Review on Distortion Generation method applied on Wireless Capsule Endoscopy

Tan Sy NGUYEN

April 1, 2021

Wireless capsule endoscopy (WCE) offers a painless and noninvasive way to identify diseases in the gastrointestinal (GI) tract. The quality of endoscopic images directly affects doctor to diagnose diseases. Due to the limited illumination of WCE system and convolution and waving natures of GI tract, the captured images often have dark areas and suffer from low contrast and severe detail loss. Therefore, image enhancement is necessary to improve the quality of endoscopic images. In this report, I would like to show us some popular types of distortions as well as how to generate them. From this study, we can implement and make a artificial distortion set which is as close as possible to reality.

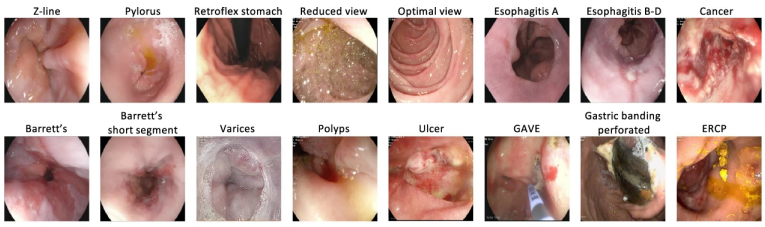
1 HyperKvasir Dataset

In the first part, I would like to describe briefly the HyperKvasir [1] dataset[1] which will be the main data used for our system. The images and videos in HyperKvasir were collected prospectively from routine clinical examinations performed at a Norwegian hospital from 2008 to 2016. The images are retrieved from the Picsara image documentation database (CSAM, Norway), a plug-in to the electronic medical record system, in 2016. As a first step, 4,000 of these images were labeled into eight different classes by medical experts and published as the Kvasir dataset22. The dataset was later extended to 8,000 images. Using Kvasir, researchers all over the world have started developing different ML models and AI systems for GI endoscopy. Note that endoscopic data is hard to retrieve from the health care systems, approvals from medical committees are hard to get, medical experts have limited time, and there are no efficient tools to label such data. Therefore, with HyperKvasir, both the amount of labeled medical data is increased for supervised learning and also a large amount of unlabeled data is released. The new dataset contains 110,079 images and 374 videos from various GI examinations, resulting in 1 million images and frames in total.

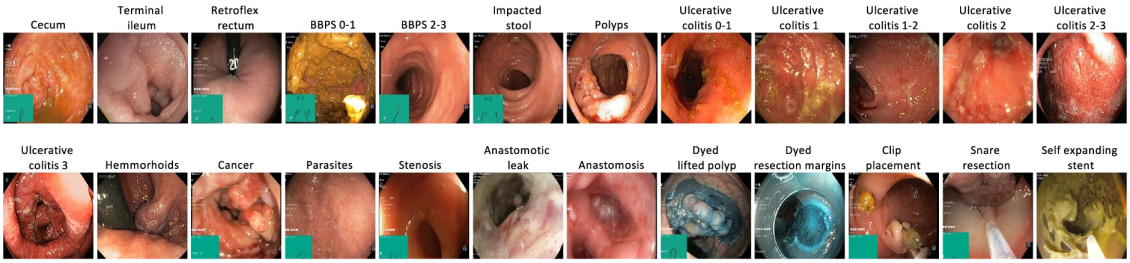
All the various labeled classes are shown in Fig. 2, i.e., 16 classes from the upper GI tract and 24 classes from the lower GI tract. In total, the dataset contains 10,662 labeled images stored using the JPEG format and more than 100,000 unlabeled images, where Fig. 1 shows the 23 different classes representing the labeled images and the number of images in each

[1] https://datasets.simula.no/hyper-kvasir/

Figure 1: Overview about HyperKvair dataset



(a) Upper GI tract.



(b) Lower GI tract.

Figure 2: Image examples of the various labeled classes for images and/or videos.

class and the corresponding resolution in Fig. 3. A CSV file is provided (image-labels.csv) giving the mapping between the image (file name) and the labeling for each image. These classes are structured according to location in the GI tract and the type of finding.

- Anatomical landmarks: Anatomical landmarks are characteristics of the GI tract used for orientation during endoscopic procedures. Furthermore, they are used to confirm a complete extent of the examination. Landmarks exist both in the upper GI tract (esophagus, stomach and duodenum) and in the lower GI tract (terminal ileum, colon and rectum). However, in the small bowel, there are no specific landmarks to be used for topographical localization of a lesion.

- Pathological findings: All parts of the gastrointestinal tract can be affected by abnormalities or findings due to disease. Most pathological findings can be seen as more or less obvious changes in the intestinal wall mucosa. These findings are classified according to the Minimal Standard Terminology, defined by the World Endoscopy Organization.

2 Distortion generation

2.1 Additive white Gaussian noise (AWGN)

Additive white Gaussian noise (AWGN) is a basic noise model used in information theory to mimic the effect of many random processes that occur in nature. Digital imaging is widely used in applications such as medical, biometrics, multimedia,...etc. In many cases, images are transmitted through Internet from one point to another. During image acquisition and transmission, factors such as moving objects, sensor quality, and channel interferences may result in additive noise. Parameters such as noise mean and variance provide noise

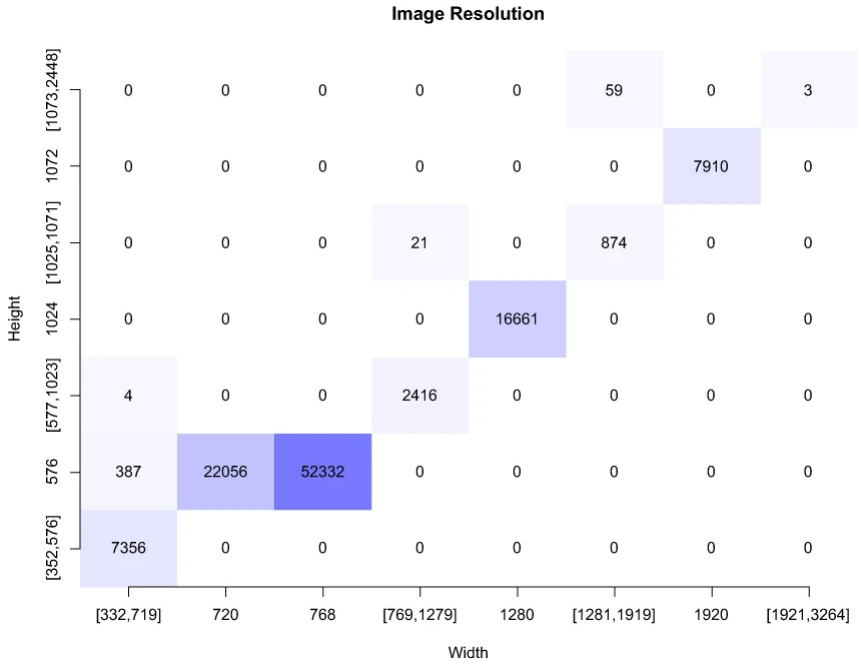


Figure 3: Resolution of the 110,079 images in HyperKvasir.

characteristics of AWGN. Image independent noise can often be described by an additive noise model, where the recorded image $f(i,j)$ is the sum of the true image $s(i,j)$ and the noise $n(i,j)$:

$$f(i,j) = s(i,j) + n(i,j) \tag{1}$$

The noise $n(i,j)$ is often zero-mean and described by its variance σ_n^2. The impact of the noise on the image is often described by the signal to noise ratio (SNR), which is given by:

$$SNR = \frac{\sigma_s}{\sigma_n} = \sqrt{\frac{\sigma_f^2}{\sigma_n^2} - 1} \tag{2}$$

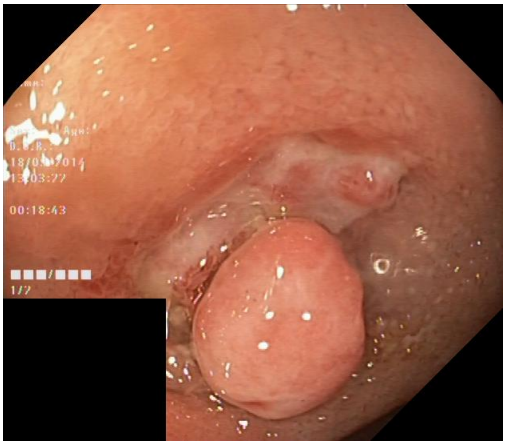
where σ_s^2 and σ_f^2 are the variances of the true image and the recorded image, respectively.

2.1.1 Result

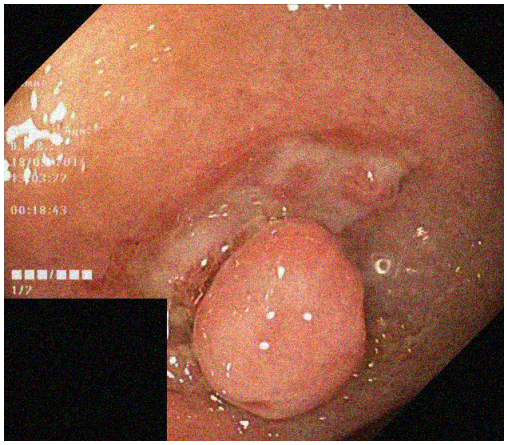
The result of image applied AWGN is present in Fig. 4

2.2 Low contrast

The whole examination process of WCE will last for 8 hours and produce about 5,000-10,000 images per person. However, the images taken by WCE system are not as distinct as those taken by traditional endoscopy due to the following reasons. Firstly, due to the volume restriction of the WCE, the battery capacity is limited. Hence, some images are not taken under sufficient illumination. Secondly, complicated circumstance of gastrointestinal tract and moving imaging method also lead to a poor image contrast. The limited sensor spatial



(a)



(b)

Figure 4: Input (a) and output (b) images showing the effect of AWGN distortion created by the above theory. Note that we can change the level of noise.

resolution and photometric sensitivity of the capsule camera does not allow the clinician to get the expected quality.

To adjust the contrast of the image, I used a simple theory to adjust the level of color of each pixel in the observed image. Two commonly used point processes are multiplication and addition with a constant:

$$g(x) = \alpha f(x) + \beta \tag{3}$$

The parameters $\alpha > 0$ and β are often called the gain and bias parameters. Sometimes these parameters are said to control contrast and brightness respectively. You can think of $f(x)$ as the source image pixels and $g(x)$ as the output image pixels. Then, more conveniently we can write the expression as:

$$g(i, j) = \alpha \cdot f(i, j) + \beta \tag{4}$$

where i and j indicates that the pixel is located in the i-th row and j-th column. To perform the operation $g(i, j) = \alpha \cdot f(i, j) + \beta$ we will access to each pixel in image. Since we are operating with BGR images, we will have three values per pixel (B, G and R), so we will also access them separately.

Increasing (/ decreasing) the β value will add (/ subtract) a constant value to every pixel. Pixel values outside of the [0 ; 255] range will be saturated (i.e. a pixel value higher (/ lesser) than 255 (/ 0) will be clamp to 255 (/ 0))

The histogram represents for each color level the number of pixels with that color level (Fig. 5). A dark image will have many pixels with low color value and thus the histogram will present a peak in his left part. When adding a constant bias, the histogram is shifted to the right as we have added a constant bias to all the pixels.

The α parameter will modify how the levels spread. If $\alpha < 1$, the color levels will be compressed and the result will be an image with less contrast (Fig. 7).

2.2.1 Result

The result of image applied low/ high contrast is present in Fig. 7

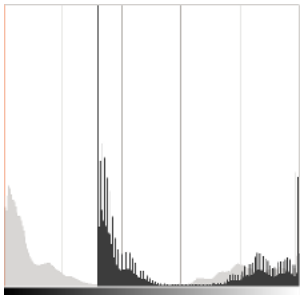


Figure 5: In light gray, histogram of the original image, in dark gray when brightness = 80

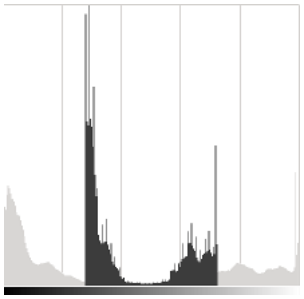
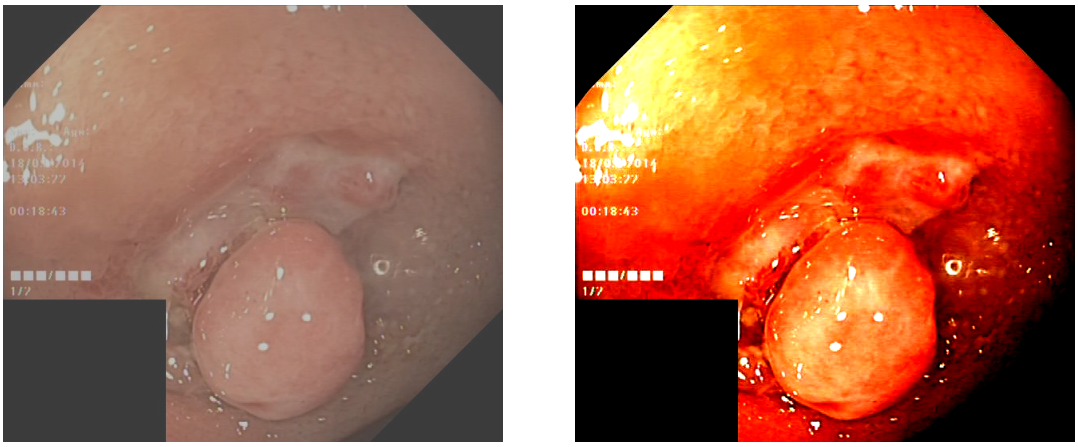


Figure 6: In light gray, histogram of the original image, in dark gray when contrast < 0



(a) (b)

Figure 7: Low contrast output (a) and high contrast output (b) images showing the effect of contrast adjustment created by the above theory.

2.3 Blur

Because WCE acquires images during a slow squirm process and transmits them from inside of the body by a wireless transmitter, the received images are often blurred due primarily to the complicated environment of the intestine and intrinsic restrictions of the equipment in terms of image acquisition and transmission. This in turn imposes difficulties for accurate and effective diagnosis. Capsule endoscopes are usually equipped with fisheye lenses that have small depths of field. Blurred images may be obtained due to fast camera motions with

low frame rates and the use of the wrong lens focus. These reasons caused two types of possible occurred blur (Motion blur and Defocus blur)

2.3.1 Motion blur

To create the Motion Blur, I have used a Motion Blur Filter applying motion blur to an image boils down to convolving a filter across the image. Sample 5×5 filter filters are given below.

Vertical

$$\frac{1}{5}\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Figure 8: Motion blur vertical filter

Horizontal

$$\frac{1}{5}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 9: Motion blur horizontal filter

The greater the size of the filter, the greater will be the motion blur effect. Further, the direction of 1's across the filter grid is the direction of the desired motion. To customize a motion blur in a specific vector direction, e.g. diagonally, simply place the first along the vector to create the filter.

2.3.2 Defocus blur

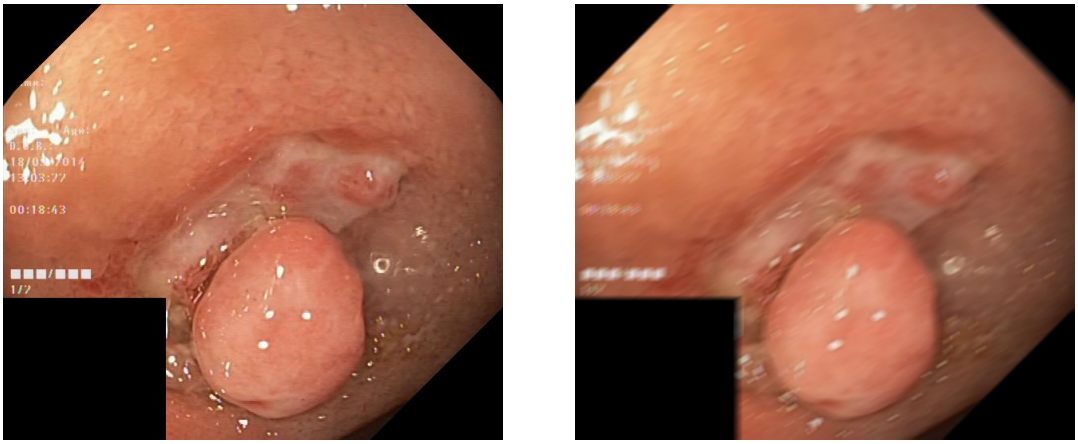
The idea to create the defocus blur is the same to motion blur but I used a different type of filter. Concretely, I created a Gaussian Kernel. After that, Gaussian Blur is done by convolving an image with a above normalized box filter.

$$G_0(x, y) = Ae^{\frac{-(x-\mu_x)^2}{2\sigma_x^2} + \frac{-(y-\mu_y)^2}{2\sigma_y^2}} \tag{5}$$

where μ is the mean (the peak) and σ^2 represents the variance (per each of the variables x and y)

2.3.3 Results

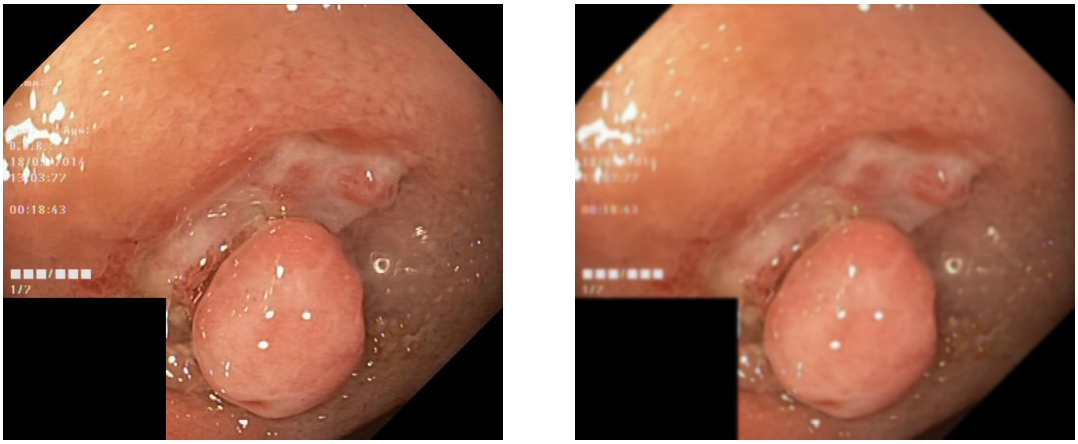
Motion blur



(a) (b)

Figure 10: Input (a) and output (b) images showing the effect of motion blur created by the above theory. Note that we can change the level and the direction of motion.

Defocus blur



(a) (b)

Figure 11: Input (a) and output (b) images showing the effect of defocus blur created by the above theory. Note that we can change the level of blur

2.4 Uneven illumination

In WCE, it is clear to know that the lighting modules consist of a bundle of flexible glass fibers that can each have a diameter as small as 30 microns. They provide shadowless illumination because their output angle exactly matches the 120-degree imaging angle of a camera. When making hundreds of thousands of single-use endoscopes, it becomes extremely

time-consuming and expensive to handle and install glass optical fibers for each instrument. Therefore, LEDs and plastic fibers became a new solution. However, they typically have output angles that are greater or less than before imaging angle of the camera, which can produce uneven lighting or shadows that degrade image quality. Especially thin endoscopes such as bronchoscopes, rhinoscopes, and cystoscopes have very little space for illumination, leading to extremely limited outer diameters for light guides. It made us to be careful with this type of distortion.

VALUE (OR BRIGHTNESS) in HSV color space will be used to adjust to illumination of original image. To be clearer, **VALUE (OR BRIGHTNESS)** works in conjunction with saturation and describes the brightness or intensity of the color, from 0 to 100 percent, where 0 is completely black, and 100 is the brightest and reveals the most color.

To simulate this kind of distortion, a scheme has been established as follow:

1. Convert the original image from RGB color space to HSV color space.
2. A circular-gradient mask will be generated which represent the uneven illumination distortion.
3. In HSV color space, channel V will be extracted and applied above mask.
4. Convert the distorted image back to RGB channel.

The most important step is step 2. We have to create the mask as close as possible to a real image. Because if the light source is not straight to the view direction, the illumination mask will be changed to present correctly the shadow sign.

2.4.1 Results

I would like to show two different result which present us the two different shadow shape for uneven illumination.

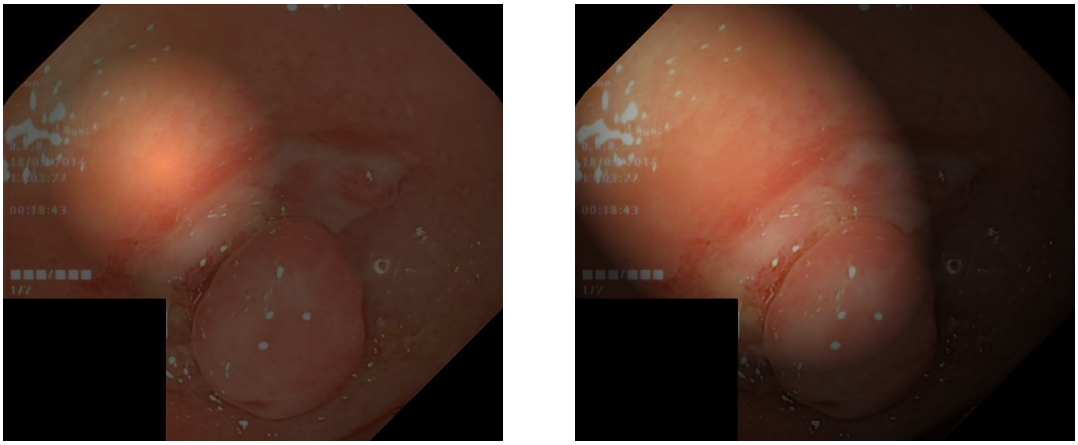
2.5 Geometric distortion

Endoscopes usually have severe barrel distortions. An endoscope needs a short focal length and a wide FOV in order to observe a broad area with minimum moving or bending of the endoscope, which is essential for steady and smooth manipulation of the endoscope because of the restricted space and degrees of freedom of movement and the limitation in hand-eye coordination during surgical cases. However, lenses used in endoscopes usually have a short focal length (a few millimeters only) and a wide FOV (ranging from 100 to 170 deg), which inevitably causes severe distortions. Typically, endoscopes exhibit barrel distortions.

In this part, I will present us a simple method to simulate the barrel distortion on an image using the physical and mathematical theory.

Firstly, I will explain the image formation from a geometrical point of view. Specifically, I will cover the math behind how a point in 3D gets projected on the image plane.

To understand the problem easily, let's say you have a camera deployed in a room. Given a 3D point $\mathbf{P}$ in this room, we want to find the pixel coordinates (u, v) of this 3D point in the image taken by the camera. There are three coordinate systems in play in this setup. Let's go over them.



(a) (b)

Figure 12: Output with 'circle' shape (a) and 'oval' shape (b) images showing the effect of uneven illumination created by the above theory. Note that we can change the direction, position and illumination level of distortion.

2.5.1 World Coordinate System

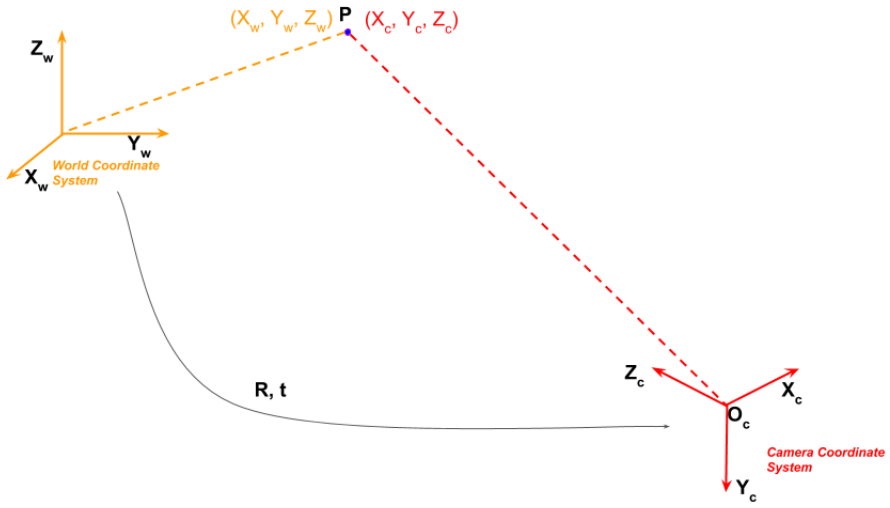


Figure 13: The World Coordinate System and the Camera Coordinate System are related by a Rotation and a translation. These six parameters (3 for rotation, and 3 for translation) are called the extrinsic parameters of a camera.

To define locations of points in the room we need to first define a coordinate system for this room. It requires two things:

- Origin : We can arbitrarily fix a corner of the room as the origin (0,0,0).
- X, Y, Z axes : We can also define the X and Y axis of the room along the two dimensions

on the floor and the Z axis along the vertical wall

Using the above, we can find the 3D coordinates of any point in this room by measuring its distance from the origin along the X, Y, and Z axes.

This coordinate system attached to the room is referred to as the World Coordinate System. In Fig. 13, it is shown using orange colored axes. We will use bold font (e.g. $\mathbf{X_w}$) to show the axis, and regular font to show a coordinate of the point (e.g. X_w).

Let us consider a point P in this room. In the world coordinate system, the coordinates of P are given by (X_w, Y_w, Z_w). You can find X_w, Y_w, and Z_w coordinates of this point by simply measuring the distance of this point from the origin along the three axes.

2.5.2 Camera Coordinate System

The image of the room will be captured using this camera, and therefore, we are interested in a 3D coordinate system attached to this camera.

If we had put the camera at origin of the room, and align it such that its $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ axes aligned with the $\mathbf{X_w}$,$\mathbf{Y_w}$, and $\mathbf{Z_w}$ axes of the room, the two coordinate systems would be the same.

However, that is an absurd restriction. We would want to put the camera anywhere in the room and it should be able to look anywhere. In such a case, we need to find the relationship between the 3D room (i.e. world) coordinates and the 3D camera coordinates.

Let's say our camera is located at some arbitrary location (t_X, t_Y, t_Z) in the room. In technical jargon, we can the camera coordinate is translated by (t_X, t_Y, t_Z) with respect to the world coordinates.

The camera may be also looking in some arbitrary direction. In other words, we can say the camera is rotated with respect to the world coordinate system.

Rotation in 3D is captured using three parameters — you can think of the three parameters as yaw, pitch, and roll. You can also think of it as an axis in 3D (two parameters) and an angular rotation about that axis (one parameter).

The world coordinate and the camera coordinates are related by a rotation matrix $\mathbf{R}$ and a 3 element translation vector $\mathbf{t}$. It means that point $\mathbf{P}$ which had coordinate values (X_w, Y_w, Z_w) in the world coordinates will have different coordinate values (X_c, Y_c, Z_c) in the camera coordinate system. We are representing the camera coordinate system using red color. The two coordinate values are related by the following equation.

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \mathbf{R} \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} + \mathbf{t} \tag{6}$$

Notice that representing rotation as a matrix allowed us to do rotation with a simple matrix multiplication instead of tedious symbol manipulation required in other representations like yaw, pitch, roll.

Sometimes the expression above is written in a more compact form. The 3×1 translation vector is appended as a column at the end of the 3×3 rotation matrix to obtain a 3×4

matrix called the Extrinsic Matrix.

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = [\mathbf{R}|\mathbf{t}] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{7}$$

where, the extrinsic matrix $\mathbf{P}$ is given by

$$\mathbf{P} = [\mathbf{R}|\mathbf{t}] \tag{8}$$

A 3D point(X, Y, Z) in cartesian coordinates can written as $(X, Y, Z, 1)$ in homogenous coordinates. More generally, a point in homogenous coordinate (X, Y, Z, W) is the same as the point $(X/W, Y/W, Z/W)$ in cartesian coordinates. Homogenous coordinates allow us to represent infinite quantities using finite numbers. For example, the point at infinity can be represented as $(1, 1, 1, 0)$ in homogenous coordinates.

2.5.3 Image Coordinate System

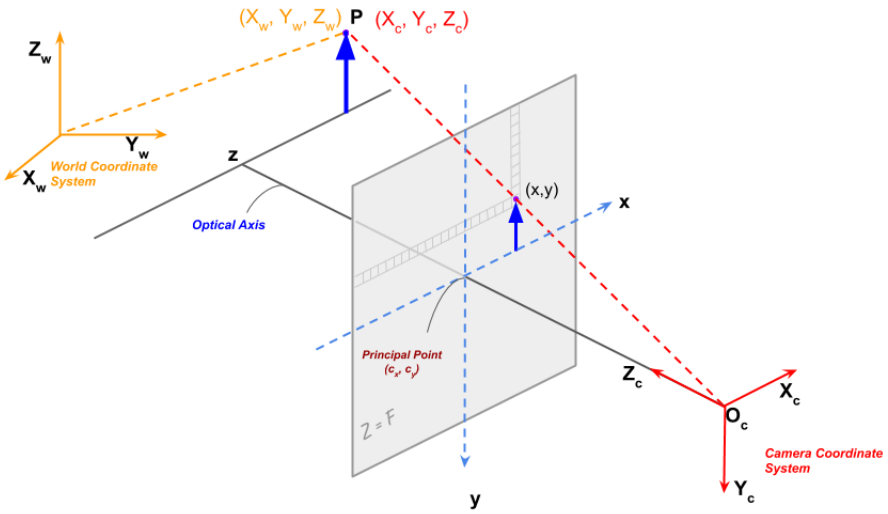


Figure 14: The projection of point $\mathbf{P}$ onto the image plane is shown.

Once we get a point in 3D coordinate system of the camera by applying a rotation and translation to the points world coordinates, we are in a position to project the point on the image plane to obtain a location of the point in the image. In the image above, we are looking at a point P with coordinates (X_c, Y_c, Z_c) in the camera coordinate system.

The optical center (pin hole) is represented using $\mathbf{O}_c$. In reality an inverted image of the point is formed on the image plane. For mathematical convenience, we simply do all the calculations as if the image plane is in front of the optical center because the image read out from the sensor can be trivially rotated by 180 degrees to compensate for the inversion

The image plane is placed at a distance f (focal length) from the optical center. Using high school geometry (similar triangles), we can show the project image (x, y) of the 3D point (Xc, Yc, Zc) is given by.

$$x = f\frac{X_c}{Z_c} \tag{9}$$

$$y = f\frac{Y_c}{Z_c} \tag{10}$$

The above two equations can be rewritten in matrix form as follows

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \tag{11}$$

The matrix **K** shown below is called the Intrinsic Matrix and contains the intrinsic parameters of the camera. The above simple matrix shows only the focal length. However, the pixels in the image sensor may not be square, and so we may have two different focal lengths f_x and f_y. The optical center (c_x, c_y) of the camera may not coincide with the center of the image coordinate system (Fig. 15). In addition, there may be a small skew γ between the x and y axes of the camera sensor. Taking all the above into account, the camera matrix can be re-written as.

$$\mathbf{K} = \begin{bmatrix} f & \gamma & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{12}$$

However, in the above equation, the x and y pixel coordinates are with respect to the center of the image. However, while working with images the origin is at the top left corner of the image. Let's represent the image coordinates by (u, v).

$$\begin{bmatrix} u' \\ v' \\ w' \end{bmatrix} = \begin{bmatrix} f & \gamma & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \tag{13}$$

where

$$u = \frac{u'}{w'} \tag{14}$$

$$v = \frac{v'}{w'} \tag{15}$$

From above equations, projecting a 3D point in world coordinate system to camera pixel coordinates is done in three steps.

- The 3D point is transformed from world coordinates to camera coordinates using the Extrinsic Matrix which consists of the Rotation and translation between the two coordinate systems.
- The new 3D point in camera coordinate system is projected onto the image plane using the Intrinsic Matrix which consists of internal camera parameters like the focal length, optical center, etc.

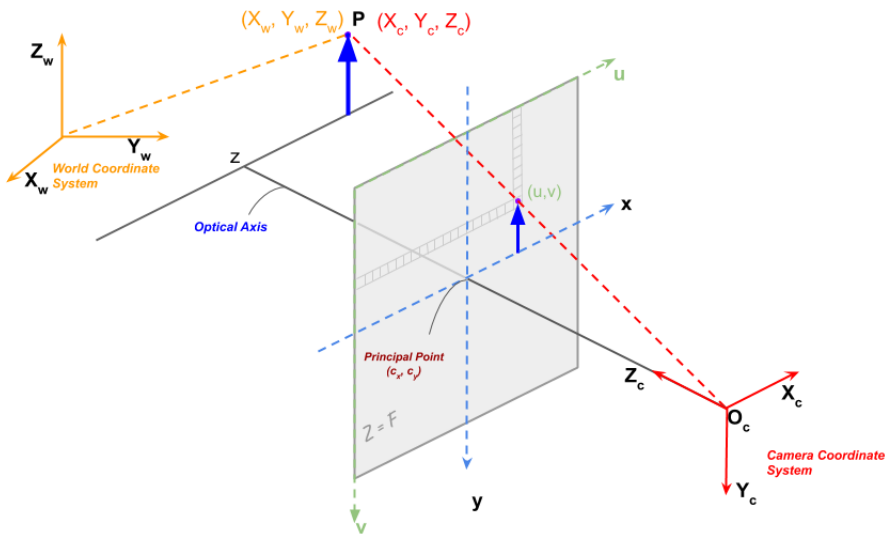


Figure 15: Shows a more realistic scenario when the image pixel coordinate system has the origin on the top left corner. The intrinsic camera matrix needs to take into account the location of the principal point, the skew of the axes, and potentially different focal lengths along different axes.

From equation (7),(8) and (13), the equations that relate 3D point (X_w, Y_w, Z_w) in world coordinates to its projection (u, v) in the image coordinates are shown below

$$\begin{bmatrix} u' \\ v' \\ w' \end{bmatrix} = \mathbf{Q} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{16}$$

where

$$\mathbf{Q} = \mathbf{K} \times \mathbf{P} = \mathbf{K} \times \begin{bmatrix} \mathbf{R} | \mathbf{t} \end{bmatrix} \tag{17}$$

and the translation matrix is:

$$\begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{18}$$

The rotation matrix is:

$$\mathbf{R} = \mathbf{R}_x \times \mathbf{R}_y \times \mathbf{R}_z \tag{19}$$

where

$$\mathbf{R}_x = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha & 0 \\ 0 & \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{20}$$

$$\mathbf{R}_y = \begin{bmatrix} \cos\beta & 0 & \sin\beta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\beta & 0 & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{21}$$

$$\mathbf{R}_z = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 & 0 \\ \sin\gamma & \cos\gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{22}$$

with α, β and γ is the adjustable parameter presenting the rotation angle in corresponding to each axis.

Now we know that a 3D point (X_w, Y_w, Z_w) in the world coordinates is mapped to its corresponding pixel coordinates (u, v) based on the above equation, where $\mathbf{Q}$ is the camera projection matrix.

Some pinhole cameras introduce significant distortion to images. Two major kinds of distortion are radial distortion and tangential distortion. Radial distortion causes straight lines to appear curved. Radial distortion becomes larger the farther points are from the center of the image. Therefore, the projection matrix will be convert to:

Radial distortion

$$u'_{distorted} = u'(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \tag{23}$$

$$v'_{distorted} = v'(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \tag{24}$$

Tangential distortion

Tangential distortion occurs because the image-taking lense is not aligned perfectly parallel to the imaging plane. So, some areas in the image may look nearer than expected. The amount of tangential distortion can be represented as below:

$$u'_{distorted} = u' + [2p_1 u'v' + p_2(r^2 + 2u'^2)] \tag{25}$$

$$v'_{distorted} = v' + [p_1(r^2 + 2v'^2) + 2p_2 u'v'] \tag{26}$$

where

$$r = \sqrt{u'^2 + v'^2} \tag{27}$$

In short, we need to find five parameters, known as distortion coefficients given by:

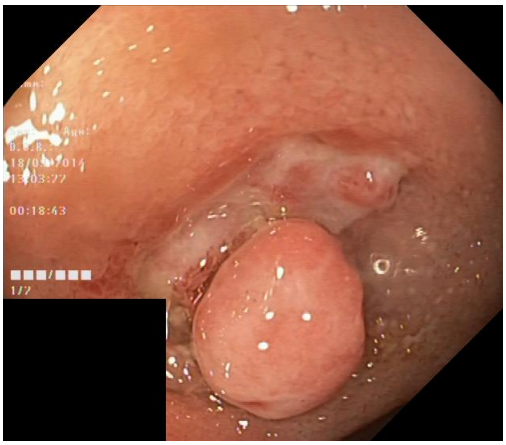
$$Distortion\ coefficients = (k_1 \quad k_2 \quad p_1 \quad p_2 \quad k_3)$$

Now, we have been able to create barrel distortion by doing the following scheme:

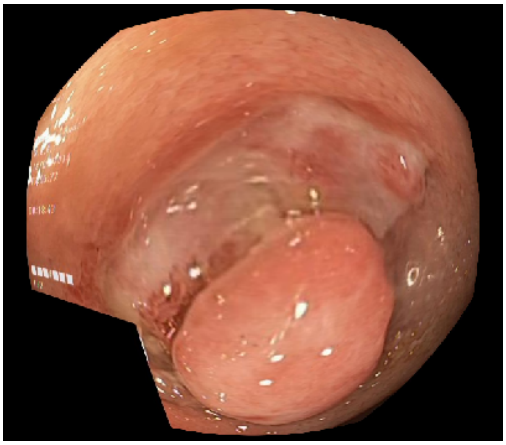
1. Create a virtual camera
2. Define a 3D surface (the mirror surface) and project it into the virtual camera using a suitable value of projection matrix. This projection matrix will be additionally affected by given distortion
3. Use the image coordinates of the projected points of the 3D surface to apply mesh based warping to get the desired effect.

2.5.4 Result

I would like to show two different result which present us the two different kinds of geometric distortion.



(a)



(b)

(c)

Figure 16: Output with 'Radial' distortion (a) and 'Tangential' distortion (b) images showing the effect of geometric created by the above theory. Note that we can change the level of distortion.

References

[1] Hanna Borgli et al. "Hyper-Kvasir: A Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy". In: (Dec. 2019). DOI: 10.31219/osf.io/mkzcq.