Quick introduction to Computer Vision

Introduction

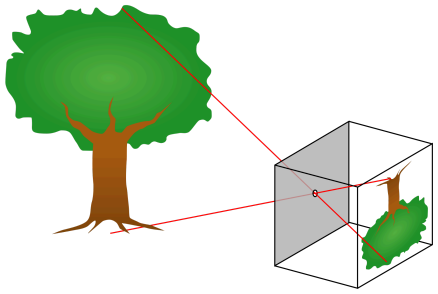
Computer vision is used in a wide range of applications. The requirements of the vision system vary significantly among these applications. For the purpose of this document, I will focus on the vision systems used on FRC Robots.

Components

Lens
Image Sensors

Lens

The lens focuses light to form an image on the plane of the image sensor. A simple lens can be created by poking a very small hole in a piece of thick paper. Light passes through the small hole to form an inverted image on a plane on the opposite side of the paper.



In most respects, the pinhole lens performs its task perfectly. Because each ray of light passes through the very tiny hole, there is only one path for it to follow, and the resulting image is very crisp regardless of the distance between the lens and the image plane or the object.

The main disadvantage of the pinhole lens is that it doesn't allow much light to pass through it onto the image plane. Any scene that isn't very well illuminated appears as a very dark image on the image plane. Increasing the brightness of the image involves a number of tradeoffs.

By increasing the size of the hole, or the aperture, the amount of light passing through the hole can be increased. Unfortunately, the larger hole allows rays of light from the scene to take multiple paths through the hole, so the image starts to become blurry.

Most lenses that you might have encountered use curved pieces of glass to focus the rays of light. These lenses increase image brightness but introduce other limitations:

- Reduced depth of field. The pinhole lens is capable of generating a clear image of any scene, no matter how far or near objects are to the front of the lens. For lenses with a larger aperture, only objects within a range of distances to the lens will be in focus. This range is called the depth of field.
- Distortion. The light passing through a pinhole lens is not bent or distorted by the refraction of lens elements, so the resulting image is a near perfect representation of the scene. Lenses with glass elements, on the other hand, cause the light to be refracted which leads to distortion. Barrel or pincushion distortion.

Image Sensors

The image sensor is responsible for converting light into an electrical signal. There are two main technologies for converting an image to electrical signals: CCD and CMOS. Both technologies are based on a grid of photosensitive pixels. You can think of each pixel similar to a water well. At the beginning of the imaging process, the wells are drained of any electrical charge. When the exposure, or integration, period begins, the drain is turned off, and light hitting the cell is converted to an electrical charge that accumulates in the well. At the end of the integration period, the amount of accumulated charge is measured. The way that this charge is measured varies between the two technologies:

- CCD Sensors typically have one, or a small number of read-out circuit(s). Each of these read-out circuits is capable of measuring the charge from one pixel at a time. When they have completed the measurement for one pixel, all of the pixels are “shifted” towards that read-out circuit.
- CMOS Sensors typically have read-out circuitry at each pixel

Below are a few differences between the two technologies:

- The read-out circuitry of CMOS sensors requires space for the sensing technology adjacent to each well. This space is not photo-sensitive, and results in a “dead” area between pixels, reducing the size of the pixel and leaving less area for photons to hit the surface, and reduces the sensitivity of each pixel. CCD sensors do not require this additional circuitry and hence are more sensitive to the light striking the surface.

- In an ideal world, the read-out circuitry at each pixel of a CMOS sensor would be identical, however, in the real world, each one of these circuits will have a slightly different bias-level and gain than the neighbouring cells. CCD sensors have only one (or a small number of) read-out circuits, each pixel is measured by the same read-out circuit and will have more uniform gain and bias levels.

- The process of shifting the charge from pixel to pixel in a CCD requires more energy than the read-out circuitry, so CCD sensors consume more power.

- One of the real benefits of CMOS technology comes from the ability to "skip" pixels in the read-out to get a lower resolution image at a higher frame rate. Often this is achieved by skipping a number of rows at the start and end of the frame, and a number of columns at the start end end of each row to give the lower resolution image. Many sensors also have the capability to "bin" or combine the output from multiple pixels into a single pixel in the output image.

- Many CMOS sensors use a "rolling shutter" which significantly reduces the amount of electronics at each pixel as some of the read-out functionality can be shared between the rows. This reduced complexity makes the sensor less complex and more economical to manufacture, but causes the rows of pixels to be "exposed" sequentially. These benefits come a the expense of "rolling shutter" effects that are visible in fast moving images like pictures of a plane propeller.

Both types of sensors introduce errors or "noise" into the resulting image. The four most common types of noise are:

- Fixed Pattern noise - caused by variations in bias of each well and read-out circuit. FPN is most notable when looking at a perfectly dark image. Although there is no light going into any of the pixels, pixels will have slightly different values. For very high accuracy applications, FPN can be measured and corrected.

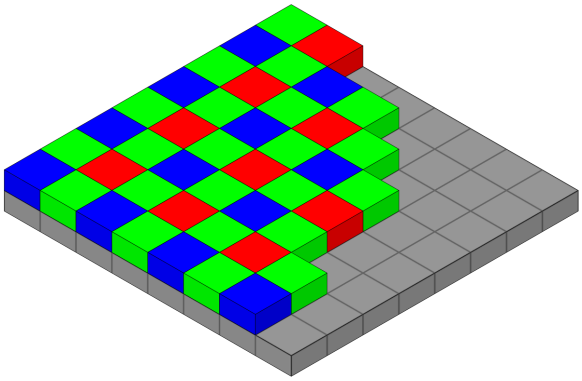
- Pixel Response Non-Uniformity or PRNU is due to the slight variation in size or light gathering capability of each pixel, allowing some pixels to gather slightly more light than others. The variability in the gain of each read-out circuit in CMOS sensors also appears as PRNU. Like FPN, PRNU can be corrected for applications requiring very high accuracy images.

- Electrical Noise - causes random noise across the image.

- Sensor Fill Noise - due to the boundaries, and in the case of CMOS sensors, read-out circuitry, the there will be small sections of the sensor that are not photo-sensitive.

Most sensors used in robotics applications are now of the CMOS variety. Sensors with a "Global Shutter" are generally preferred.

Colour vs Black & White Sensors



While it may appear that colour image sensors measure the amount of Red, Green, and Blue light at each pixel, that is not the case. In fact, colour image sensors utilize a “Bayer Filter” to block out all but one colour of light from each pixel. The Red, Green, and Blue values for each pixel in the output image are obtained by interpolating the intensities of the neighboring Red, Green, and Blue pixels. The “de-bayer” process

You will notice that there are twice as many green pixels as there are red or blue. The human eye is more sensitive to green light, so having more green pixels improves the sensor’s ability to interpolate the green values for each pixel.

Since the Bayer Filter makes each pixel sensitive to a single colour of light, colour images require more light, or longer exposure time than Black & White or “Grey-Scale” sensors.

Sensor Formats

The output from each pixel is proportional to the number of photons that land within the pixel boundary. If there isn’t enough light hitting the pixels, the image will appear dark. Similarly, if there is too much light, the image will be saturated.

For applications that require extremely high frame rates and low noise it is sometimes necessary to use specialized sensors that are much larger and have larger pixels that allow a light to strike a larger surface area. Alternatively, a larger lens with a larger aperture can allow more light to be focused on the sensor.

The sensors used for FRC robots are often ¼” or ⅓”, with pixel sizes on the order of 3uM x 3uM. DSLR cameras can take very good pictures because they have large lenses and large format sensors. These cameras can use sensors with pixels as large as 7.8uM x 7.8uM.

Exposure and Gain

Since it isn’t possible to change the size of the lens, aperture or sensor, there are 2 additional means of adjusting the intensity of an image:

- Exposure - how long the sensor accumulates light before sampling
- Gain - how much the signal is amplified

Increasing either of these parameters will make an image brighter, however, they affect the quality of the image in different ways.

For applications where the scene isn't moving quickly, increasing the exposure time will allow more light to be accumulated before being sampled. Doubling the length of exposure will typically double the number of photos captured, and double the intensity of the image. For a "still image", this will produce a high quality image with a very good signal to noise ratio, however, if the scene or camera is moving, a longer exposure time will result in the image becoming blurry.

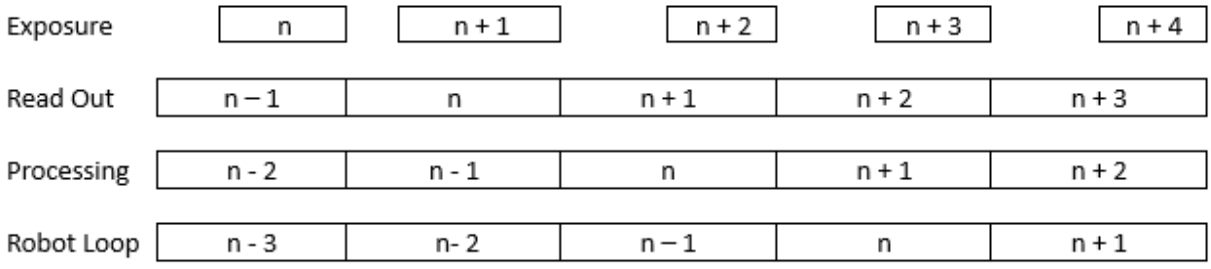
Alternatively, increasing the analog gain will amplify a smaller signal to make the image brighter. The larger gain will also amplify any noise, so the signal to noise ratio suffers.

In applications like FRC Robots tracking AprilTags, a balance between exposure times and higher gains

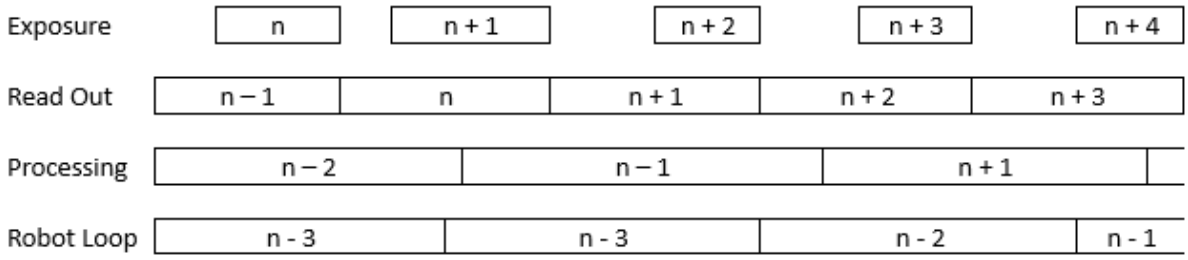
Timing and Latency

Data latency describes the amount of time that has elapsed between when data is sampled and when it is used. Video data typically has significantly more latency than any other sensors used on an FRC Robot. Some of this latency is due to the way image sensors work and also because the video data requires significant data processing that can take 10s of milliseconds. Depending on the CPU speed, it can take longer than one frame period to process the video, especially when tracking AprilTags.

In an ideal world, the processing and robot loop are well synchronized and latency is relatively short and consistent as in the illustration below.



Unfortunately, the amount of time required to process the AprilTags is often more than one frame period and the Robot Loop runs asynchronously, and at a different rate than the video system. In this case, the latency is much longer, video frames get missed, and sometimes the robot may be required to use the same data for more than a single loop as illustrated below.



There are methods to improve the latency of the data. For example, it may be better to process only 1 out of every 2 or 3 frames. This approach allows the processing to start as soon as a new frame of data is read-out from the sensors. The robot loop can be designed around knowing that new video data will only be received at the lower frequency.

If the Image Processing algorithm has knowledge of where a target (e.g. AprilTag) can appear in the image, the algorithm can “skip” the regions where there aren’t any targets. This can significantly reduce the processing time.

Sensor Calibration

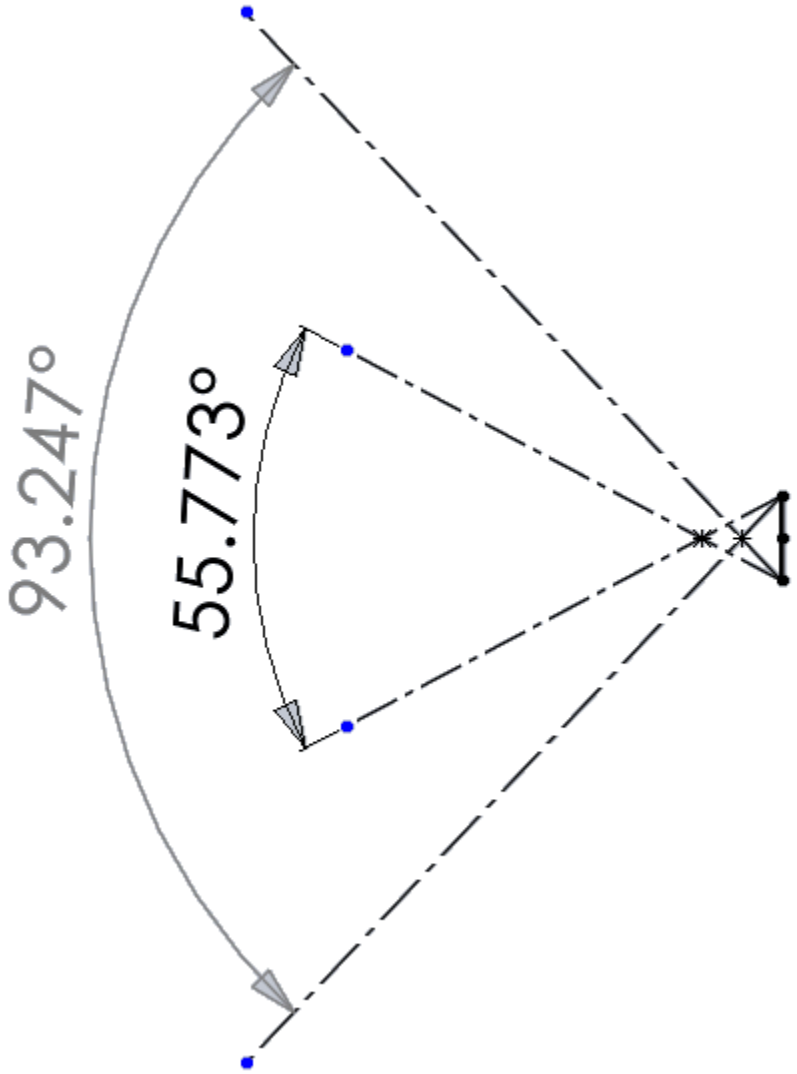
In order to obtain meaningful information from an image sensor, there are a number of parameters that must be known. The most basic parameters are often referred to as the “pin-hole” parameters. These parameters describe the relative location of the sensor to the lens:

- Focal length - the distance from the “theoretical” centre of the lens to the image plane of the image sensor. The focal length is usually expressed in terms of pixels (see Basic Sensor Math below). Since some image sensors don’t have square pixels, separate focal lengths for horizontal and vertical axis are often calculated.
- Principal Point - describes the pixel location on the image sensor where a ray of light that is perpendicular to the sensor goes straight through the lens. In most cases, this location will be close to the centre of the image sensor, but due to manufacturing tolerances, is rarely exactly at the centre.

Improving the positional accuracy of image data can be achieved by correcting for the lens distortion described above. The process typically uses a number of images of a checkerboard target that is placed at numerous locations in the field of view of the camera. Characterizing the lens and application of the lens distortion terms is outside the scope of this document but can be readily found online.

Field of View

The field of view describes the angle through which the image sensor can view the scene. The Field of View is determined by the size of the image sensor and the focal length of the lens. As the focal length increases, the field of view decreases. The image below illustrates the field of view of a sensor paired with two different lenses of different focal lengths.



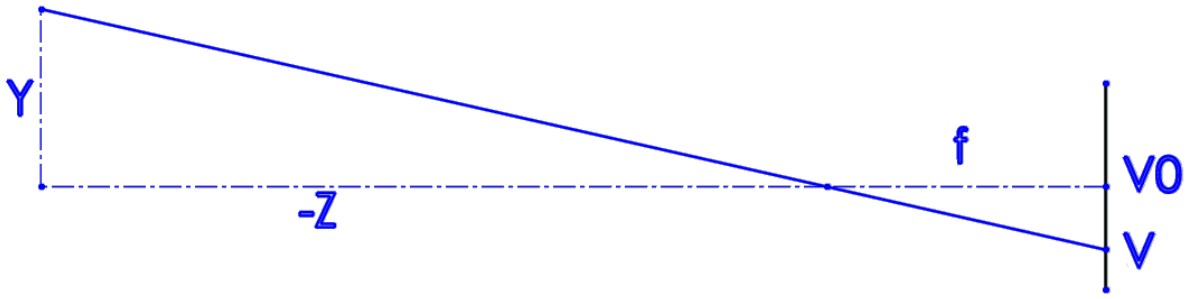
Basic Sensor Math

Calculation of where an object is located in space is achieved through basic mathematics. By determining the (u, v) location of a point on the sensor where u is the column, and v is the row, it is possible to use the principle of similar triangles to describe the location of the point in physical space. Using the pin-hole parameters described above where f is the focal length in pixels, and U0 and V0 describe the principle point, the following relations hold:

$$\frac{Y}{-Z} = \frac{V - V0}{f}$$

$$\frac{X}{-Z} = \frac{U - U0}{f}$$

The illustration below shows the relationship for the vertical axis.



Using these relationships, if for example, the height of an object is known, the Z and X locations can be computed.