The Robot's Eye

Raul Rojas

Institut für Informatik
Freie Universität Berlin
Takustr. 9, 14195 Berlin, Germany

1 Digital Vision

Autonomous mobile robots depend on video cameras to capture the images that allow them to perceive their environment. Although robots can be provided with electronic sensors superior in many cases to the biological ones, offices, streets, and working spaces have been designed for humans and contain many important visual cues that have to be decoded by a robot. Consider a robotic car driving along a street: without a visual system capable of recognizing the traffic signs, disaster would loom at every corner. Autonomous mobile robots must be able to see very much as we do, if they are to be useful in environments built for us.

In this section we review the structure and capabilities of modern digital video cameras. This is a necessary exercise for any serious investigation of robotic vision.

2 The Vision Sensor

There are two main types of commercial digital cameras: those based on a CCD chip (charge coupled device), and those based on CMOS sensors. Both are pixelated metal oxide semiconductors, but with important differences. CCDs are chips with a matrix array structure, each position representing a pixel in a rectangle. Photons which hit each pixel liberate electrons, which are collected in potential wells until they are harvested, that is, read by the electronics on the chip. The more light that hits a pixel, the more electrons that are collected there. CMOS sensors, on the other hand, use photosensitive resistors, distributed also in a matrix array. When photons hit the resistor, its resistivity changes and this can be measured with the appropriate circuit. The signal from CMOS chips can be read continuously, and this means that they are potentially faster than CCDs, which are read after a fixed integration time. However, CCDs provide a higher signal to noise ratio than CMOS chips. Therefore, they are usually part of high-quality cameras.

In general, a digital camera sensor is a planar matrix of detectors organized as an array. The important parameters for the array are the number of columns, the number of rows, and the size and form of each pixel. Although some chips have been built as tilings of hexagonally shaped pixels, most of the commercial cameras are based on rectangular arrays of rectangular pixels. The area of each pixel detector is an important parameter, because it sets a limit to the optical resolution of the camera. Pixel sizes available in digital cameras in 2006 were already approaching the diameter of the retinal rods, that is, around 2 microns. However, the spatial resolution of the total system is determined not only by the imaging chip but also by the optics. Commercial digital cameras are sold in a variety of image sizes, but a popular standard for real-time video applications is 640 columns by 480 rows (a 4 to 3 ratio).

2.1 CCD sensors

CCD sensors were invented at Bell Labs, originally with the intention of using them as memory components for computers. This application never really materialized, but in the 1970s CCD chips were already being used in video cameras. Fig. 5 shows a cross section of a CCD pixel element. When light falls on the element, electrons are freed from the semiconductor and are captured in a potential well. The more photons detected, the more electrons deposited. After a given integration time, each pixel can be read as an analog signal using the electrical connections on the chip.

The main problem for CCD chip applications, is that CCDs capture light in parallel, all pixels at the same time, but they have to be read sequentially. Usually, only the last row, the n-th, in the matrix array is read in one step, one pixel after the other. At the same time, all upper rows in the CCD array are shifted one place down, the first row to the second, the second to the third, and so on. In the next read step the $(n-1)$-th row is transmitted through a serial line to the camera electronics, and so on. When each pixel is read, the current is transformed into a digital value using an analog-digital converter of adequate sensitivity.

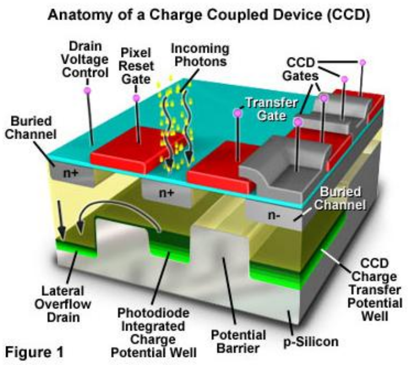

Fig. 1. Structure of a CCD pixel element

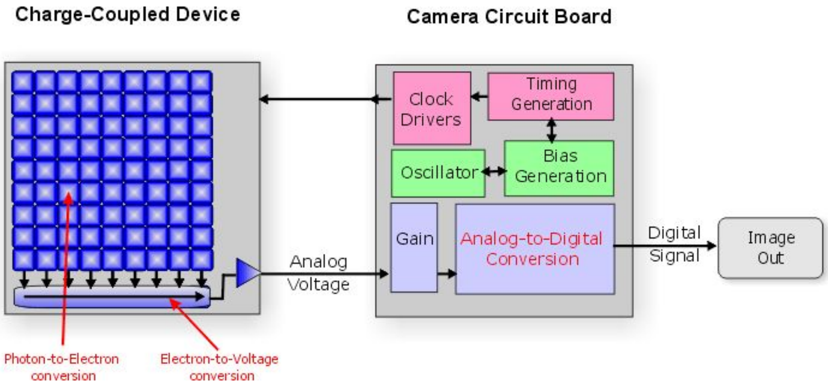

Fig. 2. A CCD chip and the read-out logic

This sequential read mode for CCD chips means that even if for an application we only need one row of information, the complete chip has to be read. Generally, it is not possible to address a single piece of data, as in RAM chips. CMOS chips, on the contrary, can be addressed in the same way as RAM chips.

There is a type of CCD chips, called interline transfer chips, in which the whole pixel array contents is read in one step to auxiliary components contained in ”interlines” (that is, arrays of condensators positioned between the CCD rows). The interlines are then read, one after the other, using transfer gates. The advantage of such an arrangement is that the CCD chip can continue collecting

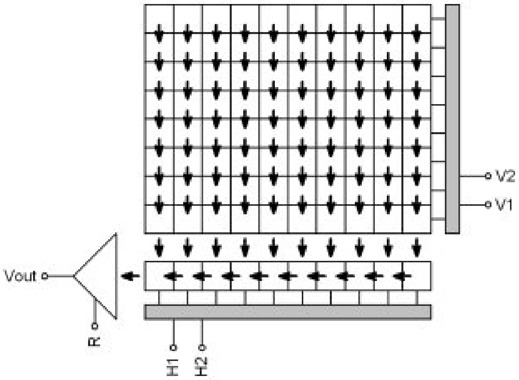

Fig. 3. Reading a CCD

light, without spending too much time in the reading process. This feature is important, for example, in astronomy.

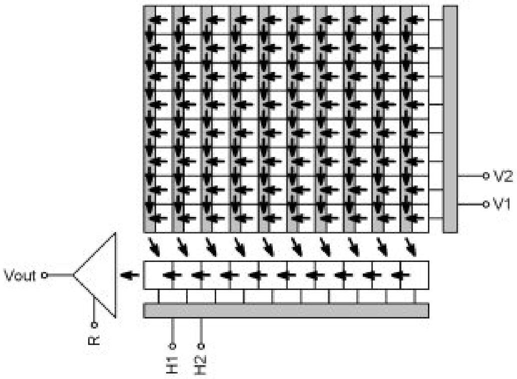

Fig. 4. Interline transfer CCD

An important parameter for CCD chips is its quantum efficiency, that is, what percentage of the photons falling on them they are able to register. CCD chips have here a similar problem as the rods and cones in the human retina: incoming light has to go through logical elements, and this diminishes the quantum efficiency of the chip. Backside illuminated CCDs collect light from the back of the chip (the light goes through a thin silicon wall), and since the logical elements are on the other side of the chip, this improves their quantum efficiency.

Due to its quantum efficiency, and because they can collect light for long periods of time with good signal to noise ratio, CCD chips are the unquestioned champions in astronomy. Fig. shows a system with 340 million pixels (40 chips), the retina of a big telescope eye.

2.2 CMOS sensors

CMOS sensors have flooded the lower tier of the digital camera market, and are improving at a rate which is making them a serious alternative to CCD chips. The main advantage of CMOS sensors is that they can be read randomly (by addressing pixels) and that the signal can be sensed continually. The electronics for handling the data stream can be etched on the same chip, so that no extra components are necessary. A CMOS camera can be built using a single chip and a lens, or simply putting a pinhole in front of the sensor. Since CMOS chips can be accessed randomly, it is possible to increase the frame rate for a small part of the image, trading off image size with frame rate.

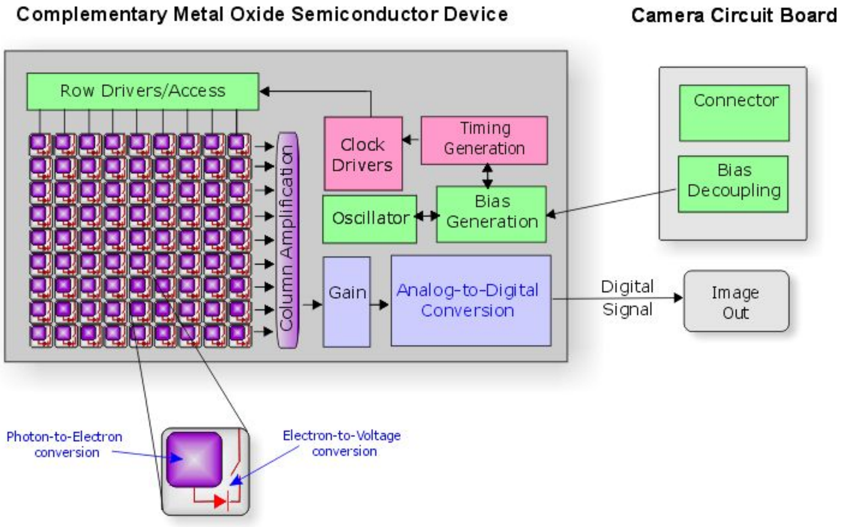

Fig. 5. Diagram of a CMOS chip

However, CMOS sensors have a lower ”fill factor” than CCD chips. This means that less area is devoted to actual photon sensing than to electronics. A fill factor of 50% means that of all the photons falling on the chip, only 50% will be measured. This disadvantage can be counteracted by etching micro lenses on the substrate of the chip. Light is then focused towards the photo sensitive part of the chip, increasing the total sensitivity.

Both CCD and CMOS sensors are sold in standard sizes given as fractions of an inch. The notation was developed for TV sensors and does not correspond to the real dimensions of the chips. Fig. 6 shows common CCD or CMOS chip sensor sizes and their corresponding side lengths in mm.

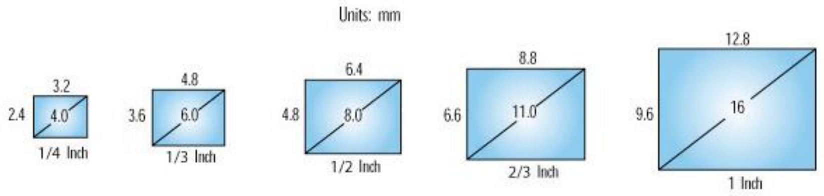

Fig. 6. Imaging chips standard sizes

CCD chips consume typically more power than CMOS chips. A 512×512 CMOS sensor consumes around 50 milliwatt of power, while CCD cameras usally get hot and are cooled for astronomy applications.

2.3 Dynamic Range

The size of the imaging chip is important because it determines, together with the lens, the angle of the visual field of the camera. There are formulas which relate chip size to camera angle.

The dynamic range of a camera refers to its capability of capturing both high and low illumination regions in the same image. The human eye has a very high dynamic range, so that we can still distinguish features inside a room when we look against the background of an open window. Digital cameras have problems with this type of images, because the shutter speed and read-out gain is set globally for the whole chip. The human retina corrects all the sensor values using a local comparison with the mean response from the receptor neighbors. The result is a logarithmic law known as the Fechner-Weber law.

In some cases, dynamic range refers to the number of bits of resolution provided by the imaging chip. CCD chips typically provide 16 bits of resolution.

Comparison between CCD and CMOS sensors

CCD	CMOS
Expensive	Cheap
Few suppliers in the world	Many suppliers
Region access	Random access to pixels
Better Signal-to-noise ratio	More stationary noise
More fill area	Less fill area
High sensitivity	Less sensitive
Large sensors available	smaller sensors
High power (can be 1000 times more)	low power
16 bits per pixel	8 bits per pixel
complex interface	simple interface
no Regions of Interest	ROI
complex read electronics	A/D conversion possible
	camera on a chip
homogeneous sensitivity	heterogeneous sensitivity
fill factor 100%	fill factor ¡ 100%, but microlens

3 The camera optics

3.1 Lenses

Common lenses are not perfect optical systems. Ideally, if we have a binary grating (alternating white and black bars), this arrangement should be projected faithfully by the optics on a film or chip. In reality, such a grating is projected with more or less diffuse borders. The point spread function of a lens tells how an ideal point of light is projected on a film. An ideal point transforms into a more or less blurred circle according to the quality of the lens.

Usually lenses are cut with spherical profiles. Such lenses can focus adequately around their rotation axis, and less accurately on the periphery. Different forms of imaging aberration are the consequence.

Some cameras use aspherical lenses, more precise than spherical ones. Such lenses are made of plastic, and their refraction properties are not as good as for glass.

New materials are always being sought by the optical industry which could offer the refractive advantages of glass, and the low weight and cost of plastics.

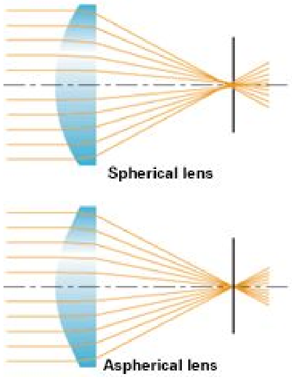

Fig. 7.

3.2 Response curve and infrared region

CCD chips are not equally sensitive in all regions of the visible spectrum. The relative spectral response curve shows how many electrons are produced in the CCD potential well by light of different frequencies. Fig. **??** shows such a curve for a certain CCD chip. Light with 600 nm wavelength produces a signal almost three times higher than light in the 400 nm wavelength region. This means that different pure colors produce different apparent luminosities in the CCD chip. This is also an effect known from the human retina, where the light receptors have different response curves for the various wavelengths.

More important for us is that CCD chips can also detect light from the non-visible parts of the spectrum, specially infrared light. Some kinds of lamps produce IR light which affects the color measurement process in a digital camera. Two colors which look identical to the human eye look different to the camera because of the IR component. To avoid this problem, the optics of most cameras has an integrated IR filter which cuts off the non-visible portions of the light spectrum. Some cheap video cameras do not have this filter and are difficult to calibrate for robotic applications in indoor scenarios.

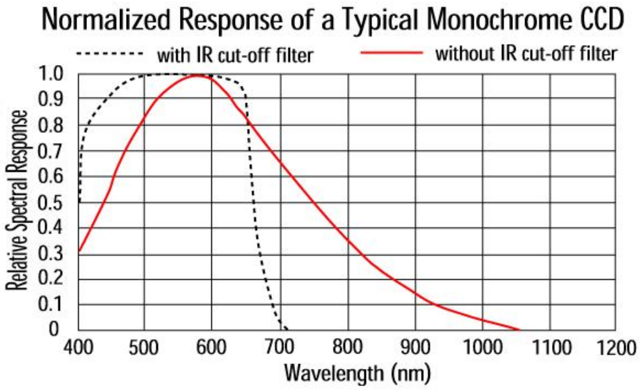

Fig. 8.

3.3 Modulation Transfer Function

With the proliferation of Megapixel digital cameras, one could think that the number of useful pixels for computer vision applications is growing exponentially while the cost of the systems is falling also at a similar rate. There is however one important component in computer vision applications where the price has not been falling as steadily, and which in many cases limits the effective resolution of the camera: it is the optics.

Lenses are not perfect, they are a compromise between manufacturing constraints and cost. Even if the imaging chip behind a system of lenses had infinite resolution, the effective resolution would be limited by the diffraction limits of the optics. Such limits are summarized in the Modulation Transfer Function (MTF) supllied by providers of optical systems for their products. Given an image with more and more parallel lines per mm, the MTF is a curve which tells us how much of the contrast is transmitted to the imaging substrate (film or chip). 100% transmission would mean that a binary signal of lines fluctuating in intensity between 0 and 1 is reproduced exactly. 50% transmission means that the white bands have become 50% gray, due to interference between the black and white lines. At 0% transmission a white and black grating is not visible anymore and appears as a gray texture.

The MTF curve is very important for such applications as OCR, where the recognition of fine details is crucial. The total MTF of a complete camera is the product of the MTF for the individual components: the optics, the imaging chip, the analog/digital converter, and some more. Fig. 14 shows the MTF for a real lens from Edmund Optics.

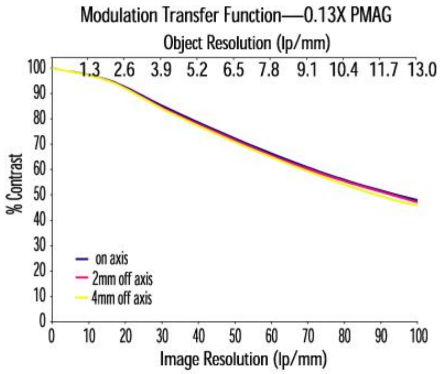

Fig. 9. Modulation Transfer Function for a real lens

The MTF curve is precisely the problem which afflicts cellular telephones with multimegapixel chips, but cheap lenses with low contrast transmission.

4 Color sensing

In general, the sensor chips used in color cameras can only provide a black or white picture. In order to detect different colors, color filters are needed. If a prism splits the incoming light, so that red, green and blue tones are separated, then three chips are needed. The three images in a so-called 3CCD camera can then be combined pixel by pixel to obtain a color image. The figure shows a prism used in a commercial camera for splitting the incoming light in three directions. In each direction a 410,000 pixels chip senses the image and provides a measurement for the camera.

In more conventional cameras, only one sensor is used but every pixel is covered by a different color filter, usually red, green or blue. This means that at each pixel only one of these colors will be measured, but the missing color information for each pixel can be extrapolated from its neighbors. The figure below shows the

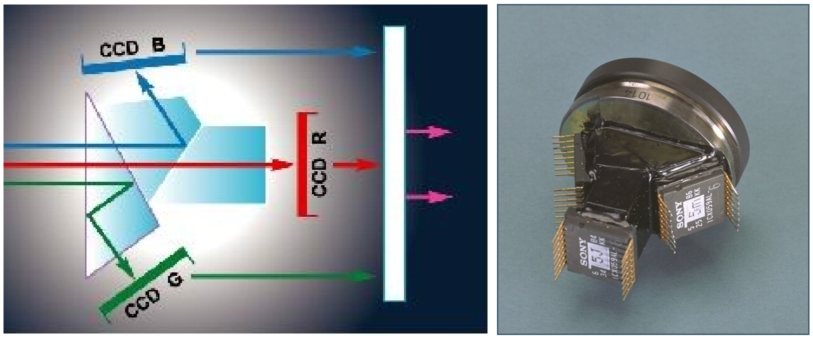

Fig. 10. Diagram of a 3CCD image splitter

arrangement for a CMOS camera. At each pixel there is the electronics needed to measure the incoming light. A photodiode transforms light quanta into a charge, which is collected in a potential well. A microlens above the pixel concentrates the light on the electronics. Below the microlens there is a red filter which lets red light through and absorbs other frequencies. This is a pixel ”specialized” in measuring red light.

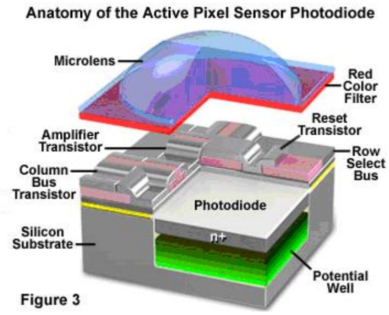

Fig. 11. CCD pixel with a microlens

Of course, not all pixels can be specialized to measure red light. Like in the retina, some pixels are covered with a read, and others with green and blue filters. The picture below shows an arrangement for a Fuji camera. A layer of microlenses is placed above a layer of filters, which in turn cover a layer of pixel sensors. In this special case the sensors are distributed in an hexagonal array.

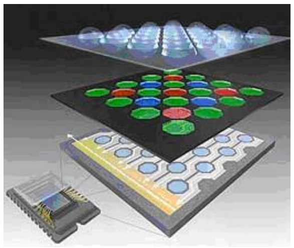

Fig. 12. Modulation Transfer Function for a real lens

More typical for video and digital cameras is a distribution of pixels as rectangular arrays. A common arrangement for the color filters is shown in the figure below, to the right. This arrangement is called a Bayer Filter, because it was invented at Kodak by one Dr. Bayer. As can be seen, in the Bayer array half of the pixels have a green filter, the color for which the human eye is more sensitive. The red and blue filters are arranged in rows and columns, with one missing pixel in each direction. At the time Bayer registered his patent, he also proposed to use a cyan-magenta-yellow combination, that is another set of opposite colors. This arrangement was unpractical at the time because the necessary dyes did not exist, but is used in some new digital cameras. The advantage of the new CMY dyes is their greater light absorption characteristics, or what is called, its quantum efficiency.

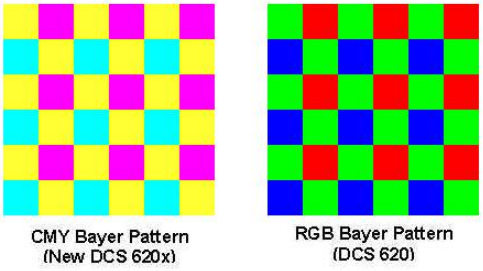

Fig. 13. Bayer and CYM filters

Once the chip has been exposed to an image, each pixel can be read. A pixels with a green filter provides an exact measurement of the green component. The red and blue components for this pixel are obtained from the neighbors. For a green pixel, two red neighbors can be interpolated to yield the red value; also two blue pixels can be interpolated to yield the blue value. There are several different methods that can be used to combine the pixel values - this process is called demosaicing. They differ in the kind of artifacts that can be produced (see below).

The graph below was produced by Kodak and shows the absorption efficiency of the different types of dyes they use. As can be seen, magenta, cyan and yellow, have better absorption characteristics.

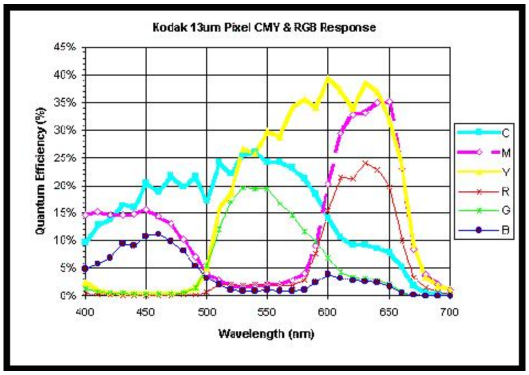

Fig. 14.

Still another popular filter is the CMYG filter, in which cyan, magenta, yellow and green filters are used. The arrangement of the filters is shown in the picture above.

As can be seen, green pixels alternate with yellow pixels, and magenta with blue. Another common filter array places columns of red, green and blue filters one after the other. In that case the horizontal resolution is a third of the vertical resolution.

4.1 Demosaicing Techniques and Color Artifacts

Once a picture has been taken, it has to be transformed in one of the standard image formats, for example RGB or S-VHS. Transforming from a Bayer filter to RGB can be done interpolating the sensor values at different places in the array.

The figure below shows the general approach. The green component of a pixel can be measured directly at any pixel with a green filter (50% of all pixels). Where the green value is missing, it can be interpolated from the four neighboring pixels with a green filter. Since the eye is more sensitive to green, more surface of the color array is used for this color. In the case of blue and red, missing information can be interpolated from the neighbors. There are two cases: the pixel has two or four direct pixels of the desired color.

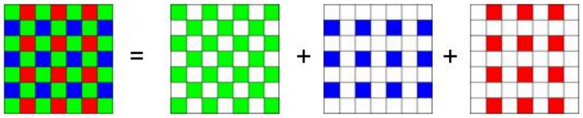

Fig. 15.

Color artifacts in the images

Color artifacts in the RoboCup images are produced by a series of different effects. The picture below shows two effects. First, due to interlacing of the video image, the even and odd lines were captured at different moments. It can be seen that there are obvious differences in shades of color. This can be due to the illumination of the room. Halogen lamps usually flicker. The flickering is almost invisible for the human eye, but it affects the digitization of the image. The second effect is the appearance of phantom colors due to interpolation in the Bayer filter. Between the ball and the white (should be white!) boundary there appears some violet. Between the green field and the white boundary we can see shades of green. The diameter of the orange golf ball is around 4.25 cm. Since each pixel measures 5 by 5 mm, the ball covers around contains around 56 pixels, of which only around half of them are of the appropriate or similar color. Therefore, small colored objects appear even smaller when only the expected object colors are selected.

Figure : False colors at the boundary of colored objects

In the next picture we can see how colors ”spill” into one another. The white patches have a blue boundary. The cyan marker (0,1,1) spills onto the green (0,1,0) marker and produces cyan (0,1,1), and the magenta (1,0,1) marker spills onto yellow (1,1,0) to produce white (1,1,1). The artifacts shown here are of the additive type.

4.2 Illumination temperature and filters

The color of objects can be perceived differently according to the spectrum of the illumination. We are not aware of this fact, because the human eye and brain excel in filtering out such perturbations, so that the perceived color of the objects can remain approximately constant. Nothing more confusing for a computer, which can perfectly "see" with its sensors that the colors reflected by objects are different according to the illumination.

Photography experts know that before taking pictures, the spectrum of the ambient light has to be determined. Incandescent and halogen lamps can be thought as systems in equilibrium, irradiating a mixture of wavelengths determined by their temperature. This is the main characteristic of a black body, a cavity in which all radiation has reached equilibrium. As shown in the picture above , the radiation escapes through a small hole in the surface. The spectrum of the black body changes according to its temperature measured in Kelvin, as shown in the diagram to the right. The temperature of illumination systems is given in Kelvin. Day light is rated at around 5000 Kelvin and has a blue tint. Early morning light has a red tint. A photographer will different kinds of filters in order to balance the light spectrum and make it appear white. This is called white balancing, and can be done by the camera itself or can be processed a posteriori in software.

The figure below shows the spectra for a tungsten lamp, a halogen lamp and sunlight. Sunlight is the more balanced of the three. The tungsten lamps tend to make the scenes appear redish or yellowish. Flurorescent lamps have a similar effect, but not as pronounced.

In the case of digital cameras, white balancing can be done by modifying the gain of each color channel. If the camera provides the computer with RGB values for each pixel, then white balancing can be done by rescaling these values. If the software is converting the values from the RGB to the HIS color space, the rescaling can be done on the fly, without losing computer time.

5 Reading the data sheet

There is special terminology which is used in the data sheets for digital vision sensors which is important to know, because it refers to the quality of the data that can be captured. Let us take as an example the data for the Philips FTT1010 sensor.

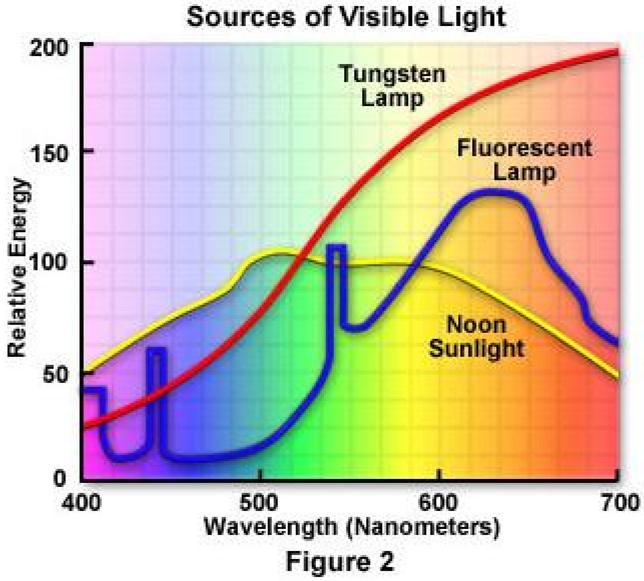

Fig. 16.

sensor type Frame transfer
format 1:1
size 12.3 mm × 12.3 mm
pixel size 12 μ × 12 μ
number of pixels 1024 *times* 1024
electron capacity 650,000 e^-
Noise 12 e^-
Dynamic 54000:1
Dark current 40 e^-/pixel/sec (at 15 C)
Quantum efficiency ¿ 30%
Antiblooming ¿ 200 times capacity
Binning Horizontal/Vertical

The format of a sensor is given in millimeters per side. This size is important because together with the lens, it determines the field of view of the camera. The FTT1010 is a chip with a quadratic form factor. The pixel size is important for the number of photons which can be captured for each pixel. The electron capacity refers to the maximum number of electrons which can be contained in each pixel potential well. This number has to be compared with the mean noise, which in this case is 12 electrons, giving a dynamic range of 54000 to 1. Since this chip encodes the data in each pixel with 16 bits, the maximum theoretical dynamic range would be 64000 to 1, which is reduced by the noise. Dynamic range: the dynamic range of a sensor chip is measured in decibels. The dark current is thermal electron noise present in the silicon substrate. It increases with temperature and can be a problem with long exposures. The quantum efficiency refers to the number of photons which can be captured in each pixel. The quantum efficiency depends on the wavelenght, and for this particular chip is listed as greater as 30%. Anti-blooming refers to some techniques used in CCD chips for avoiding that electrons spill from one pixel potential well to its neighbors. This can happen if the illumination is excessive. In this case the antiblooming construction can cope with overexposure of 200 times the pixel well capacity. H V binning - from bins, subsampling

Some other important terms used in data sheets for vision sensors are:

Color mask refers to the pattern of color filters used for measuring colors

Progressive scan The term progressive scan refers to the ability of a chip of capturing a full image at time. Video cameras for TV reproduction capture even and odd lines at different times, interlocked. This means that, when an object is moving, the position of the object in the even lines will in general be different to the position in the odd lines. This is convenient for TV reproduction, because the human eye can interpolate the movement better, but confusing for robots and computer algorithms. Progressive scan provides an instant photograph of all pixels in the same time frame.

Electronic shuttering Electronic shuttering refers to the capability of a chip of starting and ending the sensing process in its pixels through an electronic command. No mechanical shutter is needed. This provides more flexibility for capturing images.

Pixel factor Square pixel: most digital cameras use square pixels, some have hexagonal pixels.

Fill factor The fill factor is the percentage of the area of the chip that is occupied by real photo sensitive elements.

Active pixels In some CCDs and CMOS sensors not all the area of the chip is used for providing a measurement. Some extra rows and columns are available on the borders (for example for ”anti shake” control). The pixels providing readable data are called the ”active pixels”.

Fixed pattern noise CMOS chips do not have a homogeneous quality on all pixels. The readings differ from pixel to pixel, for the same illumination. This is called ”fixed pattern noise” and can be eliminated from a frame, if it is measured first.

HAD Hole accumulation diode

References

1.