Image Sensing

Shree K. Nayar

Monograph: FPCV-1-2

Module: Imaging

Series: First Principles of Computer Vision

Computer Science, Columbia University

February 21, 2022

FPCV Channel

FPCV Website

Image Sensing

Shree K. Nayar
Columbia University

Topic: Image Sensing, Module: Imaging
First Principles of Computer Vision

1

Image Sensing

Need to convert Optical Images to Digital Images (*numbers*) for computer representation and use.

Topics:

(1) A Brief History of Imaging
(2) Types of Image Sensors
(3) Resolution, Noise, Dynamic Range
(4) Sensing Color
(5) Camera Response and HDR Imaging
(6) Nature's Image Sensors

2

In this lecture, we discuss how to convert an optical image into a digital image so that a computer vision system can analyze it. We will first give a brief history of imaging, with a timeline of the major inventions that led to the modern digital camera. We argue that the most important invention in the evolution of imaging is that of the image sensor. We will describe two types of image sensors — CCD sensors and CMOS sensors — and study their characteristics, including resolution (the number of pixels in an image), noise (undesirable modifications to the image), and dynamic range (the range of brightness values that a sensor is able to measure). Then, we will discuss how image sensors can be designed to capture color, which, simply put, is the human response to different wavelengths of light.

Next, we will define a camera's response function, which determines how changing the brightness of a point in the scene affects its brightness in the image. We explain why the response function is often non-linear and show how it can be measured. We also introduce the popular technology of high dynamic range (HDR) imaging and discuss how one can use an image sensor with limited dynamic range to capture images with wider dynamic range. Finally, we will take a look at some of the fascinating image sensors that have evolved in nature, not the least of which is the retina of the human eye.

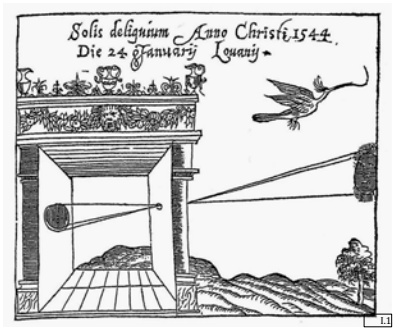
A Brief History of Imaging

Shree K. Nayar
Columbia University

Topic: Image Sensing, Module: Imaging
First Principles of Computer Vision

3

Pinhole Camera



1.1

Camera Obscura

1558

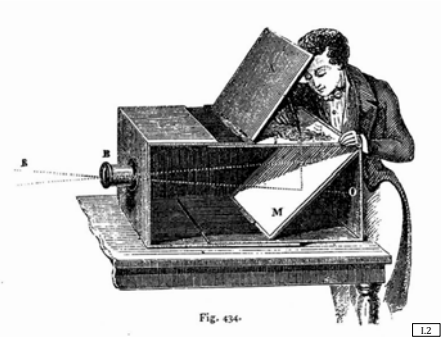
HISTORY

4

The concept of image formation dates back to 500 B.C., when Chinese philosophers were writing about the pinhole camera. Around 1000 A.D., Arabian philosophers and scientists described the pinhole camera and its properties in great detail. It was not until the 16th century though that this idea came to the West and became popular among artists. On the right is a sketch by Gemma Frisius, the Dutch mathematician and philosopher. There is a tiny pinhole in the wall on the right which projects the three-dimensional scene onto the wall on the left to create a two-dimensional image. An artist can simply walk up to the wall on the left and create an accurate sketch of the projection of the scene. The pinhole camera is powerful in terms of the sharpness of the images that it is able to create. However, only a tiny amount of light passes through it, causing the projected image to be very dim.

In the subsequent decades, the pinhole camera was enhanced with the addition of a lens, enabling the creation of brighter images. During this stage of the evolution of the camera, the emphasis was on making the imaging process more convenient for artists. In this 18th century sketch, the lens creates a vertical image which is "folded" using a 45-degree mirror such that the final image is created on a horizontal translucent sheet (tracing paper). This allows the artist to comfortably sit and make a sketch.

Lens Based Camera Obscura



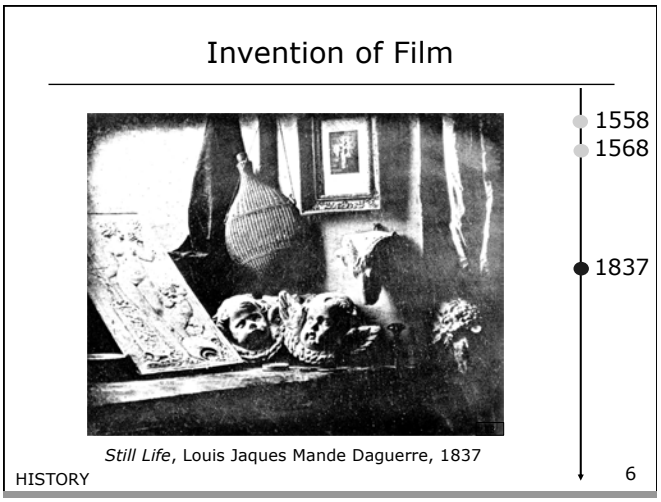
1.2

1558
1568

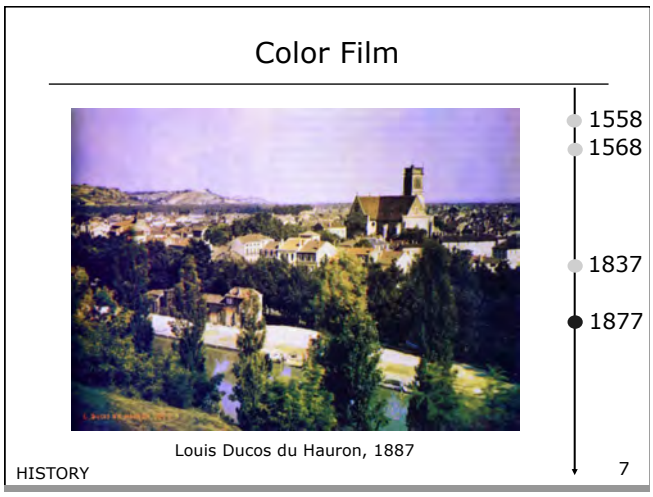
HISTORY

5

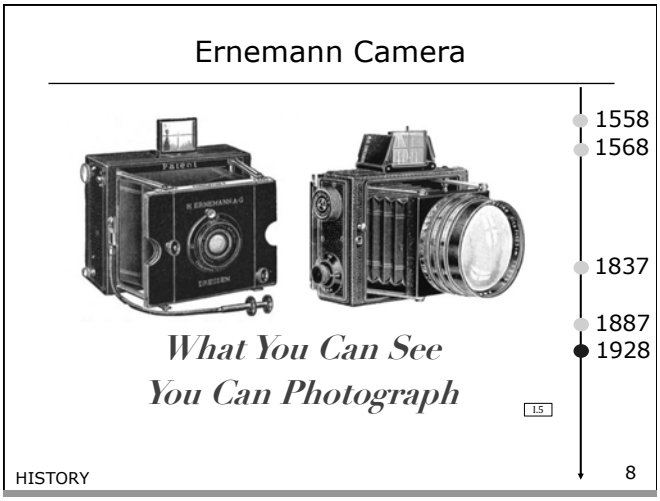
It is often argued that the most important invention in the history of photography was the invention of film. This is a photograph taken in 1837 by Daguerre who co-invented the Daguerreotype camera. For the first time, one could record a moment by simply pressing a button; there was no need for an artist. This was an incredible advancement in terms of its impact on human expression and culture. Typically, film has a layer of silver halide on it and, when exposed to light, the silver halide is converted to metallic silver. The amount of conversion at any given point on the film depends on the exposure of that point, that is, the brightness of light falling on it. Once film is exposed, it is developed using a chemical process, resulting in a photograph.



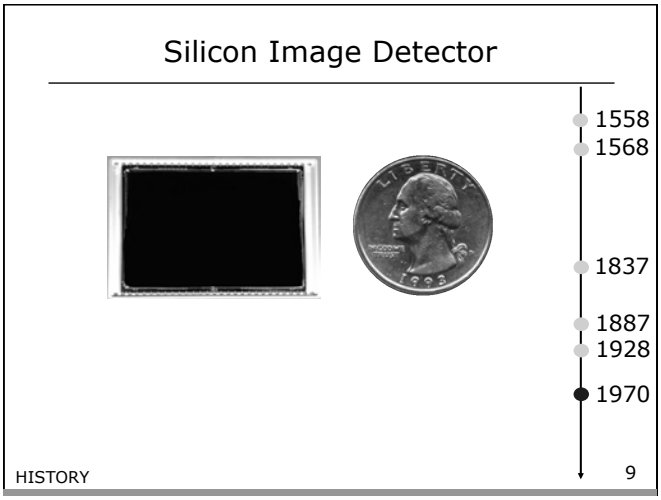
Several decades later, color photography became possible. Here is one of the first color photographs, taken in 1887. The addition of color required major advancements in chemistry — photographers still utilized a silver halide layer but with an additional layer next to it called a dye coupler, which contained various color pigments.



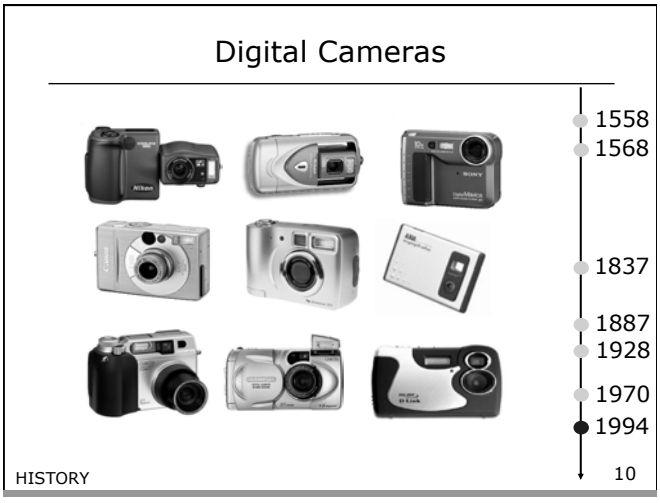
In the 1920s, consumer film cameras became available. Here is the Ernemann camera, with the advertisement proclaiming, "What you can see, you can photograph." This was a remarkable moment — for the first time, it was possible for anyone who could afford it to buy a device that enabled them to memorialize anything they could see with their eyes. Humans were now able to express themselves and communicate with each other visually in a way that was never possible before.



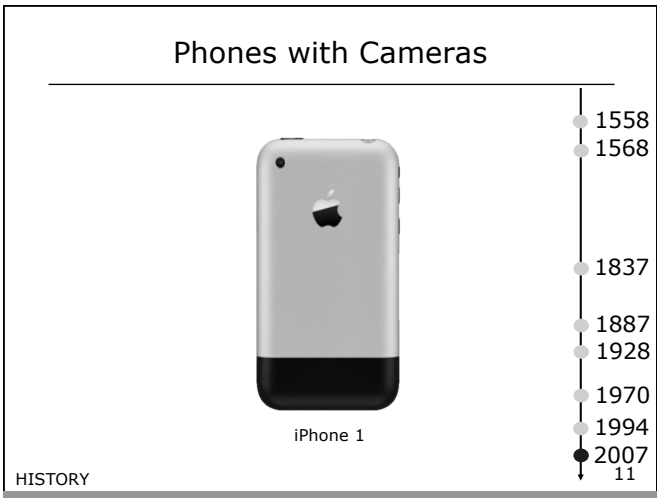
Clearly, the invention of film was a game-changer, and many consider it to be the most important invention in the history of photography. In my opinion, however, even more significant was the invention of the silicon image detector. This chip can do exactly what film can do, except it does not need to be replaced after each image is captured. Its ability to capture an infinite number of images ushered a new era of visual expression and communication.

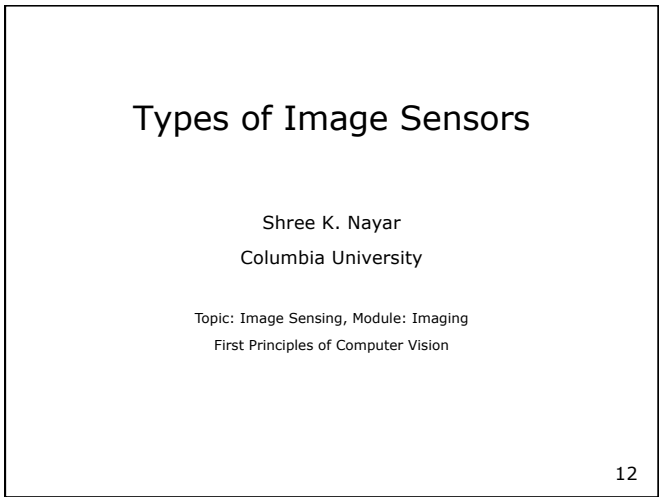


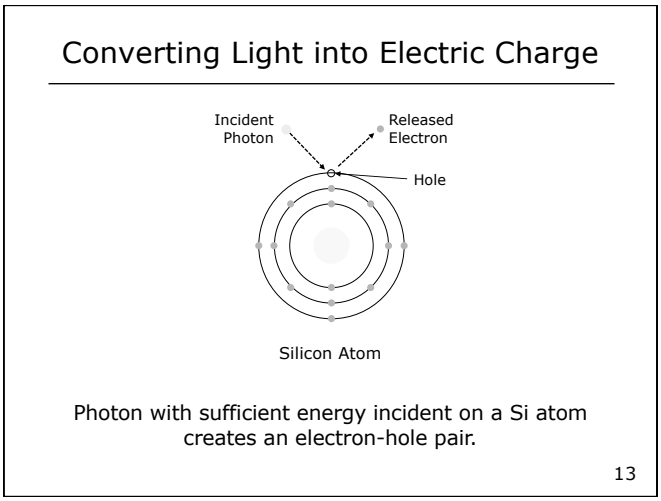
It took about 20 years for the image sensor to mature. When it did, in the early 1990s, there was an explosion of digital cameras in the marketplace. These are some of the first consumer digital cameras. In the top left corner is the Nikon COOLPIX, which was perhaps the most popular. These cameras would typically capture a 640 x 480 image, which was considered high-resolution at that time. They were also very power-hungry; after taking a few dozen photographs, the battery would need to be replaced or recharged.



The demand for image sensor technology surged, driving the industry to innovate at a rapid pace. Somewhere around the end of the 20th century, smartphones with cameras entered the market. That was a major technological milestone, as it required the miniaturization of the camera while simultaneously an improvement in terms of performance. Here you see Apple's iPhone 1, which was released in 2007 and included a tiny camera. The smartphone camera gave rise to social platforms like Snapchat and Instagram, where billions of users communicate visually on a daily basis today. The evolution of the digital camera has also served as a catalyst for the major advances in computer vision and artificial intelligence we have witnessed in the last decade.

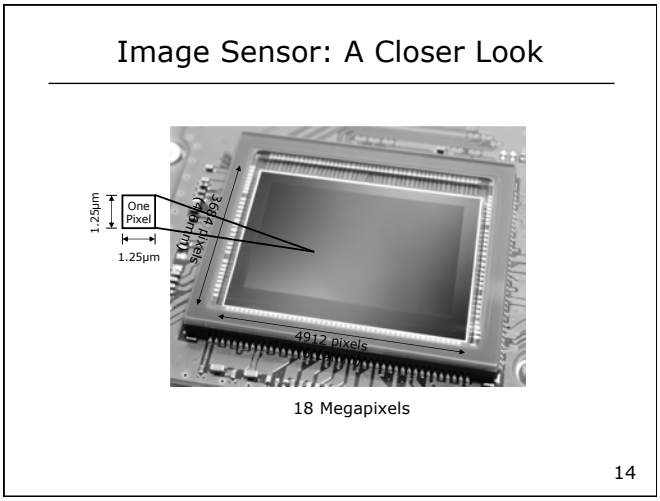




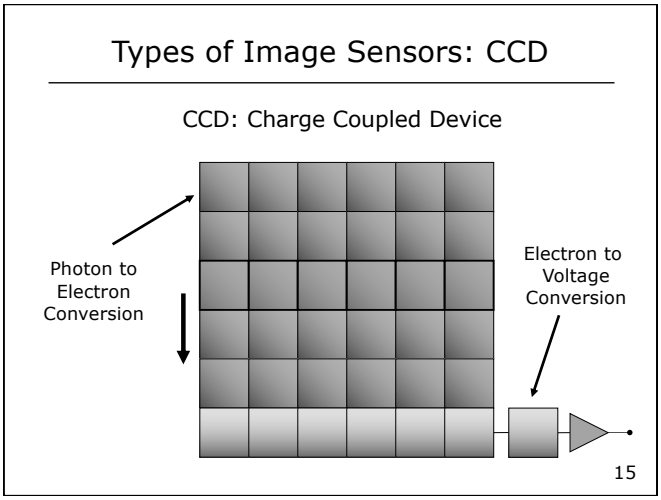


Most image sensors in use today are made of silicon, which has properties that make it well-suited for imaging applications. Shown on the right is a single silicon atom. When it is hit with a photon of sufficient energy, it releases an electron, creating an electron-hole pair. If we have a silicon crystal — a lattice of silicon atoms — and we hit it with light, the photons arriving at it cause the generation of electrons within it. There will be an equilibrium established between the photon flux and the electron flux. If we had a way to measure this electron flux, it would represent the intensity of light falling on the silicon crystal. In short, silicon does most of the work for us when it comes to image sensing — all we need is a method for converting the electrons into voltage that can be measured.

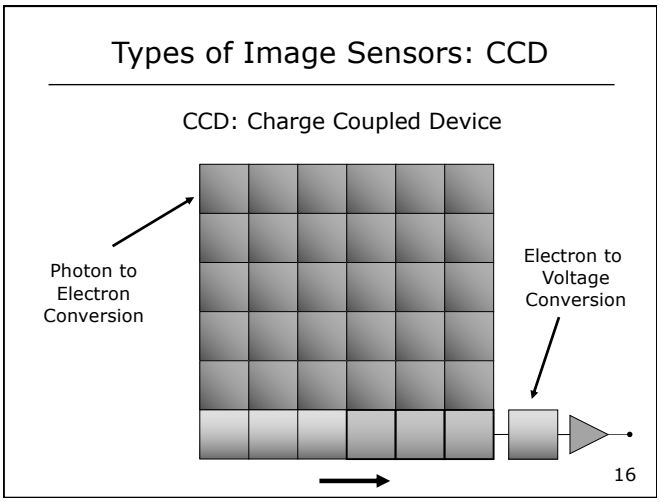
Here is an 18-megapixel image sensor. Each pixel is roughly 1.25 micrometers along each of its two dimensions. Using today's technology, 100 million pixels can easily be packed onto an image sensor. It is worth noting that image resolution does not follow Moore's law, which states that every 18 months, the computational power that can be packed into a unit area of silicon will double. With image sensors, once the pixel size is as small as the wavelength of light — approximately half a micrometer — reducing it further does not increase the true resolution of the image.



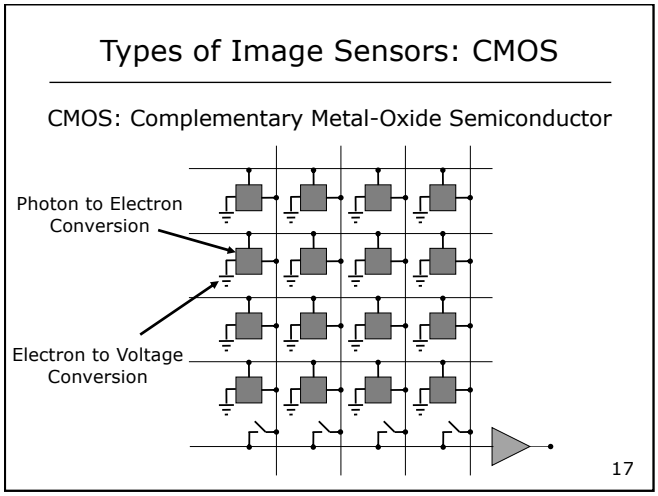
One popular type of image sensor is a Charged Coupled Device, or CCD. Each pixel has a "potential well" in which incoming photons are converted into electrons. That is, photon to electron conversion happens within the pixel. To read out the electron flux values in all of the pixels, a circuit is used that enables each row to pass its collected electrons to the next row, while that row passes its electrons to the next row, and so on.

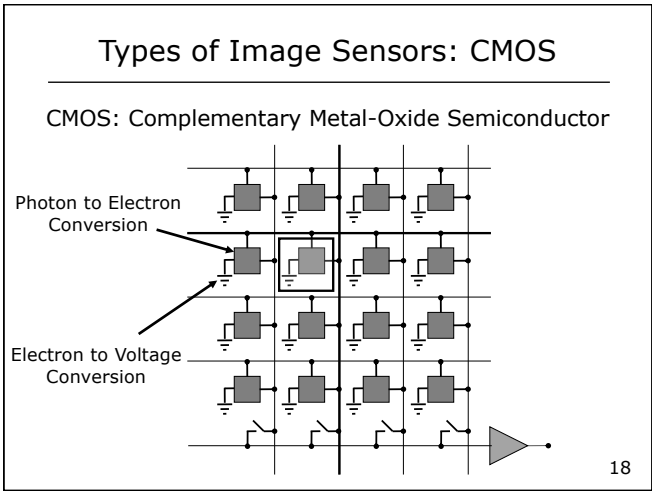


Eventually, the electrons in each row reach the bottom row. In this row, the electrons are shifted horizontally from one pixel to the next. When the electrons reach the last pixel, they are converted to an analog voltage, which is then converted to digital output using an analog-to-digital converter (ADC). The transfer of charges from one row of pixels to the next was a remarkable innovation. This transfer technique is known as "bucket brigade"; it is the same process whereby a string of people would pass buckets of water from one to the next. In the CCD, electric fields are applied to



certain regions underneath the potential wells to shift electrons from one row to the next. This technology is sophisticated as it needs to avoid losing electrons or collecting unwanted ones during the transfer process.

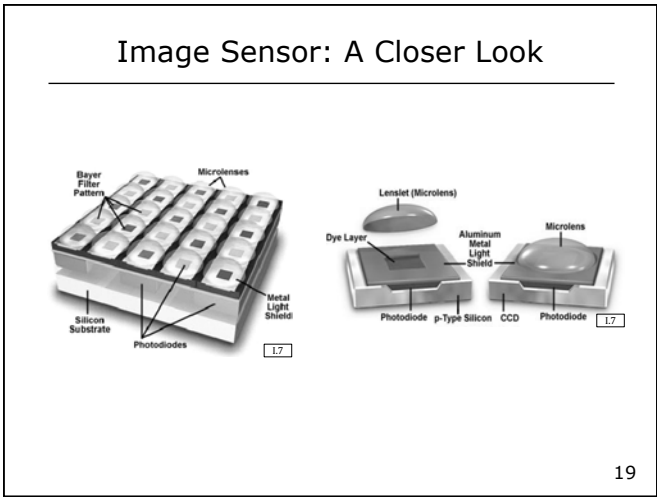




Shown here is another popular type of image sensor called CMOS (Complementary Metal-Oxide Semiconductor). Again, there are potential wells that collect light, but in this case, each pixel includes a circuit to convert electrons to voltage. As a result, voltage can be measured directly from each pixel. If we were interested in only a small region of the image, we could read out just those pixels at a much faster rate as there are fewer of them in the region as compared to the entire image.

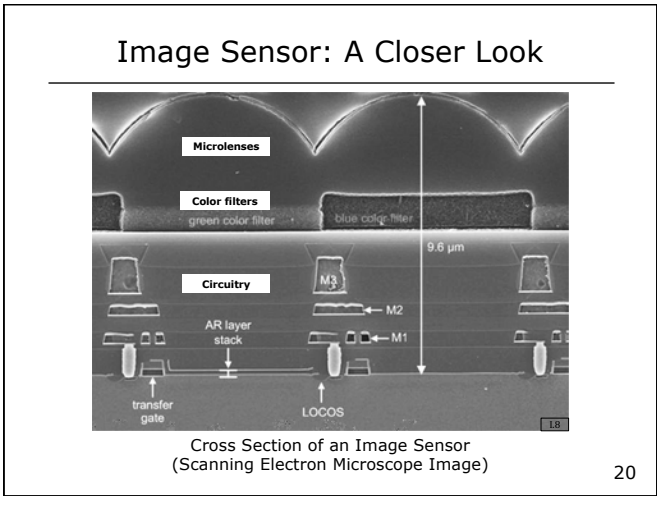
The ability to read voltages directly from pixels makes CMOS technology more flexible than CCD. However, each pixel in a CMOS sensor has a smaller light sensitive area because of the electron-to-voltage conversion circuit sitting next to it. Although CMOS and CCD are both widely used, CMOS sensors are more commonly found in consumer cameras because of their flexibility.

Seen here on the left are the potential wells (referred to as photodiodes) corresponding to pixels. A pixel has no way of determining what color of light is arriving at it; it can only convert photons to electrons. In order to measure color, we situate color filters above the pixels. Red, green and blue filters are typically used. Since only one color can be measured at any given location, each pixel has one color filter placed atop it. After the image is captured, we can take the red, green and blue values, which are distributed over the image, and interpolate them to obtain a red, green and blue measurement at each pixel.



Each pixel also has a lens sitting on top of it called a microlens. This is not the lens that is forming the image; rather, this lens takes light from the image-forming lens and focuses it onto the light sensitive area of the pixel. The reason the light sensitive area is smaller than the size of the pixel is because there is circuitry sitting around the pixel, and we want to ensure that the light falling on the circuit region is not wasted. That is, each microlens takes all of the light falling on the entire pixel area and funnels it down to the light sensitive area.

Here we see a scanning electron microscope (SEM) image of the cross section of an image sensor. We can see the microlenses and the color filters. Underneath each color filter is a potential well. Note that, in this particular sensor, the distance between the top of the microlens and the bottom of the circuitry is only 9.6 micrometers — there is an incredible amount of technology packed into this thin layer of silicon! With time, more and more technology will be integrated within an image sensor. In the future, we can expect to see additional layers of circuitry beneath the image sensor that enable them to perform various image processing and visual processing steps before the image is outputted by the sensor.



Resolution, Noise, Dynamic Range

Shree K. Nayar

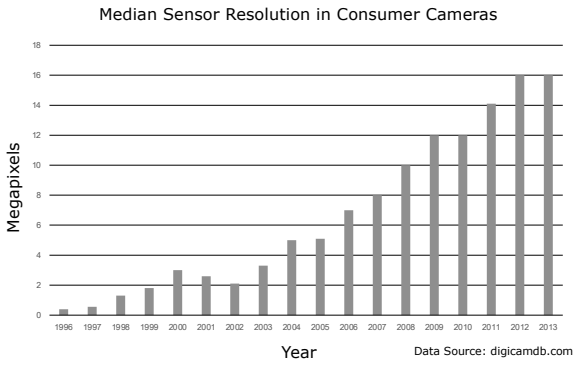
Columbia University

Topic: Image Sensing, Module: Imaging

First Principles of Computer Vision

21

Image Sensor Resolution



22

Now, let us talk about some of the key characteristics of an image sensor. We will start with resolution, which is the number of pixels in the image. In the graph on the right, we see the resolution (in megapixels) plotted as a function of the year, going from 1996 to 2013. The first digital cameras had a resolution of about a quarter of a megapixel, which by today's standard is very low. As seen in the plot, image sensor technology improved at a rapid pace, and by 2013, it was possible to capture a 16-megapixel image with a consumer camera. Today, even smartphone cameras can capture images with that resolution. For most applications, we have more resolution than we need.

Next, let us discuss the important topic of noise. Noise is unwanted modification of a signal during its capture, conversion, processing, transmission, or even storage. Clearly, in virtually all applications, we want to minimize the noise. To do so, we need to first understand the various sources of noise.

Noise in Image Sensors

Noise: Unwanted modification of signal during capture, conversion, transmission, processing.

- Photon Shot Noise (Scene Dependent)
  - Quantum nature of light
  - Random arrival of photons
- Readout Noise (Scene Independent)
  - Electronic Noise: Pre analog-to-digital conversion
  - Quantization Noise: Post analog-to-digital conversion
- Other Sources (Scene Independent)
  - Dark Current Noise: Thermally generated electrons
  - Fixed Pattern Noise: Defective pixels

23

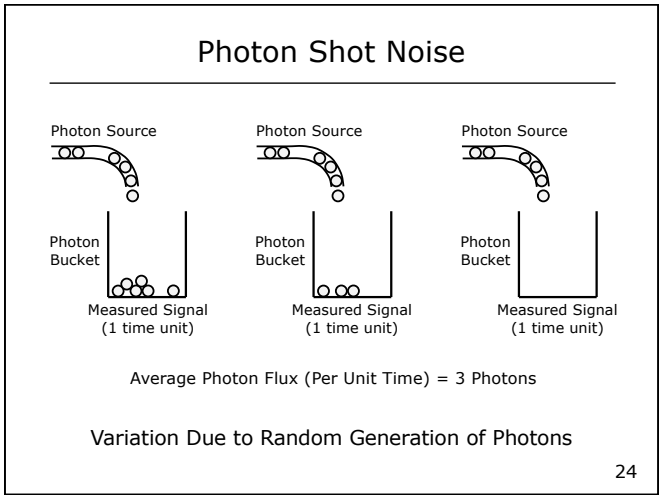
The first type of noise is called photon shot noise. This is due to the quantum nature of light itself, as a result of which photons arrive at the lens of the camera in a random fashion – akin to the arrival of rain drops on the ground. This phenomenon results in noise in the image which is a function of how bright the scene is and has nothing to do with the image sensor itself.

We know that photons that arrive at the image sensor are converted to electrons. These electrons need to be converted into a voltage and, during that process, noise is introduced by the conversion circuit. This is referred to as read (or electronic) noise. Next, the voltage is converted to a number using analog-to-digital (ADC) conversion. Although this conversion is being done intentionally, it can be regarded as

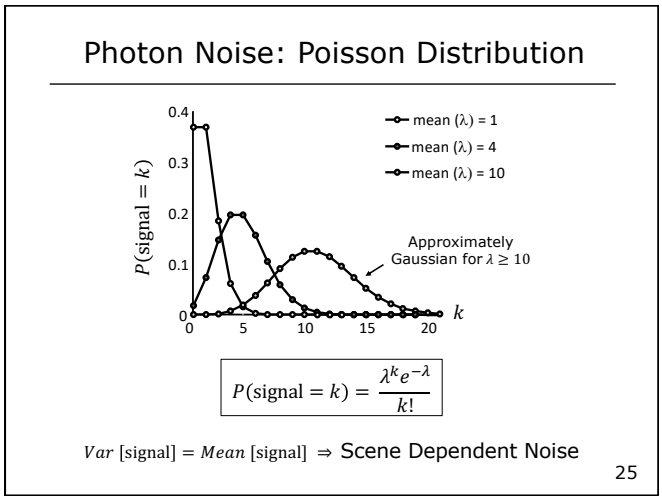
a modification of the signal and is referred to as quantization noise.

Finally, there are other sources of noise. For instance, even if we have a cap on the lens of the camera — that is, there is no light entering the camera — as the temperature of the camera is increased, there could be electrons released by the image sensor. This is referred to as dark current (or thermal) noise. We also have fixed pattern noise, which comes from the fact that no two pixels can be made identical during manufacturing — they can be expected to have slightly different responses to light.

Let us first take a closer look at photon shot noise. Say we wanted to measure the number of photons arriving at a pixel from a point in the scene. The pixel can be viewed as a bucket in which the photons are collected over a fixed time — which is the exposure time or integration time of the image sensor. In the first measurement, we may measure 3 photons. If we take a second measurement, we may get 6 photons and the third time we may get none. This is because of the random nature of the arrival of photons.

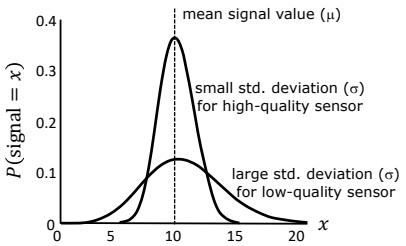


We can model the photon shot noise using the Poisson distribution. This distribution has a mean λ, which is the value we really want to measure. However, we can get other values around the mean, and the probability that we get the value k is given by the Poisson distribution shown here. In the plot, we show three distributions with different means. Note that as the mean value increases, so does the width of the distribution, and for higher mean values the Poisson distribution begins to look like a Gaussian distribution. The Poisson distribution has the interesting property that its variance is equal to its mean. The photon shot noise is clearly scene dependent as the mean value of the distribution corresponds to the brightness of the scene point. As the brightness increases, the mean increases, and so does the variance of the distribution.



Now, let us talk about read noise, which is introduced during the conversion of electrons to a voltage. This noise is often modeled as a Gaussian distribution, which has a mean and a standard deviation. Once again, the mean is the value we are interested in measuring but, due to read noise, we end up with the value *x,* the probability of which is given by the Gaussian distribution. If the image sensor is of high quality, the distribution is going to be narrow, while for low quality (noisier) sensors it is wider. In short, read noise is entirely dependent on the quality of the sensor and is independent of the brightness of the scene point.

Read Noise: Gaussian Distribution



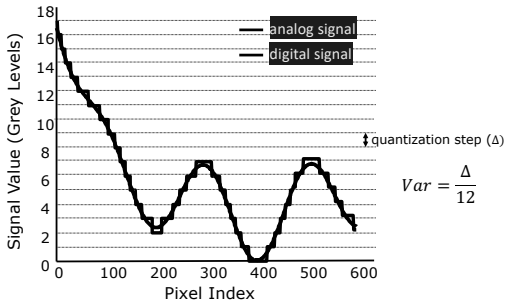
$$P(\text{signal} = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Depends on Sensor Quality (Scene Independent)

26

Next, the measured analog voltage is converted to an integer value using an analog-to-digital convertor (ADC). The end result is a discrete value that could differ slightly from the voltage. Although it is introduced on purpose, this can be regarded as a form of noise and is called quantization noise. If the step, or gap, between two consecutive discrete values is delta, it is simple to show that the variance of this noise is delta divided by 12. In present day cameras, which typically output 10 or 12 bits of brightness resolution, quantization noise tends to be small.

Quantization Noise



$$Var = \frac{\Delta}{12}$$

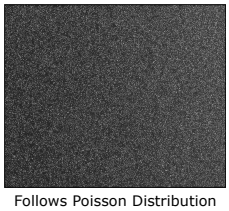
Negligible in Modern Sensors
Due to High Intensity Resolution (12-14 bits)

27

In the case of thermal noise, electrons are generated within the image sensor due to its temperature, even when there is no light arriving at the sensor. This noise is very low and is only an issue when the exposure time of the sensor is very long, which is the case in extreme low-light imaging applications such as astronomy or night time photography. For this reason, in applications like astronomy, where exposure times can be several minutes, the sensor is kept cool at a given temperature to minimize thermal noise.

Other Noise Sources

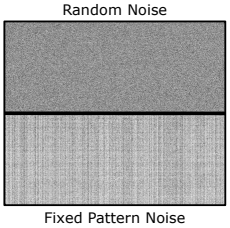
Dark Current Noise
(Thermal)



Follows Poisson Distribution

Significant Only for Long (>2 min) Exposures (Astronomy)

Fixed Pattern Noise
(Defective Pixels)



Can be Reduced by
Dark Frame Subtraction

28

Finally, there is fixed pattern noise, which is due to imprecisions inherent to the manufacturing of image sensors. No two pixels on an image sensor are going to be identical — there will be some (small) variation between the efficiencies of the millions of pixels on the sensor. In the image on the right, the bottom half shows the effect of fixed pattern noise. Even though this is the image of a scene with uniform brightness, there are differences in the brightness values measured by the pixels. These differences are extremely small and are magnified here for visualization purposes. For comparison, we show random noise in the top half of the image. Fortunately, fixed pattern noise can be compensated for by calibrating the response (or gain) of each pixel and then normalizing all subsequent measurements made by that pixel using the gain.

The last characteristic we discuss is dynamic range, which is defined as 20 times the log of the ratio of the maximum possible photon energy the pixel can measure to the minimum photon energy the pixel can detect. The maximum photon energy is determined by the potential well of the pixel. There exists a photon energy that fills up the potential well, and energy levels higher than that cannot be measured — they produce the same (maximum) electron count. The minimum detectable energy is the electron count that is detectable in the presence of noise. Note that if the signal we are trying to measure is weaker than the noise, it is not distinguishable from the noise and hence is not detectable. The unit of dynamic range is decibels.

Sensor Dynamic Range

$$\text{Dynamic Range} = 20 \log\left(\frac{B_{max}}{B_{min}}\right) \text{ decibels (dB)}$$

B_{max}: The maximum possible photon energy (full potential well)

B_{min}: The minimum detectable photon energy (in the presence of noise)

Sensor	$B_{max}:B_{min}$	dB
Human Eye	1,000,000:1	120
HDR Display	200,000:1	106
Digital Camera	4096:1	72.2
Film Camera	2948:1	66.2
Digital Video	45:1	33.1

29

Shown in the table are the dynamic ranges of some commonly used imaging systems, including the human eye. It should be noted that these are rough estimates. The human eye has a remarkable dynamic range of roughly 1 million to 1, which corresponds to 120 decibels. We see that a typical consumer digital camera today has a range of about 72 decibels, which is higher than that of film. Note that a video camera has a lower dynamic range than a still camera, since a video camera must capture images in quick succession and hence use low exposure times. As a result, the photon energy levels received by a video camera are relatively low while read noise remains the same. In short, video cameras are forced to have lower signal-to-noise ratios and hence lower dynamic ranges.

Sensing Color

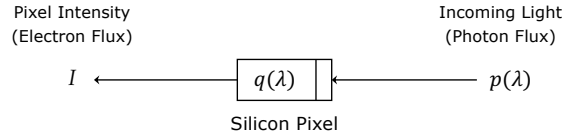
Shree K. Nayar
Columbia University

Topic: Image Sensing, Module: Imaging
First Principles of Computer Vision

30

Quantum Efficiency

Incoming light can vary in Wavelength (λ)



Quantum Efficiency:

$$q(\lambda) = \frac{\text{Electron Flux Generated}}{\text{Photon Flux of wavelength } \lambda}$$

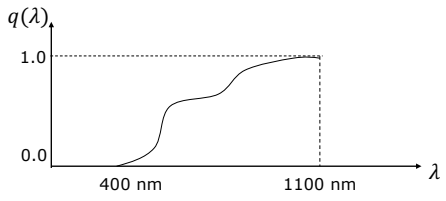
31

Now let us examine what it means to measure the color of a scene point. We can denote the photon flux from a point in the scene as *p*(λ), indicating that it is a function of wavelength λ. The photon flux arrives at a pixel, where it is converted to electron flux, *I*. The quantum efficiency of the material (silicon, for example) utilized to make this conversion is defined as the ratio of the electron flux to the photon flux. Since it is also a function of wavelength, we can denote quantum efficiency as *q*(λ).

We are especially interested in the quantum efficiency of silicon. The visible light spectrum — the range of wavelengths that are visible to the human eye — lies between 400 and 700 nanometers. At higher wavelengths, such as around 1,000 nanometers, silicon's quantum efficiency is 1, which means that every photon received is converted into an electron. However, as the wavelength decreases, the quantum efficiency does as well. Around 400 nanometers, it drops to almost zero. Thus, silicon is virtually transparent for wavelengths above 1,000 nanometers, and it becomes nearly opaque for wavelengths below 400 nanometers.

Quantum Efficiency of Silicon

Silicon is:
- Transparent for $\lambda > 1100\,nm$
- Opaque for $\lambda < 400\,nm$



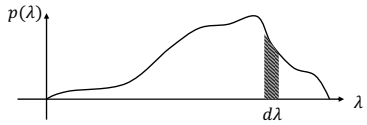
32

Now, consider a single wavelength of light, that is, $\lambda = \lambda_i$. This is called monochromatic light. From the definition of quantum efficiency, we know that the electron flux is equal to this expression [1]. However, the light arriving at the pixel from the scene point typically includes photon flux corresponding to a range of wavelengths. We refer to this as the spectral distribution *p*(λ) of the scene point. How do we determine the electron flux *I* due to a given *p*(λ)? Let us look at an infinitesimally narrow band of wavelengths from λ to λ + *d*λ. In this case, the flux arriving at the pixel is *p*(λ)*d*(λ). Therefore, the electron flux generated by the pixel for the entire spectral distribution is given by this integral [2] .This represents the total number of electrons generated due to the incoming light *p*(λ).

Assume Monochromatic Light $\lambda = \lambda_i$ with flux $p(\lambda_i)$:

$$I = q(\lambda_i)p(\lambda_i) \quad [1]$$

What if incoming light has Spectral Distribution $p(\lambda)$?



Flux (Energy) of Light with wavelength between λ and $\lambda + d\lambda$ is $p(\lambda)d\lambda$. Therefore:

$$I = \int_0^\infty q(\lambda)p(\lambda)d\lambda \quad [2]$$

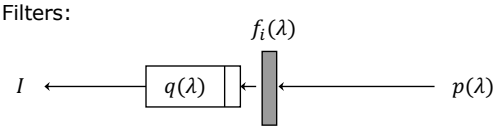
33

Given the electron flux and the quantum efficiency, can we find the spectral distribution *p*(λ)? We cannot, because there are many *p*(λ)s that can be multiplied with *q*(λ) and integrated over all λs to obtain the same *I*. To measure *p*(λ), we utilize filters placed in front of the pixel, where each filter *i* has a response $f_i(\lambda)$. Consider a filter with a response that is a delta function centered at λ_i. Note that a delta function is infinitesimally thin and infinitely tall with an area equal to 1.

$$I = \int_0^\infty q(\lambda)p(\lambda)d\lambda$$

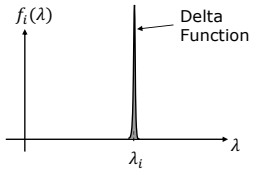
If we know I and $q(\lambda)$, can we find $p(\lambda)$?

Use Filters:



Let $f_i(\lambda) = \delta(\lambda - \lambda_i)$

Note: $\int_0^\infty \delta(\lambda - \lambda_i)d\lambda = 1$



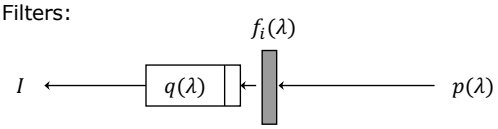
34

Now we have that *I* is equal to this integral [1] which gives us $I = q(\lambda_i)p(\lambda_i)$. That is, a filter with a narrow response centered at λ_i can be used to measure (or "pick out") the value $p(\lambda_i)$. So the question is, how many such filters do we need to recover all of *p*(λ) ? It would seem that we need an infinite number of filters. In practice, however, since *p*(λ) is almost certain to be a smooth function, it turns out that a finite number of filters suffice. We will discuss the reasons for this in our lectures on image processing.

$$I = \int_0^\infty q(\lambda)p(\lambda)d\lambda$$

If we know I and $q(\lambda)$, can we find $p(\lambda)$?

Use Filters:



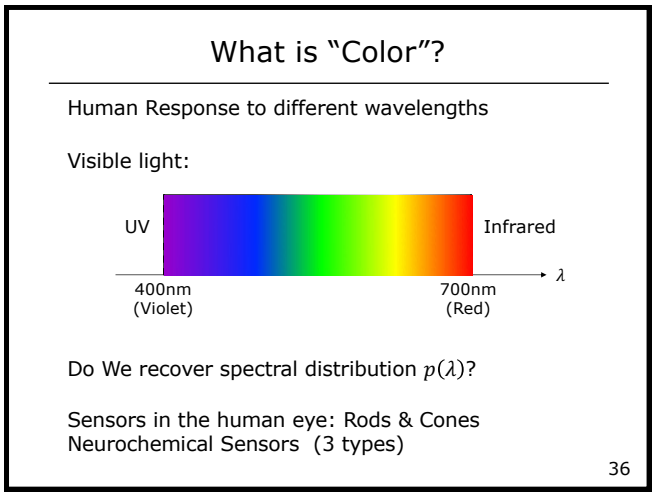
$$I = \int_0^\infty q(\lambda)p(\lambda)f_i(\lambda)d\lambda = \int_0^\infty q(\lambda)p(\lambda)\delta(\lambda - \lambda_i)d\lambda \quad [1]$$

$$I = q(\lambda_i)p(\lambda_i)$$

How many filters do we need to recover $p(\lambda)$? Infinite?

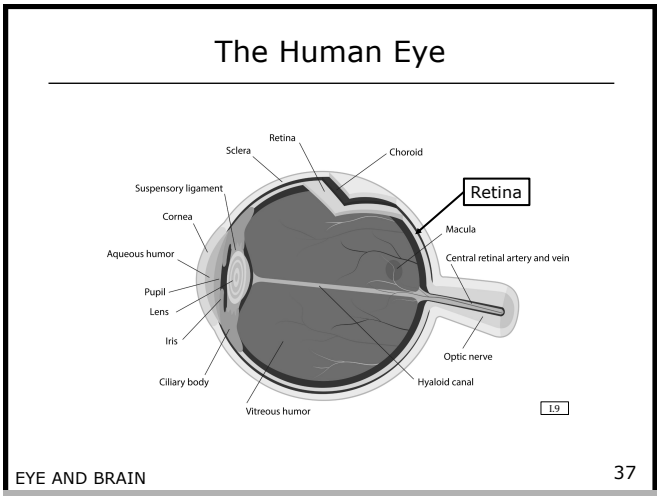
35

That brings us to the topic of color. What really is color? It turns out that color is not a physical quantity that you measure — rather, it is the human response to different wavelengths of light. We cannot see anything lower than 400 nanometers, which is ultraviolet light, or above 700 nanometers, which is infrared light. If we design cameras that can measure information outside the visible light spectrum, we can indeed develop computer vision systems that can perceive things that we cannot.

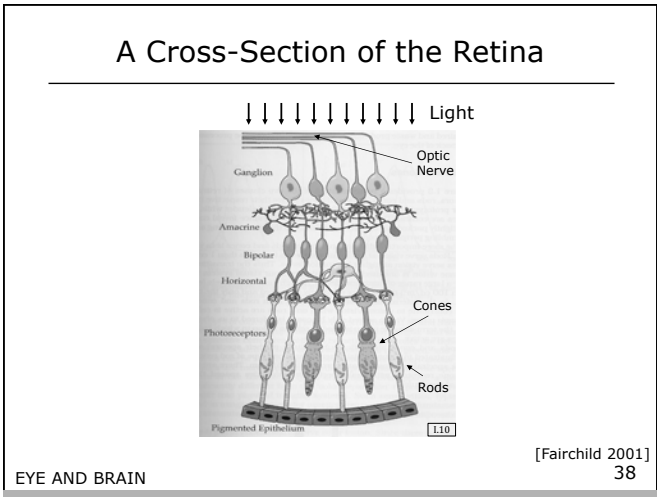


How do we, humans, measure spectral distributions that lie within the visible light spectrum? Our image sensor — the retina — utilizes two types of pixels: rods and cones. The rods are not sensitive to the color of light, but rather sense the brightness of the incoming light. Cones, on the other hand, are sensitive to color, and there are three types of cones. These are neurochemical sensors that respond to three types of color.

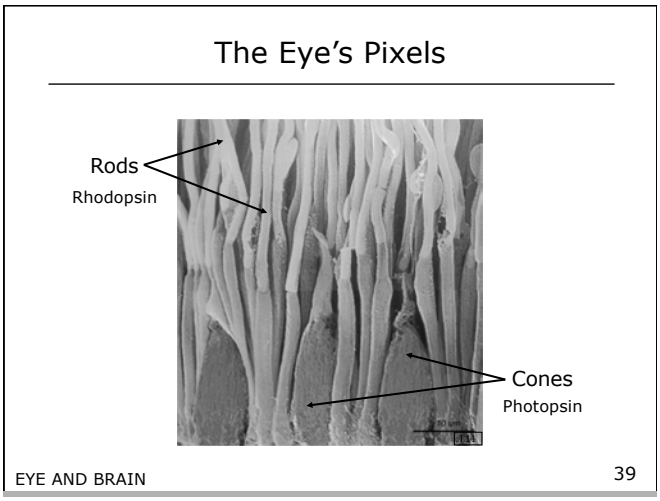
Let us take a closer look at the retina. Here is the anatomy of the human eye, which we reviewed in the lecture on image formation. The cornea acts as a lens and works with the inner lens of the eye to form an image on the retina. Unlike the planar image sensors found in cameras, the retina is curved.



In this cross-sectional view of the retina, we can see the rods and cones. Measurements made by the rods and cones are passed on to bipolar cells and then to ganglion cells. Using these cells, the retina does some early visual processing. That semi-processed image is passed through the optic nerve to the visual cortex in the brain. Which direction is light coming from in this diagram? One might assume that since the pixels (rods and cones) are sitting towards the bottom, light should be coming from the bottom as well. However, in a strange quirk of nature, light actually comes from the top and passes through the ganglion and bipolar cells before reaching the rods and cones.



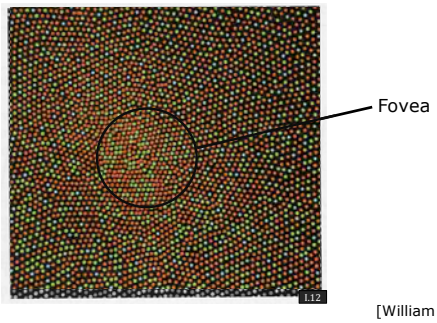
Here is a scanning electron microscope (SEM) image of rods and cones in a real retina. The rods actually look like rods, more or less cylindrical, and the cones are more conical. Both are neurochemical sensors that take in light and generate impulses that represent the intensity of that light. However, they have different proteins — rods have rhodopsin and cones have photopsin. Rods are able to measure black and white images — they are most useful in a dark environment. You may have noticed that in moonlight, you cannot discern the colors of things; rather, you perceive a very dim, colorless scene. When there is enough light, the cones are able to discern color. Vision using rods is called scotopic vision, and vision using cones is called photopic vision. These are two different "modes" that the eye operates in. Irrespective of the mode, the end result is that, after early processing by the cells in the eye, the impulses generated by our pixels (rods and cones) are sent to the brain for higher levels of visual processing



Shown here is the spatial distribution of cones on the retina. There are three types of cones – red, blue, and green – roughly corresponding to the type of light they respond to. You can see that the cones are most dense in the fovea, which corresponds to the part of the eye's field of view that has maximum acuity, or "sharpness." When you look at something in a scene, the image of that thing falls on the fovea — everything else is considered to be in your peripheral vision.

Distribution of Cones in Human Retina

Three types of cones for sensing red, green, blue



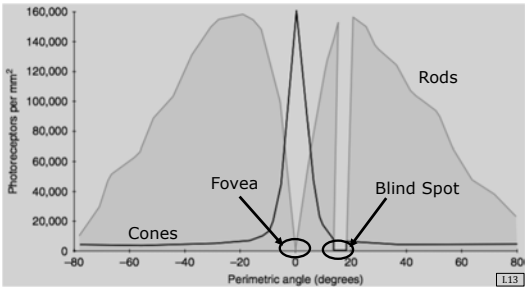
[William 1991]

EYE AND BRAIN 40

The retina has roughly 120 million rods compared to only about 7 million cones. There are very few rods in the center of the fovea; moving outwards, the number of rods rapidly increases in density and then begins to drop off. Note that there is a spot where there are no rods and cones, which is the blind spot. That is where the image is transmitted along the optic nerve to the brain. Our brain fills in visual information that is missing in the blind spot, creating the illusion of a continuous image.

Resolution of Rods and Cones

On average, ~120M rods and ~7M cones per retina.



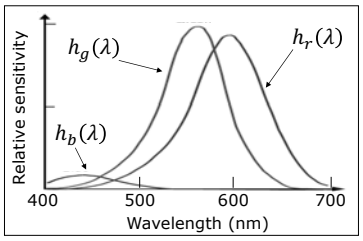
[Fairchild 2001]

EYE AND BRAIN 41

Now let us discuss the spectral responses of the cones. Since there are three types of cones, we have three different spectral responses. They measure light that corresponds to the sensations of red, green, and blue. These spectral responses are shown here and are called tristimulus curves. They are essentially the quantum efficiencies of the red, green, and blue cones.

Spectral Responses of Cones

Three Types: Different Spectral Responses



Tristimulus Curves

$h_r(\lambda)$, $h_g(\lambda)$, $h_b(\lambda)$ are near Red, Green, Blue regions of spectrum.

42

As in the case of our silicon pixel, let $p(\lambda)$ be the spectral distribution of the light falling on the retina. Using the tristimulus curves, we can obtain these three expressions for the outputs of the red, green, and blue cones. These values — *R, G,* and *B* — are called tristimulus values. The eye does not measure the complete spectral distribution $p(\lambda)$, but, rather, we have three numbers corresponding to any incoming spectral distribution.

Tristimulus Values

Three Intensities (R, G, B):

$$R = \int_{-\infty}^{\infty} h_r(\lambda)p(\lambda)d\lambda$$

$$G = \int_{-\infty}^{\infty} h_g(\lambda)p(\lambda)d\lambda$$

$$B = \int_{-\infty}^{\infty} h_b(\lambda)p(\lambda)d\lambda$$

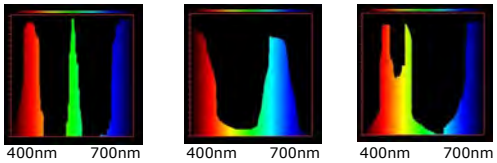
43

Since the eye only produces three values (*R,G,B*) for any given spectral distribution $p(\lambda)$, there is an entire continuum of distributions that generate the same three values. Such a class of indistinguishable distributions are called metamers. Here we can see different $p(\lambda)$s, that produce the same *R,G,B* values — 115,60, and 108, respectively. The color that we perceive in this case is a shade of purple. The existence of these metamers reveals that there are many distinctly different spectral distributions that we as humans perceive to be the same color.

Metamers

Metamers: Different $p(\lambda)$ that produce same (R, G, B)

For example, different spectra:



Same perceived color:

(R, G, B)=(115,60,108)=

44

This brings us to Young's experiment. Young found that just three wavelengths of light can be mixed to produce the sensations of virtually all the colors we are capable of sensing. He used a different projector for each of the three wavelengths. Note that where the projected circles overlap, a color is produced that is different from the original three projected colors. By simply mixing different intensities of three wavelengths, we can reproduce almost the entire gamut of colors that humans can perceive. These wavelengths are 650, 530, and 410 nanometers, and they generate sensations

The Mixing of Colors

Young's Experiment on Color Mixture

Human Sensation of nearly all colors can be produced using 3 wavelengths!

$(\lambda_r, \lambda_g, \lambda_b) = (650,530,410)nm$

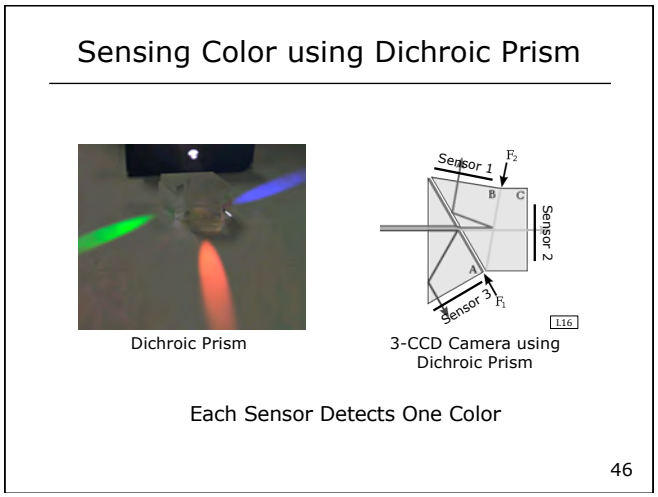
Hence, cameras and displays often use 3 filters:

(red, green, blue)

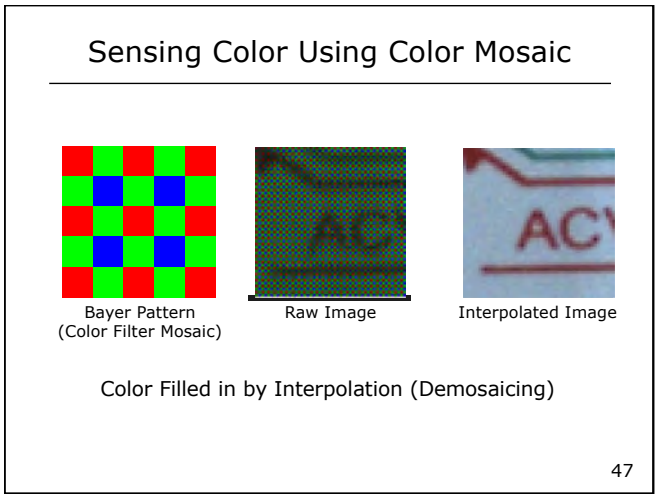
45

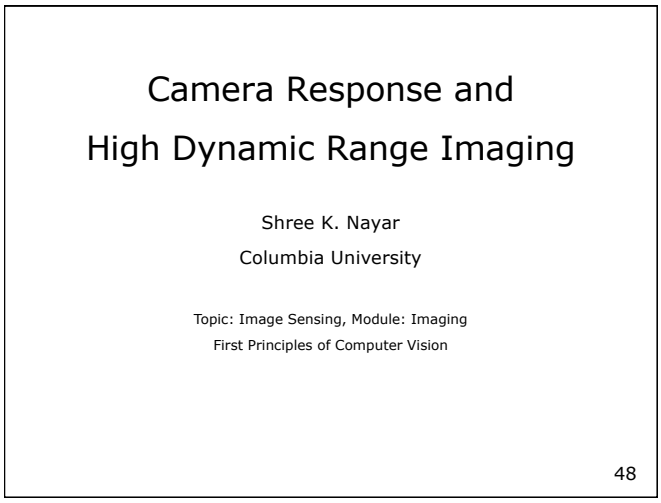
corresponding to the colors red, green, and blue. One does not need to use exactly these three wavelengths — they can be varied somewhat and yet produce a similar effect. It is Young's discovery that enables us to use just three filters – red, green, and blue — to implement color cameras and displays.

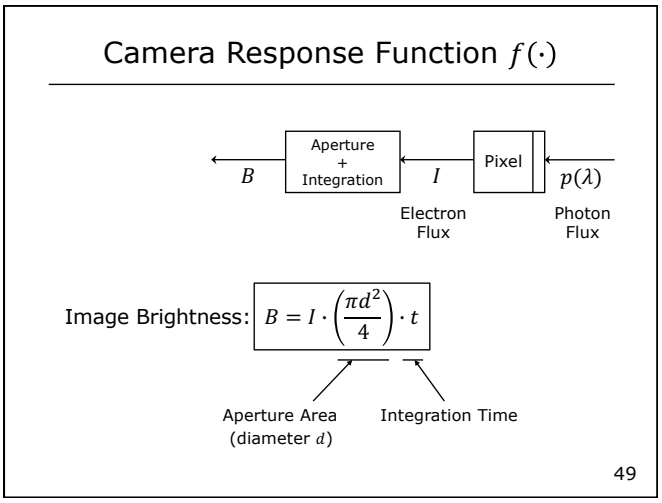
In digital cameras, color images can be captured using an optical element known as a dichroic prism. This prism is a fairly sophisticated piece of optics. When shown an image, it splits the image into three components: a reddish image, a greenish image, and a bluish image. As shown on the right, an image sensor can be placed on each of the three flat faces of the dichroic prism to capture a red image, green image, and blue image. When we stack these three images, we have red, green, and blue values at each pixel, which is referred to as a color image. The disadvantage of using a dichroic prism is that the entire imaging system tends to be bulky and requires precise alignment between the prism and the three image sensors.



Here is an alternative, more popular approach to capturing color images. In this case, a single image sensor is used where each pixel has one of three color filters in front of it. In other words, the image is measured by the image sensor through a mosaic of color filters, as shown on the left. Unlike with a dichroic prism, each pixel here only measures one color. The resulting measured image appears like the one shown in the middle and is referred to as a raw image. In this image, a pixel that measures red light, for instance, does not have the corresponding green and blue measurements. However, since its neighbors do make green and blue measurements, one can estimate the missing green and blue values of the pixel by interpolating the green and blue values provided by its neighbors. The end result is a full color image where every pixel has red, green, and blue values. This process of going from the mosaiced raw image to a full color image is called demosaicing.

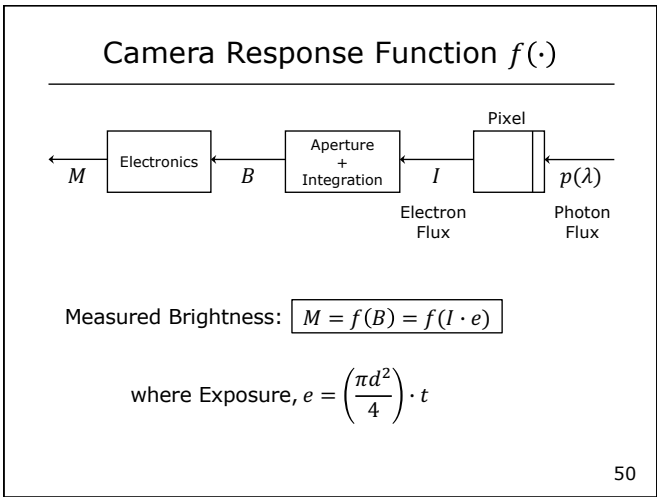






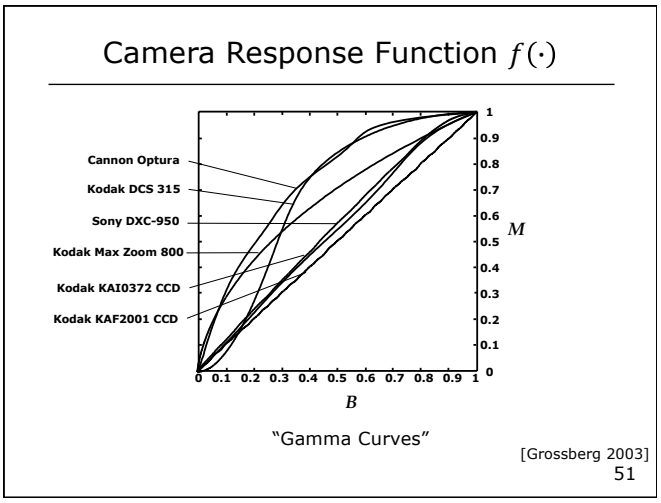
A point in the scene with a certain brightness will produce a corresponding brightness value in the image. The relationship between scene brightness and image brightness is referred to as the camera response function. While the response function is always monotonic it is not necessarily linear. Let us take a look at how we can determine this function for any given camera.

On the right, we see photon flux with spectral distribution *p*(λ) entering the pixel, which produces electron flux *I*. That, however, is not the value measured by the camera, as the electron flux will be modulated by the aperture of the lens and the integration time of the sensor (which will be short for videos, but can be longer for photographs). The image brightness *B* is therefore *I* times the "exposure," which is the product of the area of the aperture and the integration time of the image sensor. Note that the relationship between that brightness *B* and the photon flux *I* is linear.

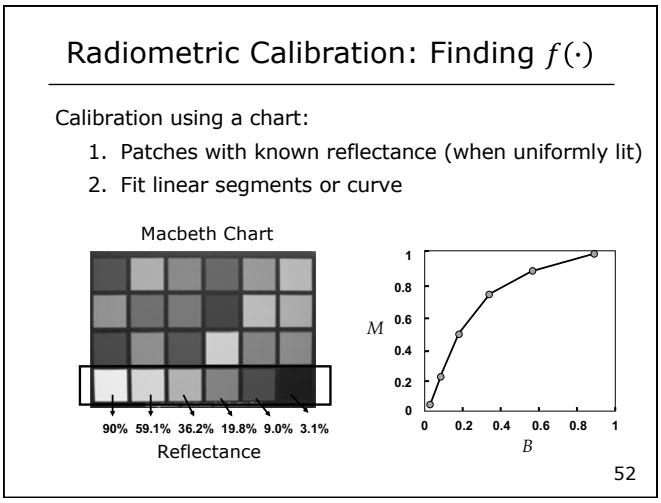


The image brightness *B* usually goes through a series of steps before being outputted by the camera. These steps include electron-to-voltage conversion, analog-to-digital conversion, as well as other image processing steps such as demosaicing. In addition, a non-linear mapping is often intentionally introduced by camera manufacturers so that the camera can measure a wider range of brightness values. Any camera has a finite dynamic range, and a wider range of scene brightness values can be mapped to this dynamic range by "compressing" some brightness values more than others. As a result, the relationship between the final output of the camera *M* and the image brightness *B* is almost always a non-linear function *f*. This is referred to as the camera's response function.

Shown here are the camera response functions corresponding to a few consumer cameras. These functions are sometimes referred to as gamma curves. In general, the response function of a camera is unknown. For some applications of computer vision, it must be measured.

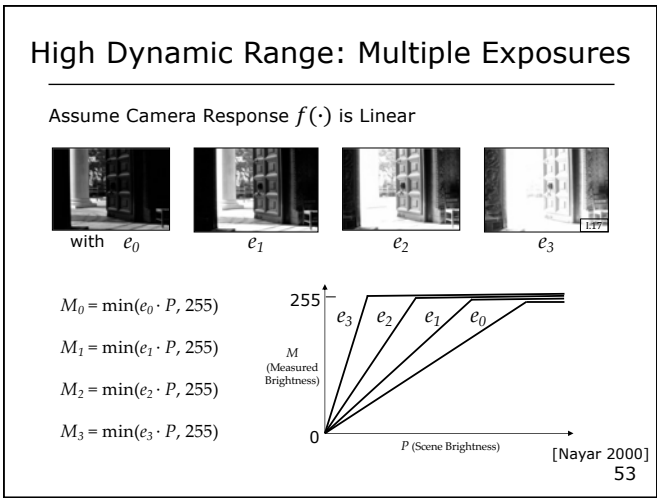


In order to find the response function of a camera, we can use a calibration chart. On the left is a Macbeth chart. Let us focus on the gray patches in the bottom row of the chart. We know the reflectance values of these patches — 3.1% for the dark one all the way to the right and 90% for the bright one all the way to the left. When this chart is lit by one or more sources that are distant from it, each point on the chart can be assumed to receive the same illumination. That is, the whole chart is uniformly lit. In this case, the brightness values of the patches in the bottom row will equal their exact reflectance values multiplied by the same scale factor. This unknown scale factor depends on various factors such as the brightness of the light sources, the gain of the camera, etc. However, irrespective of the scale, the ratios of the brightness values of the patches must equal the ratios of their (known) reflectance values. This allows us to take a single picture of the Macbeth chart using the camera and plot the relationship between the image brightness *B* and the measured brightness *M*. To remove the effect of the unknown scale factor, we can normalize all the brightness values such that the brightest value (produced by the left most patch) equals 1. This plot can be used as the response function of the camera. This process of determining the camera response function is called radiometric calibration.



Now, given any pixel value produced by the camera, we can use the response function to determine the corresponding image brightness *B*. That, in turn, is the true brightness of the corresponding scene point multiplied by the unknown scale factor. That is, we are able to determine the brightness values of all scene points up to a single scale factor. In other words, the response function, once calibrated, can be used to "linearize" the camera.

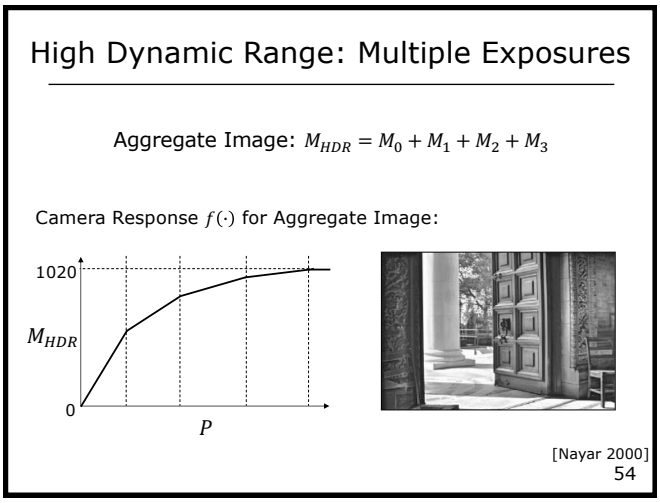
Next we describe the concept of high dynamic range (HDR) imaging. Any camera, no matter how sophisticated, will have a limited dynamic range, which is the range of brightness values it can measure. Remember that the definition of dynamic range is the range from the maximum measurable brightness to the minimum detectable brightness. Needless to say, the real world has an enormous range of brightness values. There is no camera that can capture details of both a bright sky and a dark shadow in the same image.



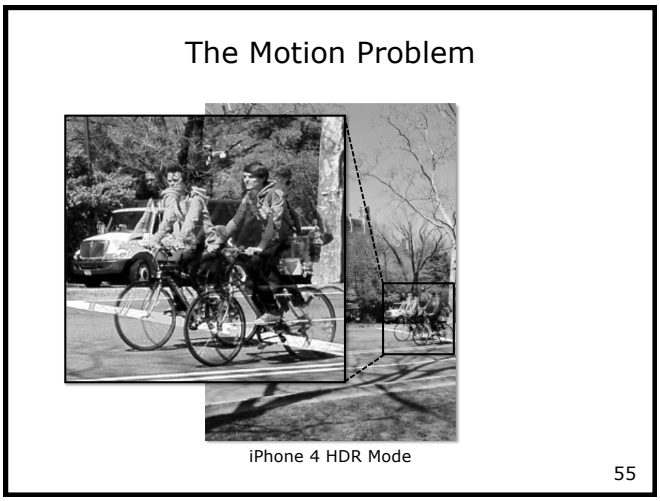
How can we enhance the dynamic range of a camera? Let us assume that the response function is known and hence the camera can be linearized. Thus, the response function is a straight line. Let us say the camera produces 8 bits of information at each pixel. The maximum value it can produce, which corresponds to the full-well capacity of the pixel, is 255. That is, the image "saturates" at 255 — there exists a scene brightness beyond which all brightness values will produce an image brightness of 255.

Imagine that we wish to capture a scene with a very wide range of brightness values. We can first take an image using a very short exposure e_0. If scene brightness is represented by *P,* the measured brightness will be the minimum of $e_0 P$ and 255, as it cannot exceed 255. The corresponding response function is seen in the bottom right, and the corresponding image is the left most one shown on the top. As expected, the bright scene regions are captured well, while the darker regions have virtually no details.

Next, we increase the exposure to emulate a response function which reaches saturation faster than in the previous case. In this image (second from the left), parts of the outdoor region begin to saturate, but the dark door becomes visible. Increasing the exposure further gives us an image in which the outside looks almost completely washed out, while more details appear inside the dark room. In the final image, taken with an even longer exposure, the outside is completely washed out — even the door is saturated — but more details within the dark room are revealed. This method of capturing multiple images of a scene is called exposure bracketing.

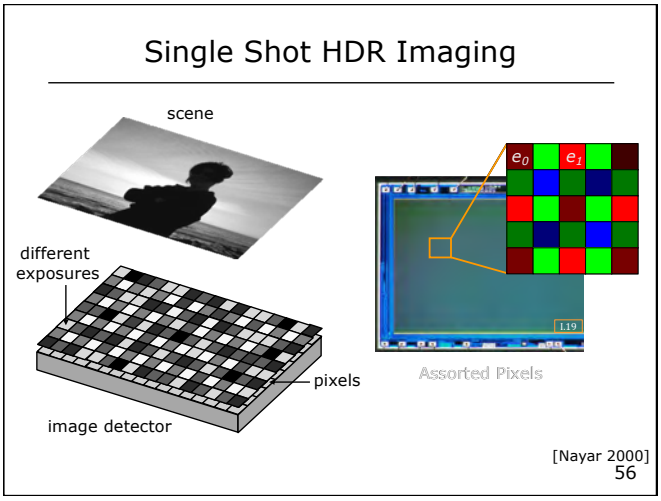


What will happen if these images are simply added together? In the summed image, the maximum measured value is 1020 — that is, 4 times 255. This image will look like one taken with another camera, a virtual camera, with a response function that looks like the plot on the left. It is simply the summation of the four response functions in the previous slide. This response function is non-linear and it compresses larger brightness values more than lower ones. Therefore, by simply adding images taken with different exposures, we can obtain an image with significantly greater dynamic range and hence more visual information. We can enhance the information in this summed image with a method called tone mapping to get the image shown on the right, where more details of the scene are revealed. This is a popular method for capturing high dynamic range images. It is used widely by smartphone cameras to enhance image quality.



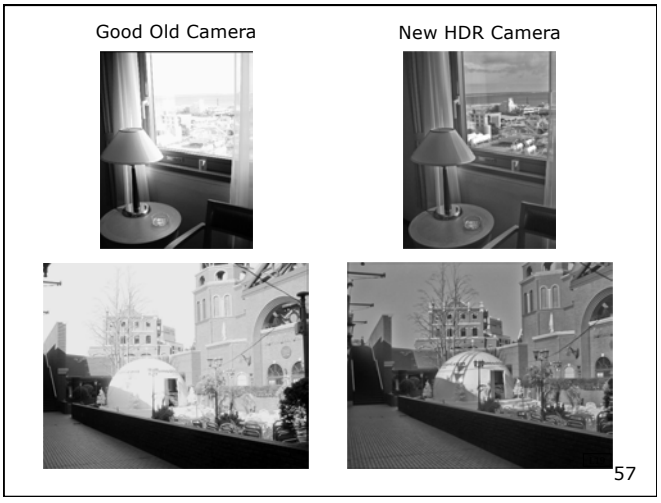
It turns out that the exposure bracketing method has a major limitation. The main assumption made by the method is that, while the multiple images with different exposures are being captured, the scene remains unchanged with respect to the camera. That is, each pixel corresponds to the same scene point during the entire capture process. Unfortunately, most scenes include objects in motion. In this example, the bicycles and the riders are in motion and hence they appear as multiple copies in the final HDR image. This is sometimes referred to as the ghosting artifact.

The only way to avoid the ghosting artifact is to capture all the information needed to compute the HDR image in a single image of the scene. Here we see a single-shot HDR method. The image sensors we have discussed thus far have pixels that have equal sensitivity to light. If we capture the scene shown here with such a sensor using a low exposure, the result would be an image in which the entire person is too dark. If we increase the exposure, the person might turn out fine, but the sky will be too bright or saturated. In both cases, there are large parts of the image that are devoid of useful information. Once this information is lost, it cannot be recovered.

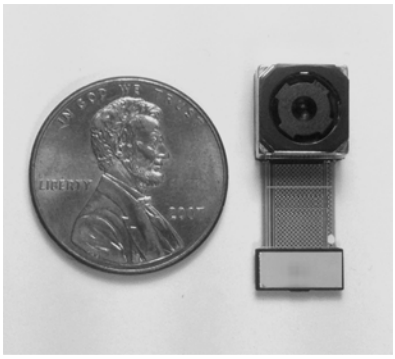


This problem can be remedied, or at least mitigated, by creating an image sensor with unequal pixels — pixels with different sensitivities to light. One way to do this is to place a "shade" with a chosen transparency on top of each pixel. Consider the case where the shades are assigned random transparencies. Then, for every dark pixel in the captured image, there is likely a neighbor that is not dark, and for every saturated pixel there is likely a neighbor that is not saturated. That is, there are no large areas in the image that are either too dark or completely washed out due to saturation. The captured image in this case looks "patterned." On the right is an image sensor with such an assortment of pixel exposures. The pixels vary not only in terms of their color filters but also exposures. Since we know the exposure of each pixel (the transparency of each shade) we can use image processing to map the captured image to a high dynamic range color image. Note that assorted pixels can also be implemented using different integration times for the pixels rather than different shades.

Let us compare the dynamic range of an assorted pixel image sensor with that of a traditional sensor. In the left column are images of two scenes taken with a DSLR camera with a traditional image sensor. Despite the high quality of the camera used, the images include large regions that are either too dark or saturated. In the right column are HDR images produced by the assorted pixel camera we described above. It is clear that these images reveal significantly more details in both the dark and bright regions.



This technology is used in many state-of-the-art image sensors, which are being used in popular phones. Here we see a camera module of the type that sits in today's smartphones, which includes the assorted pixel image sensor.



58

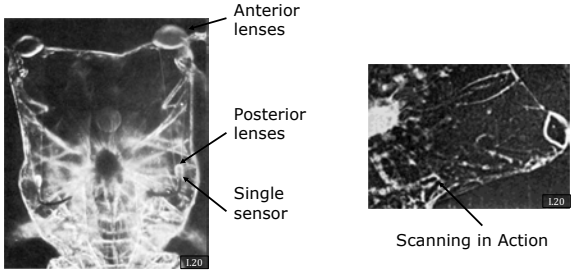
Nature's Image Sensors

Shree K. Nayar
Columbia University

Topic: Image Sensing, Module: Imaging
First Principles of Computer Vision

59

The Curious Eye of Copilia



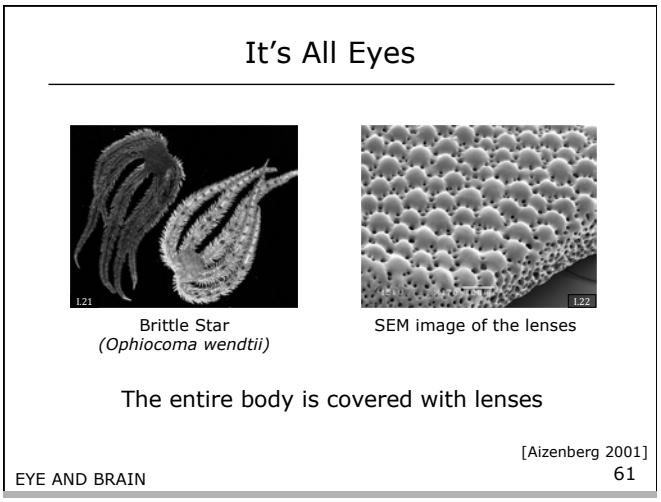
[Gregory 1964]

EYE AND BRAIN

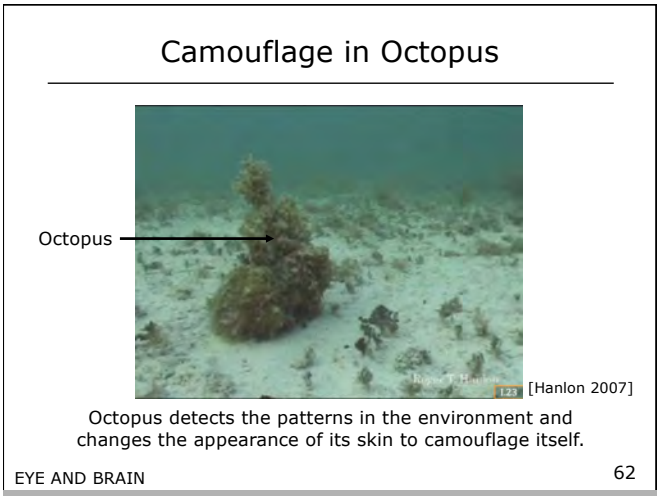
60

Let us take a look at some of the image sensors that nature has created. Here is an interesting example — the curious eye of the Copilia, which is a crustacean. It is similar to a plankton but has a long tail, which is not seen here. Shown on the left is its head, which includes two eyes. Each eye includes two lenses — an anterior lens, which is the large external lens forming an image, and a posterior lens, which has a single pixel (a single receptor) attached to it. The combination of the posterior lens and the single receptor mechanically scans the image formed by the anterior lens. On the right is a video of the scanning mechanism in action. It is remarkable that nature has developed an eye with a mechanical scanner to capture two-dimensional images.

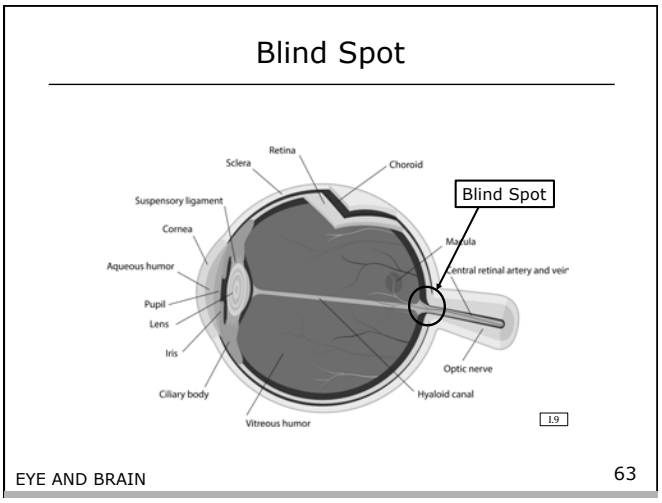
It was long believed that the brittle star did not possess any eyes. Since it does not have a brain — only a nervous system — it was a mystery to biologists as to how the brittle star was able to navigate the space around it and not fall prey to predators. Then, about twenty years ago, it was discovered that the entire body of the brittle star is covered with lenses. On the right is a scanning electron microscope image of a small piece of the body. Each one of the tiny bumps is a lens made of transparent calcite, roughly 1/20 millimeters in diameter. Each lens focuses light from a cone in the scene onto a nerve bundle. Therefore, the entire body acts like a flexible camera and is able to measure the spatial distribution of the light around it.



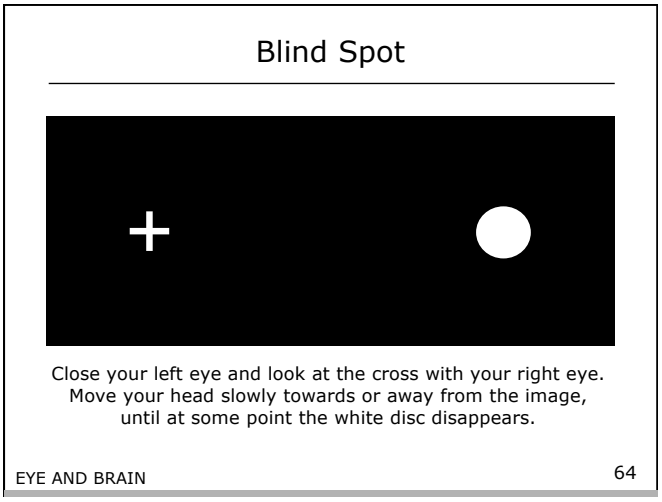
A truly incredible creation is the skin of the octopus. It includes a large number of chromatophores, which are little sacks with pigments in them. If one of these sacks is pulled to change its shape, the color of the light that the sack reflects also changes. Using this remarkable mechanism, the octopus is able to camouflage itself by controlling the visual texture of its skin to match that of its surroundings. In this video, there is an octopus sitting on the shrub, but it is not visible as it has taken on the texture of the shrub. When the camera gets close to the shrub, the octopus reveals itself and swims away.

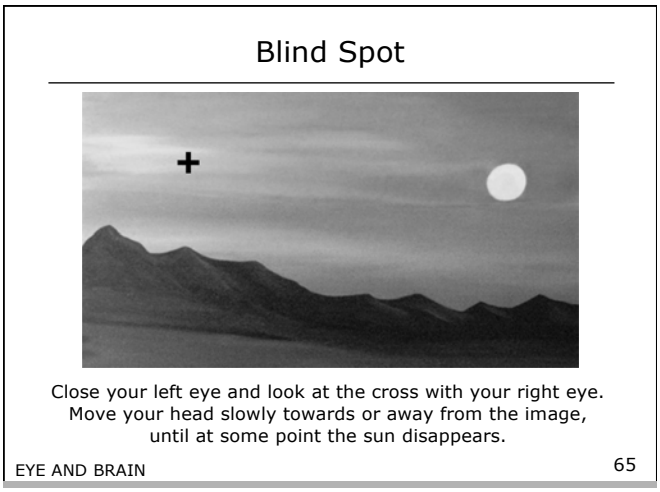


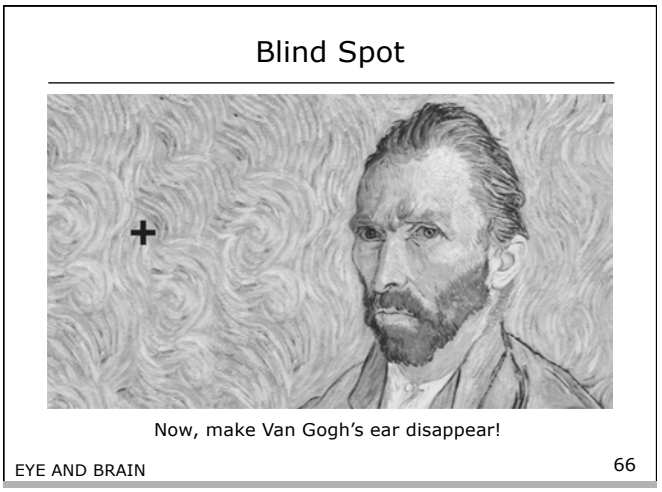
Let us return to the human eye. We have discussed the retina and the fovea in detail. As mentioned earlier, at the location where the image sensed by the retina is transmitted through the optic nerve to the brain, there is a blind spot — a patch on the retina that is devoid of rods and cones. This blind spot is not obvious to us because our brain fills in the missing information. It is often said that one must be careful of the blind spot when driving, as objects that appear within it will not be visible to the driver.



Here is a simple experiment you can do to find your blind spot. First, print the image with the cross and the disc so that it fills a letter-sized sheet of paper. Place the sheet about a foot in front of you, shut your left eye, and look at the cross on the left with your right eye. Now, slowly move the sheet towards and away from you. At some distance of the sheet, the white disc on the right will vanish. It vanishes when the image of the disc falls entirely within your blind spot.







Here are a couple more tricks you can play with yourself. Using the method described above, you can make the sun disappear in the image on the left. If you are into dark humor, you can make Van Gogh's ear disappear in the right image.

References and Credits

Shree K. Nayar

Columbia University

Topic: Image Sensing, Module: Imaging

First Principles of Computer Vision

67

References: Textbooks

Computer Vision: Algorithms and Applications
Szeliski, R., Springer

Computer Vision: A Modern Approach
Forsyth, D and Ponce, J., Prentice Hall

Robot Vision
Horn, B. K. P., MIT Press

68

References: Papers

[Aizenberg 2001] J. Aizenberg, A. Tkachenko, S. Weiner, L. Addadi and G. Hendler. "Calcitic microlenses as part of the photoreceptor system in brittlestars." Nature, 2001.

[Clark 2006] R. N. Clark. "Digital Camera Sensor Performance Summary". http://www.clarkvision.com/articles/index.html

[Fairchild 2001] M. D. Fairchild. Color Appearance Models. John Wiley & Sons Inc., 2001.

[Gregory 1964] R. L. Gregory, H. E. Ross and N. Moray. "The Curious Eye of Copilia". Nature, 1964.

[Gregory 1990] R. L. Gregory. Eye and Brain. Princeton Science Library, 1990.

[Grossberg 2003] M. D. Grossberg and S. K. Nayar. "What is the Space of Camera Response Function?". CVPR, 2003.

[Halon 2007] R. Hanlon. Cephalopod dynamic camouflage. Current Biology 17 (11), 2007.

[Hubel 1987] D. H. Hubel. Eye, Brain, and Vision. Scientific American Library, 1987.

69

References: Papers

[Kaiser 1996] P. K. Kaiser. The Joy of Visual Perception: A Web Book. http://www.yorku.ca/eye/

[Litwiller 2005] D. Litwiller. "CMOS vs. CCD: Maturing Technologies, Maturing Markets". Photonic Spectra, August 2005.

[Mann 1995] S. Mann and R. Picard. "Being 'Undigital' with Digital Cameras: Extending Dynamic Range by Combining Differently Exposed Pictures", Proc. Of IST's 48th Annual Conference, May 1995.

[Mitsunaga 1999] T. Mitsunaga and S. K. Nayar. "Radiometric Self Calibration". CVPR, 1999.

[Nakamura 2006] J. Nakamura. Image Sensors and Signal Processing for Digital Still Cameras. CRC Press, 2006.

[Nayar 2000] S. K. Nayar and T. Mitsunaga. "High Dynamic Range Imaging: Spatially Varying Pixel Exposures". CVPR, 2000.

[Nayar 2002] S. K. Nayar and S. G. Narasimhan. "Assorted Pixels: Multi-Sampled Imaging with Structured Models". ECCV, 2002.

70

Image Credits

I.1 Gemma-Frisius, 1545. Public Domain.
I.2 https://en.wikipedia.org/wiki/File:Camera_Obscura_box18thCentury.jpg . Public Domain.
I.3 Louis-Jacques-Mande Daguerre, 1837. Public Domain.
I.4 Louis Ducos du Hauron, 1877. Public Domain.
I.5 Ernemann Camera. HistoricCamera.com. Used with permission.
I.7 http://hamamatsu.magnet.fsu.edu/articles/microlensarray.html
I.8 TechInsights, Inc. Used with permission.
I.9 Purchased from iStock by Getty Images.
I.10 Color Appearance of Models. M.D. Fairchild. John Wiley & Sons, Ltd. Used with permission.
I.11 https://commons.wikimedia.org/wiki/File:ScanEMphoto.jpg. Licensed under License Art Libre (France).
I.12 Mark D. Fairchild. Licensed under CC BY SA 3.0.
I.13 Color Appearance of Models. M.D. Fairchild. John Wiley & Sons, Ltd. Used with permission.

71

Image Credits

I.14 Eye and Brain, 4th Edition. Richard L. Gregory. Princeton University Press. Used with permission.
I.15 https://commons.wikimedia.org/wiki/File:Color_Separation_Prism.jpg. Dicklyon. Public Domain.
I.16 https://commons.wikimedia.org/wiki/File:Dichroic-prism.svg. Colin M.L. Burnett. Licensed under CC BY SA 3.0.
I.19 Courtsey of Sony, Inc.
I.20 The curious eye of Copilia. R.L. Gregory et al. Nature, 1964. Used with permission.
I.21 J. Aizenberg. Used with permission.
I.22 J. Aizenberg. Used with permission.
I.23 R. Hanlon. Used with permission.

72

Acknowledgements: Thanks to Nisha Aggarwal and Jenna Everard for their help with transcription, editing and proofreading.

References

[Szeliski 2022] Computer Vision: Algorithms and Applications, Szeliski, R., Springer, 2022.

[Forsyth and Ponce 2003] Computer Vision: A Modern Approach, Forsyth, D and Ponce, J., Prentice Hall, 2003

[Horn 1986] Robot Vision, Horn, B. K. P., MIT Press, 1986.

[Gregory 1966] Eye and Brain, Gregory, R., Princeton University Press, 1966.

[Aizenberg 2001] J. Aizenberg, A. Tkachenko, S. Weiner, L. Addadi and G. Hendler. "Calcitic microlenses as part of the photoreceptor system in brittlestars." Nature, 2001.

[Clark 2006] R. N. Clark. "Digital Camera Sensor Performance Summary". http://www.clarkvision.com/articles/index.html

[Fairchild 2001] M. D. Fairchild. Color Appearance Models. John Wiley & Sons Inc., 2001.

[Gregory 1964] R. L. Gregory, H. E. Ross and N. Moray. "The Curious Eye of Copilia". Nature, 1964.

[Grossberg 2003] M. D. Grossberg and S. K. Nayar. "What is the Space of Camera Response Function?". CVPR, 2003.

[Halon 2007] R. Hanlon. Cephalopod dynamic camouflage. Current Biology 17 (11), 2007.

[Hubel 1987] D. H. Hubel. Eye, Brain, and Vision. Scientific American Library, 1987.

[Kaiser 1996] P. K. Kaiser. The Joy of Visual Perception: A Web Book. http://www.yorku.ca/eye/

[Litwiller 2005] D. Litwiller. "CMOS vs. CCD: Maturing Technologies, Maturing Markets". Photonic Spectra, August 2005.

[Mann 1995] S. Mann and R. Picard. "Being 'Undigital' with Digital Cameras: Extending Dynamic Range by Combining Differently Exposed Pictures", Proc. Of IST's 48th Annual Conference, May 1995.

[Mitsunaga 1999] T. Mitsunaga and S. K. Nayar. "Radiometric Self Calibration". CVPR, 1999.

[Nakamura 2006] J. Nakamura. Image Sensors and Signal Processing for Digital Still Cameras. CRC Press, 2006.

[Nayar 2000] S. K. Nayar and T. Mitsunaga. “High Dynamic Range Imaging: Spatially Varying Pixel Exposures”. CVPR, 2000.

[Nayar 2002] S. K. Nayar and S. G. Narasimhan. “Assorted Pixels: Multi-Sampled Imaging with Structured Models”. ECCV, 2002.

[Nayar 2022B] Image Formation, Nayar, S. K., Monograph FPCV-1-1, First Principles of Computer Vision, Columbia University, New York, February 2022.

[Nayar 2022E] Image Processing I, Nayar, S. K., Monograph FPCV-1-4, First Principles of Computer Vision, Columbia University, New York, March 2022.

[Nayar 2022F] Image Processing II, Nayar, S. K., Monograph FPCV-1-5, First Principles of Computer Vision, Columbia University, New York, March 2022.