

A paper published by Northlight Images, written by David B. Goldstein.
Article Copyright ©2009 David B. Goldstein
Version 2.0

All technical queries should be directed to the author, who may be reached at:
savingenergygrowingjobs@yahoo.com
The web version of this article can be found at:
http://www.northlight-images.co.uk/article_pages/guest/physical_limits.html

Physical Limits in Digital Photography

By David B. Goldstein

Abstract

Digital photography has made immense progress over the past 15 years, particularly in *increasing resolution* and *improving low-light performance*. But continuing progress in these aspects of performance is not possible, because cameras are now approaching the limits imposed by the basic nature of light.

This article looks at the physics that explains these limits. It shows how the fact that light has the characteristics of both particles and waves puts caps on resolution and “film speed” that are not far beyond the performance of current high-end dSLRs, and are already impacting point-and-shoot cameras.

Resolution is *limited by the wave phenomenon* of diffraction: the current generation of full-frame sensors begin to be diffraction-limited at about f/10 and above, and the best APS sensors at about f/7. More pixels can only improve resolution and contrast at wider apertures and if new lenses are up to the task. Pocket cameras begin to be diffraction limited at f/2.4 and above, which is faster than all but one available lens.

High ISO performance is *limited by the number of photons available per pixel*. ISOs above about 3200 will always result in noise, no matter how much sensors are improved. Actually, the dynamic range of current dSLRs implies that noise in the deepest shadows is limited by photon count even at low ISOs.

These limits change the nature of tradeoffs that photographers have been making concerning film or sensor size, depth of field, and sensitivity.

I. Introduction

Quantum mechanics requires that the wave and particle natures of light or matter cannot be seen at the same time. But this only applies for a single observation; if we make large numbers of observations at once, we can see both wave-like and particle-like properties at the same time.

In the last few years, technologies have emerged in consumer markets that allow anyone to see evidence of both the wave and particle nature of light at the same time. The technologies are those of digital photography.

This paper explores how the limitations imposed by quantum mechanics can be seen by examining digital photos taken with state-of-the-art current equipment. It will also show the converse—that the potential for improving the resolution of digital photos and the ability to take them in low light is low—that technology is now pushing the physical limits.

At the heart of a digital camera is a bundling of several million photon detectors arranged in a pattern that allows the user to see:

- The particle nature of light. By looking at the variation in color and light level from pixel to pixel when the sensor is illuminated by a uniform source, the user is seeing the differing number of photon counts; and
- The wave nature of light. This is visible in the form of diffraction by a fuzzy, low contrast image.

Indeed, both of these phenomena are so pronounced that they limit the “film speed” and resolution of digital cameras to levels within a factor of two or so of the performance of current high-quality product offerings. This paper computes, in a simple approximate form, the physical limits to film speed and resolution for popular sizes of cameras, and shows how these limits are binding—or nearly so—at current levels of engineering. The approximate form is appropriate because, as will be demonstrated, the arbitrariness of useful criteria for sharpness and noise make it useless to try for more accuracy of calculation.

II. Film Speed

What is the limit on signal to noise based on fundamental physics?

This calculation looks at the number of photons incident on a pixel and shows that current technology is close to physical limits. Film speeds of ISO 3200 are so close to physical limits that large increases are impossible without compromising other parameters.

The calculation begins by determining how many photons of light are available to be measured by a digital sensor (or film). The energy of a photon of light at $\lambda=500$ nm—the mean wavelength of visible light—is given by $E=h\nu$; since $h = 6*10^{-34}$, and $\nu=c/\lambda$ or $6*10^{14}$ sec⁻¹,

$$E=3.6*10^{-19} \text{ joules per photon.}$$

How many photons are incident on a pixel of sensor?

To start with, sunlight is 1.37 kW/m² at the top of the atmosphere, or a little less than 1 kW/m² at ground level with normal incidence of sunlight. (Normal [90°] incidence provides the

brightest possible light.) Of this light, about 45% is visible, the rest being mainly infrared and a little ultraviolet. So the visible light on a square meter of illuminated object is about 400 W. A very bright object will reflect almost all of this light.

But the meaningful criterion for visible noise is how the camera renders the darker parts of the image.

The ratio of the lightest light level that a sensor (or film) will capture to the darkest is called “dynamic range” by photographers, and is expressed as an exponent of 2, called an “f-stop”. The dynamic range of the most advanced digital cameras at a high sensitivity—ISO 3200—is about 8 f-stops¹, or a ratio of 2⁸. The issue of dynamic range is complex, but this estimate will prove surprisingly robust in calculating the number of photons incident on a pixel in the darkest renderable shadows regardless of the sensitivity of the sensor or film to light, at least with current technology. So the range of light we need to analyze is 1/256 of the 400 W/m²; or 1.5 W/m².

With each photon at energy of 3.6*10⁻¹⁹ joules, this is a photon density about 4 *10¹⁸ photons per second per square meter.

How many of these get to the pixels? The answer depends on the size of the sensor or film, so I will address different sizes in turn.

Full frame 35mm sensors

Assume a state-of-the-art 35mm camera with 21-24 megapixels² (this assumption matters) taking a picture at its highest film speed, ISO 3200 (this assumption would appear superficially to matter, but turns out to be relatively unimportant) with a 50mm lens and the object one meter away, taking a picture at f/16 and 1/3200th of a second, the correct exposure for ISO 3200 (these assumptions don't matter but are made just to make the derivation easier to follow).

The field of view of the lens at 1 m is 720 x 480 mm, or .35 square meter, so the photon density is about 1.5*10¹⁸ photons per second. This photon density represents the darkest shadow capable of being captured with resolution by a professional-grade camera at this “film speed”.

Each point in this field of view reflects into a hemispheric solid angle (optimistically; some reflectors will reflect into a full sphere); the amount of light that can be transmitted to the film is the reflected light times the ratio of the solid angle of the lens aperture as seen from a point on the reflector to a hemisphere. A 50mm lens at f/16 has an aperture diameter of 3mm and a radius of 1.5 mm, so the ratio is $\frac{\pi * 1.5^2}{2\pi * 1000^2}$, or 10⁻⁶.

¹ A typical dynamic range for film is about 7 f-stops (an f-stop is a factor of 2); pocket digital cameras perhaps have a little less. As of spring 2009, the four most advanced digital cameras had dynamic ranges of about 8 stops at ISO 3200 (see dxomark.com). So this paper provides a range based on these limits.

² These spec's apply to the Canon 1Ds Mark III and 5D Mark II, both at 21 megapixels, and the Nikon D3X and the Sony Alpha 900, both at 24 megapixels; the pixel size is the same or smaller (see text below) for smaller-frame cameras (image size about 15 mm by 22 mm) produced by numerous manufacturers.

So the photon density on the aperture is about $1.5 \cdot 10^{12}$ photons per second³. These photons strike 21-24 megapixels at a rate of $7 \cdot 10^4$ photons per pixel per second.

For a proper exposure at 1/3200 second, the reception rate is 20 photons per pixel for an exposure. This is small enough for noise (statistical variation in the number of photons registered per pixel) to be evident even with perfect photon detection.

This rate is quite robust to changes in assumption. Had we chosen a different combination of shutter speed and aperture, we still would have gotten the same answer for the proper exposure. While we chose the film speed to see how much a speed of ISO 3200 is pushing hard against the limits of physical laws, it turns out that with all of the professional cameras we examined, the dynamic range is a very strong negative function of the sensitivity.

Thus, as sensitivity increases from ISO 100 or 200 to ISO 3200 (or higher in some cameras), *the dynamic range drops nearly as fast as the sensitivity increases*. Thus as the camera gains 5 f-stops in sensitivity, the dynamic range drops 3 to 4.5 stops⁴. So the increase in sensitivity in the deepest shadows is actually only $\frac{1}{2}$ -2 stops. Most of the increase in sensitivity comes from boosting the highlights and midrange: the darkest shadows rendered at low sensitivity are almost as dark (less than factor-of-two difference in one case; about factor-of 4 in the other). The photon reception rate for the darkest renderable shadows is not that different at different “film speed” sensitivities!

Looking at this phenomenon the other way makes it more intuitive. If the sensitivity in the darkest shadows is limited by photon noise, then as we move to lower and lower ISOs, we have the opportunity to increase dynamic range by extending the upper limit of pixel sensitivity. The upper limit is defined by the characteristics of today’s photon detectors and the electronics⁵ and not by hard physics limits.

This reception rate of about 20 photons per pixel is already a noise risk, since the signal to noise ratio for random events such as photon strikes is \sqrt{n} : this is a S/N ratio of 4.5—not very good. (A S/N of 4.5 assures with high probability (but not near-certainty) that an observed signal is real, but it allows lots of variation from pixel to pixel of the image darkness when the light incident is actually the same. This makes for *grainy, speckled* photos.)

Noise is more than an aesthetic problem, though. Different people have different opinions about the merits or demerits of a grainy-looking photo, so if this were the only problem, a signal-to-noise ratio of 4.5 might not be a problem.

³ This estimate is not a strong function of the angle between the subject and the lens.

For narrow angle lenses, the difference between the solid angle and a linear approximation is trivial. For wide angle lenses, the aperture as seen from a large angle from normal to the film plane as a circle, not a foreshortened ellipse, so the solid angle is about the same for closer-to-normal incidence. Any error in this assumption is negligible compared to the assumption of perfectly diffuse reflection.

⁴ Data on dynamic range of real cameras are taken from dxomark.com.

⁵ The mechanisms for this limit are described in

<http://www.clarkvision.com/imagetdetail/digital.sensor.performance.summary/>

But there is a more important issue. At only slightly lower signal-to-noise ratios such as 2 or 3, which are what results when color detection is added to the calculation (see discussion below), the viewer starts to have trouble discerning what is a part of the picture and what is noise.

For example, when looking at a picture taken at ISO 3200 that includes a starry sky (and you CAN do this—I have some handheld pictures taken at f/1.4 that show clearly identifiable constellations) is that red dot another star? an airplane? or noise? Are those bright red and green dots on my daughter’s cheeks sparkle makeup or an artefact? Is that black spot on my walnut table a bug or noise?

As the signal-to-noise ratio drops further, the confusion extends beyond mere details like these and gets to the fundamental problem of what is visible at all (or visible but not really there). So the limits I describe are binding without regard to the viewer’s aesthetic preferences.

And this theoretical number is reduced fourfold by the fact that the calculation above implicitly assumes that we don’t have to worry about color. But in fact, we do care about color, and the current technology for digital photography measures only the existence of the photon, not its energy (color). And even color-responsive sensors would only work to reduce noise if the light incident on the pixel was monochromatic, which is not always the case. So the signal to noise calculation above is only valid for black-and-white photography.

Current sensor technology detects color by sending the photons through filters. Each pixel sees light filtered by either a red filter or a blue filter or one of two green filters. Thus the number of photons that make it through the filter channels is at most 1/3 of the number calculated above, or 6 photons.

Even with perfectly transmissive lenses and 100%-effective sensors, the signal to noise ratio for red and blue will be 2.4 ($=\sqrt{6}$), so color noise will be a big problem in the deep shadows and a significant problem even in the milder shadows. Note that color noise (called “chrominance noise” by photographers) means that we have to be concerned about how nearby pixels rendering a uniform subject will have random variations in color as well as in brightness.

These issues are plainly visible to the photographer. For example, the difference between black-and-white photos and color is evident in lower noise for monochrome. This phenomenon has been noted in numerous camera reviews in the popular press. The author has also noted this by taking pictures at an effective ISO of 25,000 on a 21 megapixel camera⁶ and looking at them on a computer screen. See Figure 1 for an example⁷. This example was taken at ISO 800 to minimize any possible questions that the noise might be due to heavy electronic amplification; the low photon count was achieved by underexposing somewhat such that the dark shadows were at the low end of the dynamic range.

⁶ This is done by intentionally underexposing 3 stops with the ISO setting at 3200.

⁷ Figure 1 was taken with a Canon 50 D, which has a pixel size of about half the size calculated here, so the number of photons per pixel under these conditions would be about 3.

The conclusion then is that it will be nearly impossible to increase film speed without increasing noise dramatically in the deep shadows and significantly in the middle dark shadows. Speed can only come at the expense of reducing the pixel count by making each one bigger. Or at the expense of dynamic range compared to what is expected of quality photography in color.

This latter conclusion appears to be largely a consequence of the underlying physics. If the ability to amplify the signal in the darkest shadows is limited by photon variation, then *there is a fixed limit to how dark an object can be and still be “visible”/rendered at all*. Exposure at a given ISO is determined by measuring the light reflected from an 18% reflectance grey card in the bright part of the picture. If we try to make that 18% card brighter and brighter on the photo, but cannot make the darkest shadows any brighter, most or all of the increase in film speed can only be achieved at the bright end of the brightness histogram; the dark end won't change. Thus the ratio of darkest black to whitest white must deteriorate as fast as the bright end is amplified.

What about those other assumptions? First the ones that don't matter. The 50 mm lens doesn't matter because if you doubled the focal length (without changing the sensor size), the aperture at f/16 would be twice as wide and have 4 times the area, but the field of view would be $\frac{1}{4}$ the area of the reflector, so the photon density on the aperture would be the same. So the lens focal length doesn't matter. Suppose the reflector were 2m away instead of the assumed 1m. Then the field of view of the lens would be twice as big, the area 4 times as big, but the solid angle of the aperture would be $\frac{1}{4}$ the size. So the number of photons incident on the aperture would be the same.

Next the ones that do. If the camera has $\frac{1}{2}$ the megapixels, then each receives twice the photons and the noise problem is $\frac{1}{2}$ as bad (or more correctly $1/\sqrt{2}$ times as bad). Thus the camera can use twice the ISO speed for the same noise level. The number of photons needed for a good exposure is inversely proportional to film speed, so lowering the film speed to ISO 100 increases the photon count 16 times and improves the signal to noise ratio 4 times. BUT as noted above, with current technology, the dynamic range improves almost as fast as film speed, so the photon count in the darkest shadows is not much higher at low sensitivity. The darkest shadows are darker, but they are just about as noisy.

Smaller sensors for interchangeable lens cameras

For more typical size moderately priced digital interchangeable-lens cameras, the pixel count is typically 10-15 megapixels for a sensor 1.5 to 2 times smaller in linear dimensions. So the typical pixel density is the between the same and 2 times higher, meaning the film speed limit is the same to half as high. Most of these cameras have lower dynamic range than their top-of-the-line competitors, so the signal-to-noise ratios in the darkest shadows at lower ISO will be slightly worse than full frame, but for the higher speeds they will be up to $\sqrt{2}$ worse.

Pocket cameras

If the sensor is 1/5 the size of 35mm, typical of point-and-shoot cameras, a lens with the same angle of view has 1/5 the diameter of the assumed 50 mm lens, so the area of the aperture is 1/25 as large. Thus it captures only 1/25th the number of photons. Most cameras of that size use 8 to 12 or 15 megapixels, so the number of photons per pixel would be about 1/10 the level calculated above.

This is so troublesome that we can see why pocket cameras may not even offer ISO 1600 and have visible problems with noise at 400 and even lower. (The cameras that offer high ISO (3200 to 6400) do so by combining pixels, effectively meaning that there are about 1/4 as many of them as you would expect, reducing resolution by a factor of about 2.)

In practice, the noise problem isn't quite as much worse than this calculation suggests, but that is because the dynamic range of pocket cameras is much less than for the larger and more expensive alternatives. We can't quantify here how much difference this makes because measured data on dynamic range are not published in photography magazines or manufacturer literature for such products.

This lower dynamic range affects how noisy the overall picture will look. But it does not mitigate the problem in the darkest shadows: pocket cameras are limited to about 10 times brighter light than full frame.

Ways Around the Problem

The noise problem is fundamental, in that a pixel of a given size only receives the number of photons calculated here. But noise can be reduced at the expense of resolution by averaging over adjacent pixels. Most cameras and essentially all computer software allow noise reduction by losing information.

Noise could also be reduced by binning pixels, and some cameras do this already. If four pixels are binned, obviously the combination receives four times the number of photons. If the four are chosen to have each of the four color channels, we don't lose as much information on color as one might think.

Smaller pixels also look less noisy. Imagine a ten by ten array of pixels where one pixel is now. Each pixel will now be exposed to, most likely, one or zero photons in the deep shadows. So on a pixel level, the noise will be extreme. But at the level of a viewer, the tiny pixels will be like a pointillist painting, and the noise will be less bothersome.

One can envision smart software that bins pixels only in the shadows, allowing higher resolution in those parts of the picture that are bright enough to support low noise.

It appears possible, at least in theory, to overcome at least some of the factor-of-3 loss due to color filtration by finding a way to detect color without filters. Current sensor technology

detects photons through allowing them to penetrate into a silicon semiconductor and transfer their energy into exciting electrons into the conduction band.⁸ The mean depth for absorption is a function of photon energy, which is inversely proportional to wavelength. Light at the blue end of the spectrum has a mean absorption depth an order of magnitude shorter than light at the red end. Thus one could imagine a sensor that measures the depth at which photons were absorbed rather than just the fact of absorption. Photon absorption is a statistical process, so this method, even if feasible to implement in practice, might not provide sufficient discrimination between colors to be practical.

Vision

This discussion is based on the physical properties of light, not on the technology of cameras, so it applies equally to other forms of imaging, including the human eye. And, of course, the eyes of other animals as well.

If noise is a problem at the resolution of a normal lens on a camera, it is equally a problem for vision. So my observation that a digital camera with a f/1.4 lens at 1/30th of a second can see more shadow detail than my eye can is not an accident, it is a consequence of the fact that both eye and camera are facing the same photon limits and the eye has a somewhat slower lens. The eye attempts to solve this problem in very low light by using receptors that are not sensitive to color. Dark vision is also less clear, meaning that “pixel binning” must be going on as well. And the “shutter speed” of the human eye can adjust to include longer exposures⁹. One analysis estimates the ISO speed of a dark-adapted eye is about 800¹⁰.

Even with these adaptations, there are still hard limits to how well an owl, for example, can see in the dark.

III. Diffraction

It is well known from elementary physics texts that the minimum distance between images that can be resolved when they are cast through a single round hole aperture is given by the angle $\Theta = 1.2 \lambda/d$, where λ is the wavelength and d is the diameter of the aperture. (This is true for small angles as described next). This affects different film or sensor sizes in different ways.

Full frame 35mm sensors

For a typical light wavelength of $\lambda=500\text{nm}$ and a 35mm camera with a 50mm lens set at f/10, $d=5 \text{ mm}$ and $\Theta=1.2 \times 10^{-4}$. This angle corresponds to an image size of $6\mu\text{m}$ at 50mm, or 1/4000

⁸ <http://www.clarkvision.com/imagedetail/digital.sensor.performance.summary/index.html>

⁹ <http://clarkvision.com/imagedetail/eye-resolution.htm>

¹⁰ Ibid.

of the vertical dimension of the sensor (24mm—a “35mm camera” has a sensor size of 24 by 36 mm).

However, the diffraction limit referred to above is not the limit of how small an image can be rendered *sharply*—it is the limit to how small a detail can be distinguished *at all*. (Actually, as discussed below, sophisticated image processing can extend this limit somewhat.) It does not consider that the quality of the image will be greatly degraded. In technical terms, this loss of information and contrast is referred to as the Modulation Transfer Function or MTF. Readers not familiar with this concept can search on-line for information at their appropriate level of scientific depth. While an MTF of 10% or 5% or perhaps even less can still allow information to be seen, discussions of human perception of sharpness discount the importance of resolution when the MTF drops below 35% or even 50%¹¹, so this discussion will focus on resolving points with good contrast.

The Airy disk of a circular slit consists of a widening and reshaping of the image of a point source, spreading the central point but also consisting of a series of secondary rings. All of this divergence from a point adds noise to the image; it is seen as reduced contrast within dimensions noticeably larger than Θ .

At an angle of $\Theta = 1.2 \lambda/d$, the first minimum of the diffraction pattern of one point source overlaps the maximum of the adjacent one’s pattern; this conventionally has been accepted as the minimum angular separation required to distinguish two separate sources, and is referred to by physicists as the Rayleigh criterion. (Actually, Rayleigh developed this criterion subjectively for use with visual observations. When astronomers began using film, they found that with careful image processing one can distinguish two sources at slightly closer spacing than this, but you have to know that you are looking for two (as opposed to three or ten) separate points).

But the contrast difference between the two adjacent maxima is minimal. To render two adjacent sources with clarity (minimal loss in contrast) would require at a minimum that the *first minima* of the diffraction patterns overlap. This doubles the spatial frequency of the tightest pattern that can be rendered clearly as a result of diffraction. That is, it reduces the number of vertical lines that a camera needs to resolve at the assumed conditions from 4000 vertical lines to 2000.

This can be compared to the limit on sensor resolution imposed by the Nyquist-Shannon theorem, called the Nyquist frequency. For signals that are being sampled, such as by a digital camera sensor with a finite pixel pitch, the maximum frequency that can be resolved without loss of information is twice the sampling frequency. To resolve 2000 vertical lines at the Nyquist frequency requires 4000 by 6000 pixels. So the needed number of megapixels for

¹¹ For example, a 2009 lens review in a widely read on-line publication rated the lens in question by graphing its MTF as a function of position within the picture. The level of MTF selected for the graph was 50%. See http://www.dpreview.com/lensreviews/canon_24_3p5_tse_c10/page3.asp

sharp-looking (high contrast) pictures (at the assumed aperture of $f/10$) is about 24. This is about the number of megapixels currently found in professional 35mm digital cameras in 2009.

The falloff in MTF at point spacings closer than this is illustrated by a graph at the bottom of the following website:

<http://clarkvision.com/imagdetail/does.pixel.size.matter/index.html>

To see the results calculated here, look at the curve for the Canon 30D, which has the pixel spacing of $6\mu\text{m}$ employed above, and see how at the quickly the MTF curve drops above the diffraction criterion I have used.

Contrast (that is, low MTF) is important: it is not something you can fix by just adjusting the contrast slider in Photoshop or setting the sharpening for high levels at high frequencies. Noise from low MTFs is not the same as noise due to photon statistics, but it has the same effect on the physics: noise represents an irretrievable loss of information. In this case, noise represents the diversion of light from where it should be in a perfect photo to places where it doesn't belong. If there is any signal already at the place it doesn't belong, or even if there is NO signal but you aren't sure that there is supposed to be none, the noise can get confused with a real image and there is no way it can be post-processed away.

For example, if I look at an extreme enlargement (we ARE looking at *limits*, which imply such enlargement) of the eyes of my model and see a gradual shading of brown to greenish, is that what the model's eyes really look like? Or does she have distinct spots of green in her iris that are blurred due to diffraction? Are the small bright red spots on the butterfly wing in Figure 4a below iridescence, or noise? Image processing cannot answer questions of this sort¹².

¹² Image sharpening could only help increase apparent MTF or resolution only if we knew *a priori* that the image was going to get contrast-reduced at a scale of, say, 7 microns, and corrected at this spatial frequency, or sharpening radius. But diffraction decreases MTF on a sliding scale basis over a range of frequencies, so we would have to sharpen the picture repeatedly at ever increasing radius, based on knowing the MTF of the particular lens at the particular place on the image, object distance, and lots of other things. Even if we could do this, we would still be sharpening a spurious image some of the time.



Figure 4a
Butterfly with Iridescent Wings



Figure 4b
Butterfly with green wings

(This is the same image as Figure 4a except with noise reduction post-processing. Which represents the real butterfly?)

This implies that to make full use of the sensor capabilities of the best 2009 cameras already in production requires the use of apertures less than about $f/10$. And since the quality of the optics limit the resolution of most lenses today when apertures are much larger than $f/5.6$ or $f/8$, this suggests that about 25-40 megapixels is about the current practical limit for a 35mm camera, a limit that will apply until there are some breakthroughs in optics that can perform better at $f/4$ and faster than they do at $f/8$ or 10. As cameras go above this limit, they can still resolve slightly more fine detail, but the value of such detail in terms of better pictures, even at extreme enlargement, is in the region of quickly diminishing returns.

This effect of *seeing* that pictures are noticeably less sharp beyond the diffraction limit I have employed here has been noted in blogs¹³, where sample pictures show that detail close to the diffraction limit can be observed but looks fuzzy. The author has noted it as well, observing such evident degradation in image quality at f/22 that he no longer uses f-stops smaller than f/10. This can be seen in Figures 1-3. This empirical observation aligns well with the rough concept described above where the diffraction limit for clear and crisp pictures from a 21-24 megapixel camera at f/10 is 2000 vertical lines, which is consistent with the calculated Nyquist frequency of 4000 vertical pixels.

The other visual evidence that this is going on is that when I look at extreme enlargements of my sharpest pictures, the limit imposed by the optics seems comparable to the limit imposed by pixilation. For wider apertures, it looks like a finer pixel grid would improve the image, but not by a lot.

Most discussions of diffraction in photography concentrate on the Rayleigh criterion—the bare ability to distinguish between two points being separate. This corresponds to the first minimum of the Airy disk of one point overlapping the central maximum of the adjacent one. It allows astrophotographers to see when the image they are studying corresponds to one star or two. But as I argued previously, this criterion is not as useful for conventional photography because the contrast is so low when it is barely met. But, while not as useful, it does have some meaning, so it is worth analyzing the effect of diffraction at the Rayleigh limit as well.

This discussion begins to get complicated, because two seemingly identical phenomena come into play and produce opposite conclusions. The first is that image processing allows us to see detail at extremely low contrast, especially if we have *a priori* knowledge of what we are looking for. This would imply that more megapixels than the criteria calculated above would be useful.

The second is that different methods of image processing allow us to extract detail from a grid of pixels which is beyond the Nyquist limit. This phenomenon cuts in the opposite direction: it implies that the number of pixels needed for a level of performance that records detail at the diffraction limit is less than assumed above. Thus the diffraction limits are more severe than I have calculated, and the number of megapixels needed for photography are less than the results above.

I will only address these issues briefly. The bottom line is that the two effects largely cancel out, but the effect of image recovery at low contrast appears slightly larger. Thus, the value of additional pixels beyond the limits calculated above diminishes rapidly, but is not trivial. More pixels will indeed contribute to greater visible resolution, but the loss of quality due to lower MTF begins to be important at the limits calculated above.

¹³ Two of the better examples are: <http://www.luminous-landscape.com/tutorials/understanding-series/u-diffraction.shtm> and <http://www.cambridgeincolour.com/tutorials/diffraction-photography.htm>. Also see the review noted above in note 11, which uses the word “diffraction” repeatedly in reference to f-stops smaller than f/8.

What if we looked at what it would take to resolve images with finer detail (and also evidently at lower contrast) closer to the Rayleigh limit? Several different phenomena must be examined to answer this question and, as noted, many of them are offsetting.

The first is that the Rayleigh criterion occurs at twice the spatial frequency calculated above. So to resolve all data up to a frequency corresponding to 4000 lines—the Rayleigh criterion--would require a Nyquist frequency of 8000 vertical lines, corresponding to 100 megapixels.

The Rayleigh criterion was derived based on a simple model that correctly predicted what astronomers could see. More recent astrophotographic techniques allow stars to be distinguished up to the point that MTF drops to zero. This is about 20-25% closer spacing than the Rayleigh criterion, and is referred to as the Dawes limit¹⁴. If we wished to use this as the criterion for resolution, then the required sensor resolution would be about 150 megapixels. It is also possible for astronomers to detect whether a star image is a single star or a binary star even if there is no separation between the two adjacent maxima: the form of the merged maximum can still be indicative of a binary subject.¹⁵ But there is a catch to the latter method: you have to know in advance that you are looking for two closely separated points. If you have no *a priori* information about what the subject is, this method won't work. So it is pretty much useless for normal photography.

But these effects are counterbalanced by the ability to process images for finer resolution than suggested by straightforward application of the Nyquist-Shannon theorem. At frequencies above the Nyquist frequency, we can still see *some* of the image data, but not all. And some of what we are missing can be recovered by image processing.

It is easier to derive information from a frequency above Nyquist if you know in advance what you are looking for. A sharp edge or a single line much thinner than one pixel can be seen clearly if you know that this is what the subject is, and these examples occur so often in real pictures that it would be surprising if RAW deconvolution algorithms did not account for them.

Of course, above-Nyquist frequencies generate artefacts. But for a current style digital camera with a Bayer filter for color detection, we are stuck with artefacts anyway, even well below Nyquist. This is discussed in the next section below. The end result is that sensors can in many cases resolve details slightly or even significantly above the Nyquist frequency.

A *practical* upper limit would be double the number of pixels compared to the high-contrast diffraction limit--as opposed to increasing them four times--in order to capture essentially all the information available, according to several bloggers who have looked at this issue in real photos¹⁶. So the diffraction would limit the needed level of megapixels to about 50 at f/10.

¹⁴ http://www.lichta.de/astro_article_mtf_telescope_resolution.php

See also <http://clarkvision.com/imagedetail/scandetail.html#diffraction>

¹⁵ Ibid.

¹⁶ One of the more detailed explanations is Brain P. Lawler, "Resolving the halftone resolution issue: how many dpi does it take to make an lpi?"

Real images have all sorts of spatial frequencies. The upper limit of resolving power for full information reproduction is not as important with a range of frequencies as it would appear from a resolution test that contains all of the information at the high spatial frequency. In many cases, high spatial frequencies are needed only to render sudden changes in intensity such as sharp edges rather than actual objects, just as high frequencies of sound are needed primarily to allow us to hear the difference between a violin and a piano rather than because any musical notes are at these frequencies.

In this case, especially, the Nyquist-Shannon theorem is not fully applicable. The theorem refers to point sampling of a continuous image. But pixels are not points: they sense average light intensity over a finite area. Especially when the high frequency signal one is trying to render is an “overtone” of a lower frequency fundamental (such as an attempt to make a sinusoidal signal into a square wave), this averaging over a finite space makes it easier to detect “beyond-Nyquist” frequencies.

So far, this discussion has assumed monochromatic light. For red light, the diffraction is about 40% worse, and so for multi-colored light points, diffraction will produce color fringing with red on the outside. This makes for a visually fuzzier image than for a single-wavelength approximation. Color creates even more interesting problems, as I will discuss below.

This difference between practical limits due to diffraction and theoretical limits seems to parallel the old film-versus-digital debate. Film can resolve more detail than digital at the Rayleigh limit (or similarly low MTF limit imposed by lens aberration) close to or perhaps beyond 24 megapixels full frame. But it resolves such detail at very low contrast. At more modest spatial frequencies, film is noticeably worse than digital in MTF/contrast. The net effect is that the perceived crossover point where digital looks sharper than film is in the range of ~3-6 megapixels.

This surprising result may account for the fact that the sudden transition from film to digital was largely unforeseen. Photographers were waiting for the ultimate resolution of digital to surpass film and unwilling to recognize that digital pictures looked sharper long before that threshold was reached. One review of one of the first 3 megapixel cameras noted as early as 2000 that the results surpassed those of Velvia film in quality.¹⁷ Yet the magazine *Popular Photography* predicted as late as 2001 that there was still a healthy future for small-format film photography¹⁸. And while the magazine did in fact note in its review of the Canon 1Ds and 1Ds Mark II that these cameras produced superior pictures to those of film cameras despite slightly lower resolution, its editors did not generalize these findings to a conclusion that good MTF at moderate spatial frequencies was more important than greater-than-zero MTF is at the highest frequencies.

¹⁷ http://luminous-landscape.com/reviews/cameras/d30/d30_vs_film.shtml

¹⁸ Michael J. McNamara. “Film vs. Digital: Is film about to get knocked out of the picture by digital? Or is the fight of the century just warming up?” *Popular Photography*, March 2001.

One presentation argued, in the context of video photography, that the human perception of sharpness was proportional to the integral of the square of the MTF as a function of frequency—to the area under a curve of MTF-squared graphed by frequency¹⁹. This observation would corroborate these findings—arguing that high MTF at moderate frequencies, which digital is better at than film, are more important to perceived sharpness than how high the MTF extends at contrasts of 10% or so.

After scanning some 18,000 slides, I concur. I found that my very sharpest slides, taken in contrasty sunlight on Kodachrome 25, are comparable to digital shots at about 6-10 megapixels, at least in terms of resolution (but they are worse in rendering contrast at these resolutions). But for more typical subjects, the crossover point is closer to 3-4 megapixels or less. And for higher speed film or for lower-contrast subjects, it is at 2 megapixels or even less.

Other than the article I cited in this context, and a few more (at least one by the same author), this difference was not really noticed in the photo magazine and website world, and digital took over from film in the marketplace before the experts had a clear consensus as to why.

This experience reinforces my argument that the most relevant calculation of the limits to sharpness imposed by diffraction is based on the high-contrast onset-of-diffraction limit that I focus on here.

So the practical “speed limit” for 35mm cameras is not that much higher than the current level of about 24 megapixels for a perfect f/10 lens. In order to make use of higher resolution sensors, we will require both the use of wider apertures than f/10 (at all times) and probably the invention of better optics that can produce more sharpness at wider apertures than is typical of current lenses. And of course, as shown above, each doubling of megapixels cuts the photon count in half. And at the 21-24 megapixels calculated above, we are pushing the issue on film speed and noise, so greater resolution comes at the expense of noise and speed.

This is not to say that pixel count never should, or will, go higher. This calculation suggests that significantly higher pixel count could produce noticeably sharper pictures with truly outstanding optics operated at about f/5.6 or wider. So the megapixel race may not stop dead at current levels. However, the rate of increase should slow, since improvements in optics and some retraining of photographers concerning depth of field are both necessary to realize much gain.

A Digression on Digital Camera Sensor Design

This discussion is oversimplified in a number of ways that interact with each other, relating to the form of the image we are trying to record, the way that digital camera sensors record color

¹⁹ http://media.panavision.com/ScreeningRoom/Screening_Room/Demystifying_Part1_480p.html

information and reduce artefacts caused by regular image details that interfere with the regular pattern of pixels, and the color of the images.

To analyze these effects we would have to construct a matrix of calculations that looked at the following issues in all possible combinations:

- Image characteristics: what sort of image detail are we trying to record? The discussion above implicitly assumed that the detail we were trying to distinguish is two point sources located closed together, or equivalently, a pattern of points located on square grid. The calculation would be different for:
 - Parallel lines a fixed distance apart. The diffraction formula would no longer have the factor 1.2 in this case.
 - A thin line (imagine a spider web strand) that we want to record as a line rather than have it disappear into the background.
 - A grid of points of unequal intensities
 - A sharp edge. When we look at spatial frequencies, the ability to render sharp edges or fast patterns of gradation from light to dark depends on the ability to see regular patterns at a higher spatial frequency than what we want to record. This is analogous to the need to register higher sound frequencies to reproduce a specific wave shape. Thus to distinguish a 2000 Hz note on a flute from the same pitch on a violin depends on the ability of the stereo to reproduce sounds at 4000, 6000, and even 20,000 Hz. This would be mathematically equivalent to the case of looking at two points or two lines except for its interaction with the other issues discussed next.
- Color. The discussion above assumes light is monochromatic with a wavelength of 500 nm. This is greenish light. But the author (and the reader?) was imagining white light when visualizing this calculation and its effects. So we would actually have to consider at least three cases:
 - White light. In this case, we would have three concentric Airy disks: one smaller in diameter for blue light, and one larger for red light. The red disk would be about 40% bigger since red light has a wavelength as long as 700 nm. If the image we are taking is in black and white, diffraction would be 40% worse than calculated. If the image were to be in color, we would see a red ring around a white Airy disk. The camera would likely have difficulty in rendering the color edge of the ring, however.
 - Adjacent points that are the same frequency as each other. This is actually a much more interesting problem than it would appear, because of the way that color is recorded on current digital cameras, an issue that will be discussed later in this section.
 - Adjacent points that are of different frequencies
- Digital camera technology. A digital sensor uses a Bayer filter over the photon sensor, which is a 2 by 2 array of red, blue and two green filters, one color over each pixel. It also employs an anti-aliasing filter than refracts a small fraction of the light incident on any pixel to several of the adjacent pixels, so a source of a given color that only illuminates one pixel and thus is of unknown color will also illuminate (to a lesser extent) adjacent pixels with different

filters. This filtering reduces the modulation transfer function of a sensor at spatial frequencies close to or slightly below the limit of that the number of pixels can render.

A thorough analysis of these issues would require looking at each possible combination of these factors; it would far more than double the length of this paper, and so is not attempted here.

Even a simplified analysis is lengthy.

Modeling the Interaction Between an Anti-Aliasing Filter and a Bayer Filter

We can very briefly consider a few cases. The following calculations are examples of what physicists call a Fermi problem because Enrico Fermi used to write them for his students. It involves a problem in which the goal is to get the underlying structure of the solution correct while essentially making up the input data. A good solution will illustrate that even if ones guesses on the data are wrong (within limits of course) the answer remains the same. In this case, I will show that one can derive a reasonably close estimate of how an anti-aliasing filter affects resolution near the diffraction limit by showing that some superficially plausible assumptions yield unreasonable results.

Since I do not have enough data on how anti-aliasing filters work quantitatively, I will make three alternate arbitrary assumptions. I will show that only one of them makes any sense, and then continue the analysis using the plausible assumption.

An anti-aliasing filter is intended to be a low-pass filter that screens out frequencies that are above the Nyquist limit in order to reduce or eliminate aliasing. It is also used to determine color for small areas; this is a point I will get back to later. It is hard to imagine how a spatial low-pass filter could be designed other than the hypothesis generated by looking for physical descriptions on-line that refer to birefringent crystals. A birefringent crystal has two different indices of refraction, based on polarization of the incident light, and thus takes a single incident light beam and breaks it into two (or three) components that are refracted at different angles. It is easy to imagine designing a filter (or pair of filters) that sends some of the light that would ordinarily be incident on a given pixel to the two adjacent pixels. Two of these filters crossed at 90 degrees would send light to the four adjoining pixels.

Such a filter would reduce or eliminate signals at the Nyquist frequency, which corresponds to one point incident on every other pixel. So this is what the following derivation assumes.

Let's start with three possible assumptions about the anti-aliasing filter. The first assumption I will call a "strong" filter": it sends $1/6$ of the light incident on each pixel to each of the four adjacent pixels, leaving only $1/3$ to be recorded at the central pixel. It is easy to see that it will render a grid of points along a line at every other pixel as an indistinguishable grey blur, since all pixels in the line will see an intensity of $1/3$. This appears to be a good design, since it eliminates signal at and above the Nyquist frequency.

The “intermediate” filter is assumed to send 10% of the light to each adjoining pixel and retain 60% at the central point and sends 10% to each adjacent pixel.

The “weak” filter sends 5% of the light to each adjoining pixel and retains 80% at the point of incidence.

The analysis turns out to depend critically on color and the Bayer filter methodology used to measure color, and produces some surprising results.

We start out by assuming highly colored incident light. Color is a complex issue, because it is not the same thing as wavelength of light. *All* pure wavelengths are a mix of at least two primary colors, and all but three particular wavelengths are a mix of all three²⁰. Obviously any mixture will therefore contain all three colors. So we will assume an incident green light that is 70% green, 20% blue, and 10% red, and an incident blue light that is 70% blue, 20% red, and 10% green. These are intense colors—the green light appears actually to be out of the Adobe RGB and sRGB gamuts.

I constructed a simple spreadsheet that analyzes the combined effect of the anti-aliasing filter and the Bayer filter. I started by looking at a rectangular grid of point sources at the same intensity and color, spaced two pixels apart in both the x and y dimensions. This corresponds to the Nyquist spatial frequency.

The strong anti-aliasing filter spreads this image apart so that in both horizontal and vertical directions, all points register an intensity of 1/3 the incident level. So the image is completely greyed out: the filter appears to do its job of cutting off frequencies starting at Nyquist.

But it isn’t that simple. First, because the diagonal points receive no signal. So the input signal at points (0,0), (0,2), (2,0), (2,2)... becomes a solid grey along odd vertical and horizontal axes but has point of black at (1,1), (1,-1), etc., an ugly artefact.

When you consider that each of these pixels is a different color, the artefacts get far worse. If the green light is incident on green pixels, the main pattern is correctly rendered in green, but the odd pixels display noticeable red and blue fringing on the lines.

So for the following pattern of incident green light—

1	1	1	1
1	1	1	1

²⁰ Mark S. Rea, ed. The IESNA Lighting Handbook, Reference and Application. New York, Illuminating Engineering Society of North America, 2000.

p.4-6 displays 1931 CIE chromaticity diagram, which is most useful for small angles, showing that light at 380 nm, 504 nm, and about 545 nm are the only wavelengths that have a zero for one of the color coordinates.

1 1 1 1

The sensor registers—

0.233333	0.066667	0.233333	0.033333	0.233333	0.066667
0.066667	0	0.033333	0	0.066667	0
0.233333	0.033333	0.233333	0.066667	0.233333	0.033333
0.033333	0	0.066667	0	0.033333	0
0.233333	0.066667	0.233333	0.033333	0.233333	0.066667

where the numbers refer to light intensity on the pixel and the colors identify the color of the pixel with darker shades corresponding to more intense light, and really weak light left white. (The cells at 0.033 are weak red).

The “low-pass” filter, while correctly recording green points where they should be, adds some pretty bright noise in color: the blue intensity is 30% of the green intensity, and the red intensity (not colored in above) is 15%. Instead of wiping out all signal at the Nyquist frequency, the filter instead passes the signal of the holes in the grid while generating color artefacts.

These cannot be post-processed away, since the characteristics of the incident image cannot be inferred with any specificity from the recorded image. The color fringes may be due to the filter, but they may not, since the incident image could indeed be a grey or green blur along the x and y axes, or it may be a set of dots an any frequency up to or above Nyquist.

But what’s really weird is that if we displace the incident image by one pixel, the green points *stay in the same place* (which is now the wrong place compared to the image) at the same intensity; however new diagonal bright green pixels appear.

The incident green signal (now incident on non-green pixels)—

1	1	1
1	1	1
1	1	1

Produces the following image:

0.233333	0.066667	0.233333	0.033333	0.233333	0.066667
0	0.233333	0	0.233333	0	0.233333
0.233333	0.033333	0.233333	0.066667	0.233333	0.033333
0	0.233333	0	0.233333	0	0.233333
0.233333	0.066667	0.233333	0.033333	0.233333	0.066667

This is really ugly: all of the points register where there should be nothing; the actual location of the points shows up as near zero and the wrong colors.

But the situation is really worse than that.

Unless the grid is perfectly aligned horizontally, the image will eventually switch from one of these patterns to the other, generating moiré patterns that even have false color.

So I tried an example of slightly diagonal lines.

The incident green image is—

1	0.8	0.65	0.5	0.35
	0.2	0.35	0.5	0.65
1	0.8	0.65	0.5	0.35
	0.2	0.35	0.5	0.65
1	0.8	0.65	0.5	0.35

The rendition of it is—

0.033333	0.21	0.066667	0.169167	0.033333	0.134167	0.066667	0.099167	0.033333	0.040833
0.233333	0.006667	0.233333	0.024167	0.233333	0.028333	0.233333	0.019167	0.233333	0.021667
0.066667	0.21	0.033333	0.169167	0.066667	0.134167	0.033333	0.099167	0.066667	0.040833
0.233333	0.03	0.233333	0.018333	0.233333	0.019167	0.233333	0.028333	0.233333	0.005833
0.033333	0.21	0.066667	0.169167	0.033333	0.134167	0.066667	0.099167	0.033333	0.040833

So what should look like a set of three points at vertical positions 1,3,5 and then shifting gradually to 2,4,[6] instead is just a jumble with lots of green and other-color images at seemingly random places.

I ran a few cases where the image is not centered on an integer-number pixel and found yet other sets of odd color fringes and displaced central images.

I also modeled a case with intense blue light, and found similarly disturbing results.

The patterns have no apparent structure. The variation of intensity from one pixel to adjacent ones is so large and unstructured that it is difficult to imagine how an image processing system could distinguish signal from noise.

And this is for the simplest of possible images: a rectangular grid that is centered on a pixel and aligned with the x and y axes. If you can't reconstruct this image, or filter out the noise, you can't do it for any image.

We see an interesting irony here. By trying to eliminate signal at the Nyquist frequency completely, we filter out the very information we need to reconstruct the signal without generating new artefacts. When the incident signal after the anti-aliasing filter is a featureless grey, we lose the ability to infer either intensity or color. We will see in the final case analyzed—the weak filter—how this information can be reconstructed if some of the signal is allowed to pass.

The medium strength filter does not do much better than the strong filter.

For the second case above, the output is—

0.14	0.12	0.14	0.06	0.14	0.12
0	0.14	0	0.14	0	0.14
0.14	0.06	0.14	0.12	0.14	0.06
0	0.14	0	0.14	0	0.14
0.14	0.12	0.14	0.06	0.14	0.12

It is still impossible to detect what the incident image is in this case, and in most of the cases I examined, even those that use a fundamental frequency of 20% lower than Nyquist.

The weak filter is largely free from these artefacts.

For the relatively more difficult case of an incident image 20% off-center from the horizontal pixels—

0.8	0.2	0.8	0.2	0.8	0.2
0.8	0.2	0.8	0.2	0.8	0.2
0.8	0.2	0.8	0.2	0.8	0.2

the registered image puts the points in the right place—

0.233333	0.066667	0.233333	0.033333	0.233333	0.066667
0.053333	0.046667	0.033333	0.046667	0.053333	0.046667
0.233333	0.033333	0.233333	0.033333	0.233333	0.033333
0.033333	0.046667	0.053333	0.046667	0.033333	0.046667
0.233333	0.066667	0.233333	0.033333	0.233333	0.066667

Even in the cases in the spreadsheet (not reproduced here) where the maximum point in the image is displaced from where it should be, the relationships of the colors of adjacent pixels would seem to permit the creation of an algorithm that can reconstruct the original image by knowing the color and compensating for intensity reductions due to the color filters when, for example, an incident highly green signal is received on a red pixel.

This can occur because of the high ratio of direct light to light that is deflected to adjacent pixels. If all the pixels have approximately equal intensities of received light, it is impossible to distinguish between deflected photons and ones truly incident. But if pixel (1,1) receives a low intensity that is about the amount that would be calculated given *a priori* knowledge of the characteristics of the filter, it is possible to assume that the light is almost all due to the filter.

Also, if the main or brightest point is known to be receiving 80% of the light, we know that its intensity is NOT affected very much by light the filter has sent from adjacent points.

One can imagine an algorithm that determines the color of each pixel by first trying to average the color received at that site weighted by the reciprocal of 80% with the average of the four adjacent pixels weighted at the reciprocal of 5% each. That color could then be used to estimate what the absolute (before considering color) intensity for that pixel should be. Such a calculation would seem to be able to be subject to self-consistency tests and usually work out to the right answer.

If this discussion seems a little abstract, consider the following pattern--

0.07	0.16	0.07	0.08	0.07	0.16
0	0.07	0	0.07	0	0.07
0.07	0.08	0.07	0.16	0.07	0.08
0	0.07	0	0.07	0	0.07
0.07	0.16	0.07	0.08	0.07	0.16

What does this represent?

Clearly the points in light blue are local maxima, so they represent points.

What color are they? Since they are surrounded by green at an intensity of almost half of that of the maximum, and furthermore the diagonals are at 0, evidently the AA filter is deflecting a strong green component. So these points are mostly green with a little blue (and possibly some red, but we don't know).

What about the weaker maxima in red? They are likewise surrounded by green at nearly the same intensity, and the diagonals are also at 0. So likewise it is clear that the weaker maxima are only weaker because the color is weak in red. Thus they are interpretable as real maxima and not as noise.

So a simple algorithm that determines color from examining the closest 4 points could correctly place the image in the right pixels and correctly specify both colors (both those indicated in red and those indicated in blue) as mostly-green light with some (respectively) blue and red.

This is in fact the incident pattern that generated this output in the model. So the perceived color would at least be close, and the perceived intensity also about right. And the position would be right.

This argument, coupled with examination of the other simulations, suggest that even the 80/5/5/5 filter may be a little too strong.

This simulation exercise indicates that for anti-aliasing filters to work at suppressing artefacts in color as well and black-and-white, they must be relatively weak. Also that such weak filters allow lower-noise estimation of color on a pixel by pixel level.

This analysis is consistent with information on how Nyquist limits affect motion picture photography. One presentation on this issue explained that movie cameras employ strong anti-aliasing filters to eliminate almost all of the information above the Nyquist frequency²¹. It off-handedly noted that still cameras allowed higher frequency data to pass. This calculation explains why. Movie cameras do not employ a Bayer filter. Instead each pixel has a separate system for recording each primary color. So the color artefact problem analyzed above is not an issue: a strong filter can work.

These weak filters will allow some frequencies above Nyquist to pass through, whether we like it or not. This will enable the following simulation of how a high frequency incident image that is diffraction limited at the Rayleigh limit would be resolved.

Consider first a two dimensional grid of points that are barely resolved. For this case we will assume white light, containing about the same fraction of each primary color. Assume that the first point is centered on pixel (1,1) and the second at pixel (1,2.7) the third point is at (1,4.4). The points are 1.7 pixels apart, so we are trying to resolve one line pair using 1.7 pixels, or the equivalent of $4000/1.7$ or 2400 line pairs on a 24 megapixel sensor, almost at the limit measured in camera tests. This is a frequency above Nyquist by about 15%. We will ignore the vertical dimension for simplicity.

Without diffraction or anti-aliasing, pixel (1,1) would record 1 unit of light, pixel (1,2) would record .3 units, and pixel (1,3) would record .7 units. Pixel (1,4) would record .6 units and pixel (1,5) would record .7. Thus we would see the second point as distinct from the first, with *very* low contrast. (Distinguishing the second point from the third is problematic in this example, and becomes more so in the proceeding, so we will ignore it here. (The fact that it is problematic is a consequence of the signal's frequency being above Nyquist.)

With diffraction at the Rayleigh limit, the Airy disk would have a radius of 1.7 pixels. The first point would record that fraction of the energy of the disk that is within a radius of .5 pixels

²¹ See footnote 19

(plus the corners) from the center. This is a very difficult calculation to make, so I will assume that 80% of the light makes it into pixel (1,0) and the other 20% is spread among pixels (1,2) and the other three neighbors. Pixel (1,1) receives some light from point 2, but less—assume it is .02 units; it also picks up the same amount from the left, so it is now at, say, .84 units of light.

Pixel (1,2) is now up to .35 units--.3 from the point source and .05 from diffraction from point (1,1). This pixel also receives diffracted light from the second point source—more of it proportionally. So perhaps it is at .43 units. Pixel (1,3) loses a lot less of its light from diffraction to the right, and also picks up some light from the second maximum of point 1's Airy disk. I assume here that it drops only about .02 to .68 units.

Now let's consider the effects of the anti-aliasing filter. The light detection of pixel (1,1) is now down to $.8 \cdot .84$ or .68; and pixel (1,2) picks up .04 to a level of .47. And pixel (1,1) picks up .03 of light from each of points (1,2) and (1,0) due to the filter. But the filter also reduces pixel (1,2)'s own light level by 20% of .42, so it is at $.8 \cdot .42 + .04$, or .38. On the other hand, it picks up about .03 from the right to end up at .41. Pixel (1,3) loses 20% and is at $.8 \cdot .68$ or .54 plus whatever it picks up from the right. Pixel (1,4) is still at about .6.

So without considering the anti-aliasing filter, the first three pixels record .84, .43 and .68 units of light respectively. After the filter, the results are about .71, .41, .60. In both cases the first two points are weakly distinguishable, but the contrast is considerably lower with the filter's effects included. The contrast difference between the second and third point was .25 before the filter and .19 after, a reduction of an already-low MTF by one third. This is a significant degradation of sharpness but not apparently large enough to render a distinguishable detail invisible. The filter has not affected sharpness that much.

The point of this calculation (which is for white light) is that the anti-aliasing filter reduces contrast and may even take barely distinguishable points and make them indistinguishable, but it will not change the effect of diffraction on the absolute limit to digital resolution by factors like 1.4, but rather by something more in the neighborhood of 10%.

Another interesting point is the importance of the implicit assumption that the source points of light are equally intense. If point 2 were much dimmer than point 1 in this calculation, we might no longer be able to distinguish them. If point 2 were brighter, the difference would be disproportionately larger.

And looking at the other choices of images we might want to study, it is evident that the effects of sensor design are to reduce the visibility of diffraction limits by a quite small amount, smaller than the error of other approximations in this derivation.

If we repeat the calculation above with 2-pixel separation, the anti-aliasing filter reduces contrast (MTF) by only a moderate amount. Pixels (1,1) and (1,3) would be the centers of their Airy disks and the effects of the filter; so pixel (1,2) would gain brightness from both effects but remain much dimmer than 100%. The anti-aliasing filter would reduce the brightness at the centers by 20% and increase the intermediate point by 10%; this reduces the MTF by about

30% beyond what is lost to diffraction. Again, the filter has had only a modest effect on perceived sharpness. And this would be less if my hypothesis about the value of very weak anti-aliasing filters is correct.

One other trend is interesting from this type of calculation. We are trying to see small differences in light level between adjacent pixels in order to resolve closeby points. If the signal to noise ratio is small, these differences will be lost in the noise. So this implies that in the deep shadows, we will lose resolution due to both diffraction and anti-aliasing regardless of how the sensors or the RAW algorithms work.

Other sizes of film or sensor

For other sizes of “film” the diffraction limit varies linearly with sensor dimension, as seen next. As the linear dimension of the sensor increases, the focal length of the normal lens increases proportionally. This means that the aperture diameter increases linearly, and the angle limiting resolution is inversely proportional. Thus if we double the size of the sensor, the number of lines that can be resolved in one dimension doubles also.

Smaller sensors for interchangeable lens cameras

The pixel size for a smaller sensor (roughly 15 by 22 mm, with a little variation bigger or smaller by manufacturer) is about 20% smaller, so diffraction effects begin to be visible at about f/8 instead of f/10. And because of the smaller sensor size, the number of pixels at the same number of lines of resolution is about 40%; this suggests that the ~25 megapixel level of diminishing returns is equivalent to about 10 megapixels, a level that is already exceeded by several models.

As we compare camera lenses, whether they were designed for film cameras or digital, we find that the optics designed for smaller cameras offer higher resolution (in lines per mm) than those for larger cameras, so a film area that is 4 times bigger is less than 4 times sharper. But this is not true for small-sensor camera. In fact, most of the lenses available for these products are designed for and interchange with full frame sensors.

So the higher pixel density on smaller-sensor cameras takes us into new territory for exploring the limitations of the lens relative to the sensor. Reviews of the first 15-megapixel cameras have begun to point this out, showing higher levels of resolution in some cases²² but failing to see much advantage to the finer screen in most cases. In my experience, having recently bought the camera in question, a Canon 50 D, the extra pixels do result in sharper pictures if I watch the aperture, but I am using sharper lenses than were used in the reviews.

Pocket cameras

Typically the sensor size is 1/5.5 that of 35mm film, thus the diffraction limit for 2000 lines of clear, sharp resolution or 25 megapixels is f/1.8 instead of f/10. If we reduce the number of

²² See <http://www.luminous-landscape.com/reviews/cameras/50d.shtml>

lines to correspond to the current sensor resolutions of 10-12 megapixels, the diffraction limit increases to $f/2.4$. Since no current small-sensor camera has a lens faster than $f/2.0$, and only one product is faster than $f/2.8$, it means that small high-megapixel cameras are always diffraction limited and that megapixel counts much above about 12—which is currently offered on a few top of the line cameras—are almost pointless. This observation explains why typical small cameras do not even allow f -stops smaller than $f/8$: at $f/8$ the diffraction limit is 450 lines, corresponding to about 1.5 megapixels.

The author has also observed this effect: pictures taken at $f/8$ are visibly, disappointingly less sharp than those taken at wider apertures. I have started using a pocket camera with manual override to assure that I use apertures wider than about $f/5$, and preferably much wider, whenever possible. This corresponds to about 3.5 megapixels, which is less than the sensor resolution of 7 megapixels, and looks that way, since enlarged photos start to look blurry before they look pixilated.

Large and Medium Format Cameras

A 4x5 (inch) film camera has about 4 times the resolving power (in terms of diffraction limit) of a 35mm camera at the same f -stop. But most large format cameras have slow lenses and are used stopped down for depth-of-field reasons, so this increase usually is not fully realized. So if we wanted our 4x5 camera to perform well at $f/32$, we would only get 25% greater resolution than a 35mm camera at $f/10$. We would need only about 50% more pixels than for 35mm to perform at the same level.

No one yet produces a commercial 4x5 digital sensor. The largest size commercially available is 36mm x 48 mm. This is about 1.4 times the size of a 35mm sensor. So it could make use of about twice the megapixels as 35mm. The densest sensor at this size has 60 megapixels (*as of Aug. 2009*). Note that this again limits the f -stop to $f/8$ to make use of this pixel density.

Vision

As mentioned in the discussion on photons, most of this discussion is about wavelengths of light and lens and sensor dimensions. So it also applies to human and animal vision. Diffraction would appear to impose a hard limit on how eagle-eyed an eagle can be.

III. Conclusions

A. Physics conclusions

Technology today is approaching or even at the limits of resolution, film speed, and image size imposed by the laws of physics for cameras in the 35mm format or smaller. Further significant improvements in speed or resolution will require the use of larger sensors and cameras.

Thus today's equipment is sufficiently powerful to allow a physicist to see directly—without the need for instruments and calculation/interpretations beyond the camera itself and a computer monitor-- the effects of both the photon nature of light and the wave nature of light. A single picture taken at $f/22$ and ISO 3200 will display uniform unsharpness unrelated to motion blur or focus error and the noise patterns of each primary color as well as overall brightness that reflect the statistical nature of photon detection.

B. A Photographer's Conclusions

Currently available products are pushing the limits of what is physically possible in terms of resolution and low-light performance. Therefore a photographer must relearn how to make choices about what size of camera to use for what assignment, and what tradeoffs to make when shooting concerning aperture and depth of field and concerning film speed versus image quality. Past experience and rules of thumb developed in the film age will give results that are qualitatively and quantitatively different than what digital experience is teaching us.

1. Bigger equipment size/sensor size means better pictures

As equipment pushes closer to the diffraction limit, image quality is directly proportional to image size. Equipment size scales with image size: by simple dimensional analysis, the weight of equipment should scale as the third power of image size. Better pictures require heavier equipment to an even greater extent than was true in the past.

This is not to say that we are at the end of the line for improving picture quality in small cameras. What it means is that to improve picture quality significantly, we need to increase sensor size. The future challenge to camera designers is no longer to fit more pixels on a small sensor, but rather to design larger sensors without increasing the size of the camera. Or conversely, to hold sensor size constant but make the camera smaller.

2. Depth of field

Depth of field at a given f-stop is inversely proportional to image size. In the past, photographers got around that by using smaller apertures for large cameras: a setting of $f/64$ on an 8 by 10 inch camera produced the same depth of field as $f/8$ on a 35 mm camera, but allowed greater resolution. But with higher sensor resolution for 35mm digital cameras, both cameras are at the diffraction limit measured in angular terms. Therefore there is no real advantage to the large format. Taking advantage of the larger film size will now require larger apertures, limiting creative freedom for 8 x10 users who want more resolution in return for the weight and inconvenience.

3. What does speed trade off against?

With film, sensitivity traded off against many parameters: resolution, color accuracy, and grain being the most significant. With digital, the only tradeoff of importance is with noise. And even this tradeoff refers mainly to noise in the highlights and middle tones, where it is not a big

aesthetic problem even at ISOs of 800 or 3200 or even higher (depending on the aesthetic sense of the viewer). Looking at the performance of current best-in-class cameras, higher film speed comes mostly (or in some cases almost entirely) at the expense of dynamic range. While this tradeoff may be due in part to engineering choices made by the manufacturer, much of it is fundamental: at the highest dynamic range currently available, even at ISO 100 the noise in the shadows is pushing the limits of what is acceptable artistically.

4. Film and Digital

Digital equipment performs *much better* than film of the same sensor size, which is why a discussion of the limits imposed by the physics of light was not interesting in the past. This better performance, which to the author's eyes is a factor of 3 to 5 compared to the best 35mm slide film, means that past rules of thumb and tradeoff rules must be reconsidered. For example, a rule of thumb for hand-holding the camera used to be that the shutter speed must be at least the reciprocal of the lens focal length. Thus a 50mm lens could only be hand-held reliably at 1/50 second. But if the image *could* be 3-5 times sharper than film, then a three-times-faster shutter speed is needed. Or image stabilization even on normal and wide angle lenses.

Similarly, since depth of field is based on the smallest size of blur that could be considered equivalent to a point, it now is 3-5 times shallower. And not only that, but the small apertures that photographers used to use for large depth of field are no longer available due to diffraction. Large depth of field now comes at the expense of sharpness, or else requires taking multiple pictures with a tripod-mounted camera and combining them on the computer, or using a tilt-and-shift lens. This is even more the case for smaller sensors, where for pocket cameras sharpness demands the widest aperture available. Depth of field is no longer the simple creative choice that it used to be.

ILLUSTRATION OF DIFFRACTION AND NOISE IN THE SAME PICTURE

Figure 1. Enlargement of Figure 3, taken at $f/22$. Note noise especially in shadows on the right hand side of the buildings. Diffraction limits both sharpness and contrast.



Figure 2. Identical photo except taken at $f/8$, shows how much sharper the picture looks where diffraction isn't an issue. Note noise is the same because exposure is the same.



Figure 3. Full size picture from which Figures 1 and 2 are cropped. (Actually only Figure 1 is enlarged from this precise image since this was taken at f/22).

