paper is then washed and dried, and it returns to its very low sensitive state, and so it can be easily scanned to capture the negative image.

C.1 EP accelerator formula

The chemistry of the accelerating process is described in some 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 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 accelerator, 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:

doi:10.1088/2053-2571/aae504ch16

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. 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.

C.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 papers (RC): 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!

C.3 Preparing the paper

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 using a camera with a removable film holder, as described below, the paper can be loaded into the film holder immediately after the accelerator has been applied and the cover sheet of acetate put in place. This can all be done in an indoor dim room, if there is access to one near the scene being photographed. Or it can be accomplished under a black cloth, while wearing a red headlamp.

C.4 The camera and lens

The basic design of the EP/Cyano camera I present here is that of a 3 × 4 inch format press camera, like a small Crown Graphic press camera from the 1940s. Constructed out of hardwood and hobby plywood, the camera with film back measures $3\frac{1}{2} \times 5\frac{1}{4} \times 7$ inch closed, and weighs less than $1\frac{1}{4}$ pounds. The bellows is folded by hand from opaque paper, and the camera has a focus range of approximately 100–200 mm. The front standard is held in place by a thumb screw inserted through the bottom of a single slot running the length of the bed, and into the bottom-center of the front standard. Since a single screw holds the front standard in place, the front standard has swing; in fact care must be taken if one does *not* want swing. Focus and swing adjustment is achieved simply by loosening this screw, moving the front standard, and then tightening the screw. This same thumb screw is used to hold the hinged bed in place when the camera is closed. A second thumb screw is inserted from the back to lock the bed in place when open. The camera bed opens in the portrait-mode direction. A second tripod mount on one side is used to orient the camera in landscape mode (at which point front swing becomes front tilt).

These features so far described differ from those found in an ordinary press camera only in their relative simplicity, mechanical crudeness, light weight and low cost. One can get away with this here, because the goals of the photographer are different, and precise mechanical alignment and stability are often neither necessary nor a virtue for the EP or cyanonegative photographer. Furthermore, the camera need only be *merely* light tight; it need not be *really most sincerely* light tight.

This particular design is a convenient size; it is small and lightweight, but still large enough to capture a significant amount of detail. Most of the EP pictures in this book were taken with this very camera. The 3 × 4 inch format means that one 8 × 10 sheet of enlarging paper is enough for six exposures, and this translates to about $0.20 per exposure (the cost of the accelerator solution is trivial).

The Graflex Optar 135 mm *f*/4.7 lens shown in figure C.1 is a good choice for this camera. These relatively low-cost but excellent lenses were used for many years on

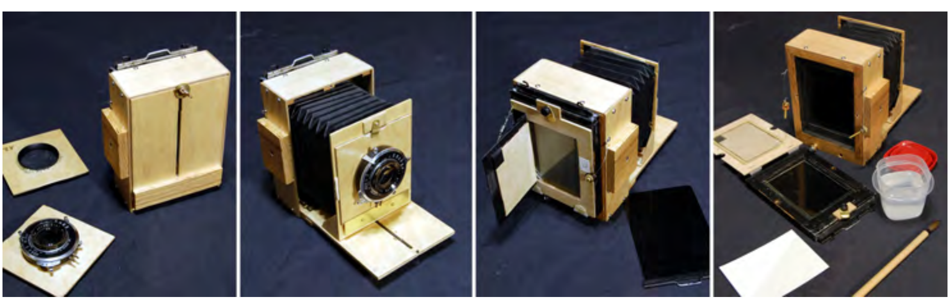

Figure C.1. Cameras and film holders for cyanonegative and EP photography. The cameras are of the simple design described in this article, with paper bellows and crude mechanics. Each 3 × 4 inch format camera weighs (with film holder but without lens) less than $1\frac{1}{4}$ pounds. The film backs are modified from $3\frac{1}{4} \times 4\frac{1}{4}$ film holders. The sensitized paper is sandwiched between the focus screen (mounted on a hinged door) and a sheet of glass.

the popular Speed and Crown Graphic press cameras, and there are many available on the used market, at a typical cost (2018) of less than \$100—much less if the shutter is broken (not a concern for EP photography). But even a simple magnifying glass can produce interesting results. Aperture stops can be made out of black paper; if the magnifying glass is stopped down to *f*/5.6 or *f*/8, daylight exposures are still easily within reach, but the image shows far fewer lens aberrations than if used with its full aperture (typically about *f*/2.5).

C.5 The film back

The heart of the EP/cyano camera is a special combination film back and focus screen that is designed like a frame for contact printing. The example described here is modified from a standard, double-sided $3\frac{1}{4} \times 4\frac{1}{4}$ inch film holder. But this same kind of back can be made from any 4×5 or 8×10 inch film holder, thus converting any view camera into an EP camera. A good choice is the oldest version of the Speed Graphic (sometimes called a 'pre-anniversary' Speed Graphic). It is lighter—and cruder—than later versions, and is usually much less expensive on the used market. The extra features on the later models are not needed for EP photography. If the bellows are damaged (thus lowering the price), a little bit of tape is all that is needed since they do not need to be perfectly light tight for EP photography.

For my modified EP film back, the light sensitive paper is sandwiched between two pieces of glass, thus ensuring the wet paper stays flat during the exposure. The rearward piece of glass is on a hinged door, and it is frosted (it is covered with strips of matte-finish cellophane tape). Thus it becomes the focus screen when no paper is present. A hinged cover is attached to the door, and swings 90° to it. During an exposure this prevents light from getting through the focus screen to the back of the film. During focusing and composition it props up to form a sun shade.

All of the printing frame mechanism is housed on one side of the double sided film holder. The other side still retains its dark slide, which can be used as a shutter. The film can be loaded on site, directly into the film holder while it is attached to the camera (under a dark cloth, for example). Alternatively, the film holder can be removed from the camera after focusing and composing, loaded separately, and then re-attached for the exposure.

Unlike FB paper, RC paper does not curl when wet, and so it is possible to cut such paper to fit in a standard unmodified sheet film holder. The accelerator can then be applied once the paper is in the holder, a protective sheet of acetate placed on top, and the dark slide put in place. And so any large format camera can be used unmodified for EP photography. The paper may still curl a little (it should be cut slightly undersized to give it room to expand), and so possibly affect the focus.

C.6 Washing, drying and scanning

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). Washing should be carried out in dim light, and the paper should be left to dry in total darkness.

Scan the paper only after it is dry. The scanning process will damage the negative, but if done carefully, the image can be captured well. A higher resolution scan exposes the paper more, as does repeated scanning. Most papers (especially RC papers) will suffer more damage from scanning if the humidity is high.

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 any time 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 = 0.004 97 furlongs, etc.

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

doi:10.1088/2053-2571/aae504ch17

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

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 $\mathrm{m\ s^{-1}}$ (called 'meters per second'). What if we want to divide $\mathrm{ms^{-1}}$ by seconds? We can do that just fine, and we get $\mathrm{m/s/s} = \mathrm{m\ s^{-2}}$ (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* too small. And so clearly, when dealing with numbers like this, we need a better way.

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	$\times 10^{12}$

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:

$$9.75 \times 10^{7} = 975 \times 10^{5} = 0.009\,75 \times 10^{10} = 97\,500\,000\,000 \times 10^{-3} \tag{D.6}$$

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.